CHARACTERIZING THE SLOPE TRADE-OFF: A VARIATIONAL PERSPECTIVE AND THE DONOHO–TANNER LIMIT

BY ZHIQI BU[1], JASON M. KLUSOWSKI[2], CYNTHIA RUSH[3], AND WEIJIE J. SU[4]

[1]*Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania, zbu@sas.upenn.edu*

[2]*Department of Operations Research and Financial Engineering, Princeton University, jason.klusowski@princeton.edu*

[3]*Department of Statistics, Columbia University, cynthia.rush@columbia.edu*

[4]*Department of Statistics and Data Science, University of Pennsylvania, suw@wharton.upenn.edu*

Sorted ℓ_1 regularization has been incorporated into many methods for solving high-dimensional statistical estimation problems, including the SLOPE estimator in linear regression. In this paper, we study how this relatively new regularization technique improves variable selection by characterizing the optimal SLOPE trade-off between the false discovery proportion (FDP) and true positive proportion (TPP) or, equivalently, between measures of type I error and power. Assuming a regime of linear sparsity and working under Gaussian random designs, we obtain an upper bound on the optimal trade-off for SLOPE, showing its capability of breaking the Donoho–Tanner power limit. To put it into perspective, this limit is the highest possible power that the Lasso, which is perhaps the most popular ℓ_1-based method, can achieve even with arbitrarily strong effect sizes. Next, we derive a tight lower bound that delineates the fundamental limit of sorted ℓ_1 regularization in optimally trading the FDP off for the TPP. Finally, we show that on any problem instance, SLOPE with a certain regularization sequence outperforms the Lasso, in the sense of having a smaller FDP, larger TPP and smaller ℓ_2 estimation risk simultaneously. Our proofs are based on a novel technique that reduces a calculus of variations problem to a class of infinite-dimensional convex optimization problems and a very recent result from approximate message passing theory.

1. Introduction. Reconstructing the signal from noisy linear measurements is vital in many disciplines, including statistical learning, signal processing, and biomedical imaging. In many modern applications where the number of explanatory variables often exceeds the number of measurements, the signal is often believed—or, wished—to be sparse in the sense that most of its entries are zero or approximately zero. Put differently, this means that a majority of the explanatory variables are simply irrelevant to the response of interest.

Accordingly, a host of methods have been developed to tackle these problems by leveraging the sparsity of signals in high-dimensional linear regression. These methods often rely on, among others, the concept of *regularization* to constrain the search space of the unknown signals. Perhaps the most influential instantiation of this concept is ℓ_1 regularization, which gives rise to the Lasso method (Tibshirani, 1996). The optimal amount of regularization, however, hinges on the sparsity level of the signal. Intuitively speaking, if the sparsity level is

MSC2020 subject classifications: Primary 62F03, 62E20; secondary 62J05, 62J07.

Keywords and phrases: SLOPE, false discovery rate, true positive rate, approximate message passing, sorted ℓ_1 regularization, phase transition.

Author names are listed alphabetically.

low, then more regularization should be imposed, and vice versa (see, for example, Abramovich et al. (2006)).

This intuition necessitates the development of a regularization technique that is adaptive to the sparsity level of signals, which is typically unknown in practical problems. To achieve this desired adaptivity, Bogdan et al. (2015) introduced *sorted* ℓ_1 *regularization*. This new regularization technique turns into a method called SLOPE in the setting of a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \tag{1.1}$$

where $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients, $\boldsymbol{y} \in \mathbb{R}^n$ is the response, and $\boldsymbol{w} \in \mathbb{R}^n$ is the noise term. Explicitly, SLOPE estimates the coefficients by solving the convex programming problem

$$\arg\min_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \tag{1.2}$$

where $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$ are the order statistics in absolute value of $\boldsymbol{b} = (b_1, \ldots, b_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ (with at least one strict inequality) are the regularization parameters. The sorted ℓ_1 penalty, $\sum_{i=1}^{p} \lambda_i |b|_{(i)}$, is a norm, and the optimization problem for SLOPE is, therefore, convex (see also Figueiredo and Nowak (2016)). As an important feature, the sorted ℓ_1 norm penalizes larger entries more heavily than smaller ones. Indeed, this regularization technique is shown to be adaptive to the degree of sparsity level and enables SLOPE to obtain optimal estimation performance for certain problems (Su and Candès, 2016). Notably, in the special case $\lambda_1 = \cdots = \lambda_p$, the sorted ℓ_1 norm reduces to the usual ℓ_1 norm. Thus, the Lasso can be regarded as a special instance of SLOPE.

A fundamental question, yet to be better addressed, is how to quantitatively characterize the benefits of using the sorted ℓ_1 regularization. To explore this question, Figure 1 compares the model selection performance of SLOPE and the Lasso in terms of the *false discovery proportion* (FDP) and *true positive proportion* (TPP) or, equivalently, between measures of type I error and power. Needless to say, a model is preferred if its FDP is small while its TPP is large. As the first impression conveyed by this figure, both methods seem to undergo a trade-off between the FDP and TPP when the TPP is below a certain limit. More interestingly, while *nowhere* on the Lasso path is the TPP above a limit, which is about 0.5707 in the left plot of Figure 1 and 0.4343 in the right, SLOPE is able to pass the limit toward achieving full power. To be sure, these contrasting patterns persist even for an arbitrarily large signal-to-noise ratio. This distinction must be attributed to the flexibility of the SLOPE regularization sequence $(\lambda_1, \ldots, \lambda_p)$ compared to a single value as in the Lasso case. Recognizing this message, we are tempted to ask (1) *why* the use of sorted ℓ_1 regularization brings a significant benefit over ℓ_1 regularization in the high TPP regime and, equally importantly, (2) *why* SLOPE exhibits a trade-off between the FDP and TPP just as the Lasso does in the low TPP regime.

1.1. *A peek at our results.* To address these two questions, in this paper we characterize the optimal trade-off of SLOPE between the TPP and FDP, uncovering several intriguing findings of sorted ℓ_1 regularization. Assuming TPP $\approx u$ for $0 \leq u \leq 1$, loosely speaking, the trade-off curve gives the smallest possible value of the FDP of SLOPE using any regularization sequence in the large system limit. To prepare for a rough description of our contributions, in brief, we work in the setting where the design has i.i.d. Gaussian entries and the regression coefficients $\beta_1, \ldots, \beta_p$ are i.i.d. draws from a distribution that takes non-zero values with a certain probability. Notably, it is generally nontrivial to define false discoveries in high dimensions (G'Sell et al., 2013), which is not an issue however in the case of independent

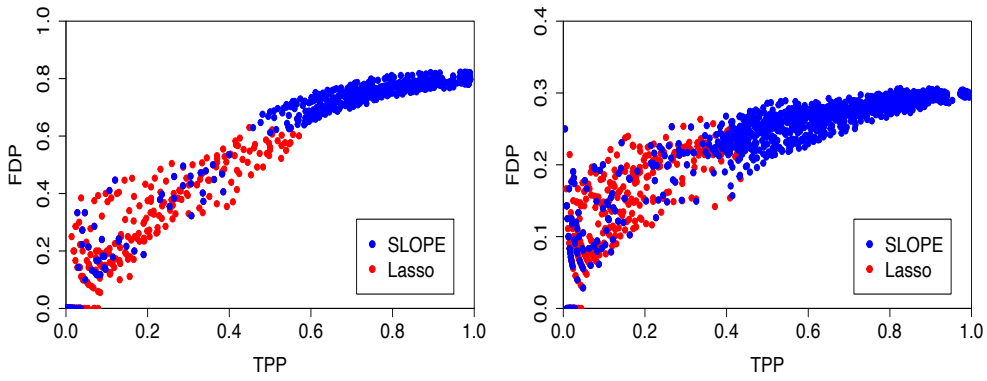

Fig 1: Comparison between SLOPE and the Lasso in terms of the TPP–FDP trade-off. Given an estimate $\widehat{\boldsymbol{\beta}}$, define its FDP $= \frac{|\{j:\beta_j=0 \text{ and } \widehat{\beta}_j\neq 0\}|}{|\{j:\widehat{\beta}_j\neq 0\}|}$ and TPP $= \frac{|\{j:\beta_j\neq 0 \text{ and } \widehat{\beta}_j\neq 0\}|}{|\{j:\beta_j\neq 0\}|}$. The SLOPE regularization sequence $\boldsymbol{\lambda}_{\lambda,r\lambda,w}$ is defined in (2.5), with varying $0 < r < 1$ and $\lambda > 0$, and $w = 0.2$ in the left plot and $w = 0.3$ in the right plot. The results of the Lasso are taken over its entire solution path, and its highest TPP is about 0.5707 in the left plot and 0.4343 in the right plot. Left: $(n, p) = (300, 1000), |\{j : \beta_j \neq 0\}|/p = 0.2$, and $\boldsymbol{w} = \boldsymbol{0}$ (noiseless); right: $(n, p) = (400, 1000), |\{j : \beta_j \neq 0\}|/p = 0.7$, and $\boldsymbol{w} = \boldsymbol{0}$. On both plots, non-zero entries of $\boldsymbol{\beta}$ are i.i.d. draws from the standard normal distribution. More specifications of the setup are detailed in Section 2. The result presents 10 independent trials.

regressors. The assumption on the signal prior corresponds to the *linear sparsity* regime. In addition, we assume that both $n, p \to \infty$ and the sampling ratio n/p converges to a constant (see more detailed assumptions in Section 2). From a technical viewpoint, these assumptions allow us to make use of tools from approximate message passing (AMP) theory (Donoho et al., 2009; Bayati and Montanari, 2011).

Breaking the Donoho–Tanner power limit. To explain the contrasting results presented in Figure 1, we prove that under the aforementioned assumptions, SLOPE can achieve an arbitrarily high TPP. Moving from sorted ℓ_1 regularization to ℓ_1 regularization, in stark contrast, the Lasso exhibits the Donoho–Tanner (DT) power limit when $n < p$ and the sparsity is above a certain threshold (Donoho, 2006, 2005). Informally, the DT power limit is the largest possible power that any estimate along the Lasso path can achieve in the large system limit. For example, in the setting of Figure 1 this power limit is about 0.5676 in the left plot and 0.4401 in the right plot. For SLOPE and a certain choice of the regularization sequence, interestingly, we show that the asymptotic TPP-FDP trade-off of SLOPE beyond the DT power limit is given by a simple Möbius transformation, which is shown by the blue curve in Figure 2. This Möbius transformation naturally serves as an upper bound on the (optimal) SLOPE trade-off curve above the DT power limit.

Lower bound via convex optimization. Next, we address the second question by lower bounding the optimal trade-off for SLOPE, followed by a comparison between the trade-offs for the two methods in the low TPP regime. To put it into perspective, the Lasso trade-off obtained by Su et al. (2017) is plotted as the green solid curve in Figure 2. Apart from the simple fact that the SLOPE trade-off is better than or equal to the Lasso counterpart, however, it requires new tools to take into account the structure of sorted ℓ_1 regularization. To this end, we develop a technique based on a class of infinite-dimensional convex optimization problems. The resulting lower bound is shown in red in Figure 2. It is worth noting that the development of this technique presents several novel ideas that might be of independent interest for other regularization schemes.

Instance superiority of SLOPE. The results illustrated so far are taken from an optimal-case viewpoint. Moving to a more practical standpoint, we are interested in comparing the two methods on a specific problem instance and, in particular, wish to find a SLOPE regularization sequence that allows SLOPE to outperform the Lasso with any given penalty parameter in terms of, for example, the TPP, the FDP, or the ℓ_2 estimation risk. Surprisingly, we prove that on any problem instance, SLOPE can dominate the Lasso according to these three indicators simultaneously. This comparison conveys the message that the flexibility of the sorted ℓ_1 regularization can turn into appreciable benefits. This result is formally stated in Theorem 3.

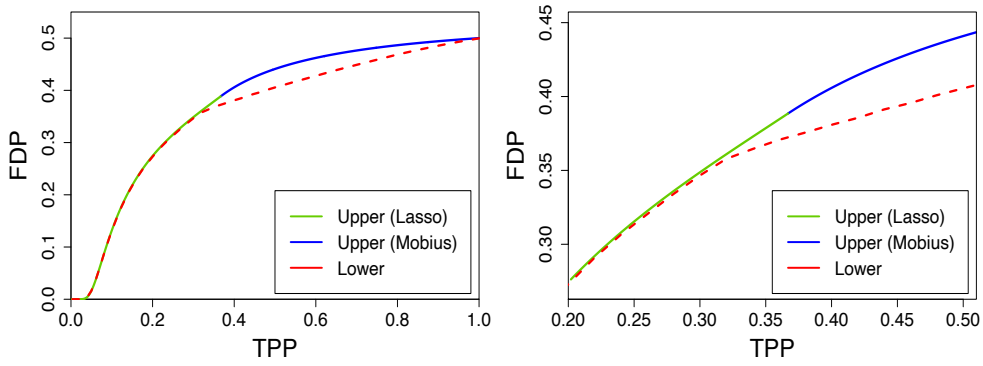

Fig 2: Illustration of the upper bound $q^\star$ and lower bound $q_\star$ for the SLOPE TPP–FDP trade-off. The right plot is the zoom-in of the left. Here $n/p = 0.3$ and $|\{j : \beta_j \neq 0\}|/p = 0.5$ (see more details in the working assumptions in Section 2). The Lasso trade-off curve shown in green is truncated at the DT power limit about 0.3669 (Su et al., 2017). The optimal SLOPE trade-off curve must lie between the two curves. Notably, the two bounds agree at TPP $= 1$.

1.2. *Organization.* The remainder of this paper is structured as follows. In Section 2, we present the main results of this paper. Next, Section 3 introduces the AMP machinery at a minimal level as a preparation for the proofs of our main results. In Section 4, we detail the derivation of the lower bound based on variational calculus and infinite-dimensional convex optimization. In Section 5, we specify the upper bound, especially the part given by a Möbius transformation above the DT power limit. We conclude this paper in Section 6 by proposing several future research directions. Omitted proofs are relegated to the appendix (Bu et al., 2022).

2. Main results. Throughout this paper, we make the following working assumptions to specify the design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, and noise $\boldsymbol{w} \in \mathbb{R}^n$ in the linear model (1.1), as well as the SLOPE regularization sequence $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$. To obviate any ambiguity, we consider a sequence of problems indexed by (n, p) with both n, p tending to infinity.

(A1) The matrix $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries. The sampling ratio n/p converges to a constant $\delta > 0$.
(A2) The entries of $\boldsymbol{\beta}$ are i.i.d. copies of a random variable Π satisfying $\mathbb{P}(\Pi \neq 0) = \epsilon$ for a constant $0 < \epsilon < 1$ and $\mathbb{E}(\Pi^2 \max\{0, \log \Pi\}) < \infty$. The noise vector $\boldsymbol{w}$ consists of i.i.d. copies of a random variable W with bounded second moment $\sigma^2 := \mathbb{E}(W^2) < \infty$.
(A3) The SLOPE regularization sequence $\boldsymbol{\lambda}(p) = (\lambda_1, \ldots, \lambda_p)$ is the order statistics of p i.i.d. realizations of a (nontrivial) non-negative random variable Λ.

Moreover, we assume that $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{w}$ are independent. Notice that the sparsity level of $\boldsymbol{\beta}$ is about ϵp and that each column of $\boldsymbol{X}$ has approximately a unit ℓ_2 norm. The noise variance σ^2 can equal 0, meaning that our results apply to both noisy and noiseless settings. In (A3), by "nontrivial" we mean that Λ is not always equal to 0. As an aside, SLOPE is reduced to the Lasso if the distribution of Λ is a unit probability mass at some positive value.

The working assumptions are mainly driven by their necessity in AMP theory (Donoho et al., 2009; Bayati and Montanari, 2011), which enables the use of the recent analysis of an AMP algorithm when applied to solve SLOPE (Bu et al., 2020) (similar analysis is given in Hu and Lu (2019) and requires similar assumptions). Regarding (A2), the condition $\mathbb{P}(\Pi \neq 0) = \epsilon$, which implies linear sparsity of the regression coefficients, is not required for AMP theory. Rather, this condition is only made so that the TPP and FDP are well-defined. Besides, the merit of the linear sparsity regime has been increasingly recognized in the high-dimensional literature (Mousavi et al., 2018; Weng et al., 2018; Su, 2018; Sur et al., 2019; Wang et al., 2019).

2.1. *Bounds on the SLOPE trade-off.* Our main result is the characterization of a trade-off curve that teases apart asymptotically achievable TPP and FDP pairs from the asymptotically unachievable pairs for SLOPE[1]. For any estimate $\widehat{\boldsymbol{\beta}}$, recall that its FDP and TPP are defined as

$$\text{FDP} = \frac{|\{j : \beta_j = 0 \text{ and } \widehat{\beta}_j \neq 0\}|}{|\{j : \widehat{\beta}_j \neq 0\}|}, \quad \text{TPP} = \frac{|\{j : \beta_j \neq 0 \text{ and } \widehat{\beta}_j \neq 0\}|}{|\{j : \beta_j \neq 0\}|}, \tag{2.1}$$

with the convention $0/0 = 0$. When it comes to the SLOPE estimator, we use $\text{TPP}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ and $\text{FDP}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ to denote its TPP and FDP, respectively.

Likewise, we define the thresholded FDP and TPP, namely,

$$\text{FDP}_\xi = \frac{|\{j : \beta_j = 0 \text{ and } |\widehat{\beta}_j| > \xi\}|}{|\{j : |\widehat{\beta}_j| > \xi\}|}, \quad \text{TPP}_\xi = \frac{|\{j : \beta_j \neq 0 \text{ and } |\widehat{\beta}_j| > \xi\}|}{|\{j : \beta_j \neq 0\}|}, \tag{2.2}$$

which reduce to FDP and TPP when $\xi = 0$. These thresholded versions of FDP and TPP are introduced purely for technical reasons, and have been used in previous work on penalized estimators like SLOPE including in Wang et al. (2020b). Specifically, the SLOPE estimator is known to possibly have many elements that are very close to zero, but not strictly equal to zero, causing the direct asymptotic analysis of the FDP and TPP defined in (2.1) to be difficult. We refer interested readers to Hu and Lu (2019, Example 3 and Figure 3) for a concrete example that illustrates such a phenomenon. Instead, we analyze asymptotic (in p) properties of FDP_ξ and TPP_ξ in (2.2) and then allow $\xi \to 0$ to recover asymptotic properties of FDP and TPP defined in (2.1).

Our main results are stated in the following two theorems, which give lower and upper bounds on the optimal SLOPE trade-off. Taken together, they demonstrate a fundamental separation between asymptotically achievable TPP–FDP pairs and the unachievable pairs over all signal priors Π and SLOPE regularization sequences $\boldsymbol{\lambda}$. Note that both the upper bound $q^\star$ and lower bound $q_\star$ are defined on $[0, 1]$ and completely determined by ϵ and δ. The expression for $q^\star$ is given in (2.9), while $q_\star$ is detailed in Section 4.

THEOREM 1 (Lower bound). *Under the working assumptions, namely (A1), (A2), and (A3), the following inequality holds with probability tending to one:*

$$\text{FDP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \geq q_\star\left(\text{TPP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}); \delta, \epsilon\right) - c_\xi,$$

[1]R code to reproduce the results, e.g., to calculate $q_\star$ and $q^\star$, is available at `https://github.com/woodyx218/SLOPE_AMP`.

where $q_\star(u;\delta,\epsilon) > 0$ for all $0 < u < 1$ and c_ξ is some positive constant which tends to 0 as $\xi \to 0$.

THEOREM 2 (Upper bound). *Under the working assumptions, namely (A1), (A2), and (A3), for any $0 \le u \le 1$, there exist a signal prior Π and a SLOPE regularization prior Λ such that the following inequalities hold with probability tending to one:*

$$\mathrm{FDP}_\xi(\boldsymbol{\beta},\boldsymbol{\lambda}) \le q^\star\left(\mathrm{TPP}_\xi(\boldsymbol{\beta},\boldsymbol{\lambda});\delta,\epsilon\right) + c_\xi(\Pi,\Lambda) \quad \textit{and} \quad |\,\mathrm{TPP}_\xi(\boldsymbol{\beta},\boldsymbol{\lambda}) - u| \le c_\xi(\Pi,\Lambda),$$

where $q^\star(u;\delta,\epsilon) < 1-\epsilon$ and c_ξ is some positive constant which tends to 0 as $\xi \to 0$.

REMARK 2.1. The probability is taken with respect to the randomness in the design matrix, regression coefficients, noise, and SLOPE regularization sequence in the large system limit $n, p \to \infty$. In relating to the assumptions made previously, this theorem holds even for $\sigma^2 = 0$, the noiseless case.

The proofs of Theorem 1 and Theorem 2 are given in Section 4 and Section 5, respectively. Most notably, our proof of Theorem 1 starts by formulating the problem of finding a tight lower bound as a calculus of variations problem. Relying on several novel elements, we further reduce this problem to a class of infinite-dimensional convex programs.

On the one hand, Theorem 1 says that it is impossible to achieve high power and a low FDP simultaneously using any sorted ℓ_1 regularization sequences, and this trade-off is specified by $q_\star$. On the other hand, Theorem 2 demonstrates that SLOPE can achieve at least the same trade-off as that given by $q^\star$ by specifying a prior Π and a regularization sequence $\boldsymbol{\lambda}$. Indeed, the proof of this theorem is constructive in that we will show that SLOPE can come arbitrarily close to any point on the curve $q^\star$ (see Section 5). Another important observation from Theorem 2 is that SLOPE can achieve any power levels, which is not necessarily the case for ℓ_1 regularization-based methods, as we show in Section 2.2.

Informally, let q_{SLOPE} denote the optimal SLOPE trade-off curve. That is, $q_{\text{SLOPE}}(u)$ is asymptotically the minimum possible value of the FDP under the constraint that the TPP is about u, over all possible SLOPE regularization sequences and signal priors (see formal definition in Section 3). Combining the two theorems above, we readily see that the optimal SLOPE trade-off must be sandwiched between $q^\star$ and $q_\star$:

$$q_\star(u) \le q_{\text{SLOPE}}(u) \le q^\star(u),$$

for all $0 \le u \le 1$. Consequently, the sharpness of the approximation to the SLOPE trade-off rests on the gap between the two curves, and throughout the paper, we refer to the gap as the function $u \mapsto q^\star(u) - q_\star(u)$. Figure 3 illustrates several examples of the two curves for various pairs of ϵ, δ. Importantly, the plots show that the two bounds are very close to each other, thereby demonstrating tightness of our bounds. In fact, the gap between $q_\star$ and $q^\star$ is an upper bound of the gap between the analytical $q^\star$ and the true trade-off q_{SLOPE}. Furthermore, a closer look at the plots reveals that the two curves seem to coincide exactly when the TPP is below a certain value. In this regard, the SLOPE trade-off might have been uncovered exactly in this regime of TPP. Future investigation is required to obtain a fine-grained comparison between the two curves.

Looking at Figure 3, the reader may initially find the non-monotonicity of the trade-off curves in ϵ as surprising. We argue that this is due to the DT phase transition: in the case of the Lasso, for fixed δ, it can be shown that the trade-off curves are monotonically increasing in ϵ; in other words, $q_{\text{Lasso}}(u;\delta,\epsilon_1) > q_{\text{Lasso}}(u;\delta,\epsilon_2)$ whenever $\epsilon_1 > \epsilon_2$. However, in some settings, we empirically observe that $\mathrm{TPP} = 1$ is achieved with a dense SLOPE estimator. When this

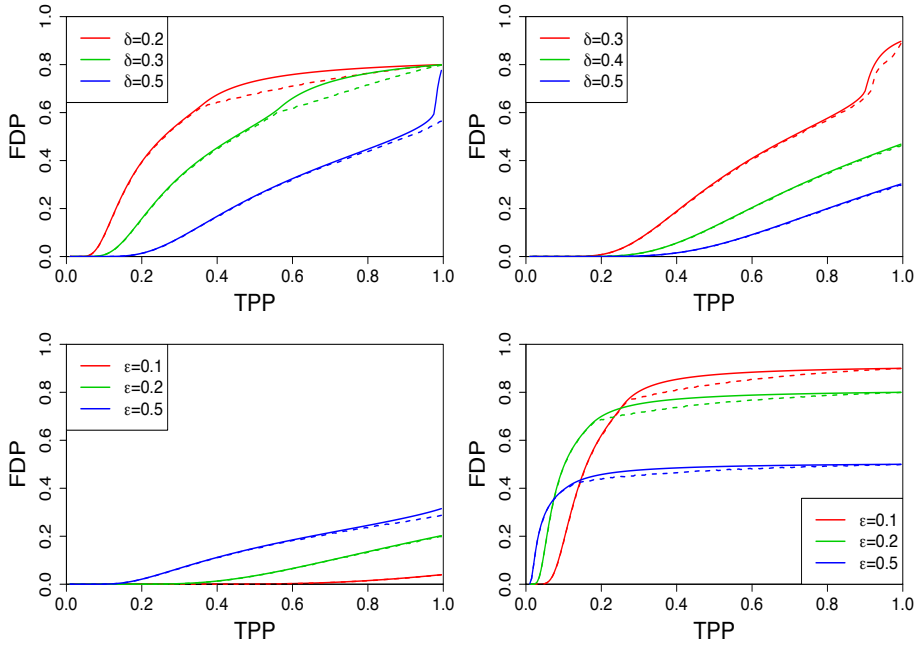

Fig 3: Examples of the SLOPE trade-off bounds $q^\star$ and $q_\star$ for different (δ, ϵ) pairs. Top-left: $\epsilon = 0.2$; top-right: $\epsilon = 0.1$; bottom-left: $\delta = 0.9$; bottom-right: $\delta = 0.1$. For a given δ, note that the trade-off for SLOPE is not monotone with respect to ϵ, which is a departure from the Lasso counterpart (see Su et al. (2017, Figure 4)). Numerically, the upper and lower bounds seem to coincide when the TPP is below a threshold (see more details in Figure 5). To give more details, in one regime with $\delta = 0.1, \epsilon = 0.5$, the maximum gap between the upper and lower bounds $\max_u[q^\star(u) - q_\star(u)]$ is less than 0.0235; whereas in another regime with $\delta = 0.5, \epsilon = 0.1$, the maximum gap is always less than 0.0056.

occurs, $q_{\text{SLOPE}}(1) = 1 - \epsilon$ and thus $q_{\text{SLOPE}}(1; \delta, \epsilon_1) < q_{\text{SLOPE}}(1; \delta, \epsilon_2)$. In words, the SLOPE trade-off at TPP $= 1$ is monotonically *decreasing* in ϵ. Therefore, the patterns may not be monotone between the TPP upper limit $u^\star_{\text{DT}}$ and 1, shifting from increasing in ϵ to decreasing in ϵ at the extreme. In short, the regime beyond DT phase limit is different for SLOPE and when SLOPE enters this regime, breaking the monotonicity in ϵ may occur.

To be complete, we remark that the message conveyed by these two theorems does not contradict earlier results established for FDR control of SLOPE (Bogdan et al., 2013, 2015; Brzyski et al., 2019; Kos and Bogdan, 2020). The crucial difference between the two sides arises from the linear sparsity assumed in the present paper, which is a clear departure from the much lower sparsity level considered in the literature. In this regard, our results complement the literature by extending our understanding of the inferential properties of the SLOPE method to an unchartered regime.

2.2. *Breaking the Donoho–Tanner power limit.* To better appreciate the trade-off results presented in Theorem 2 for SLOPE, it is instructive to compare them with the TPP and FDP trade-off for the Lasso, which is arguably the most popular method leveraging ℓ_1 regularization.

To put it into perspective, first recall some results concerning the optimal trade-off between the TPP and FDP for the Lasso. A surprising fact is that under the working assumptions,[2] the

[2]Note that, in the case of the Lasso, (A3) is replaced by the assumption that $\lambda > 0$ is a constant.

Lasso cannot achieve full power even with an arbitrarily large signal-to-noise ratio when $\delta < 1$ (that is, $\boldsymbol{X}$ is "fat") and the sparsity ratio ϵ is above a threshold, which we denote by $\epsilon^\star(\delta)$. The dependence of this value on δ is specified by the parametric equations

$$\delta = \frac{2\phi(s)}{2\phi(s) + s(2\Phi(s) - 1)}, \qquad \epsilon^\star = \frac{2\phi(s) - 2s\Phi(-s)}{2\phi(s) + s(2\Phi(s) - 1)}, \tag{2.3}$$

for $s > 0$.[3] For simplicity, henceforth (δ, ϵ) is said to be in the *supercritical* regime if $\delta < 1, \epsilon > \epsilon^\star(\delta)$. Otherwise, it is in the *subcritical* regime when $\delta < 1, \epsilon \le \epsilon^\star(\delta)$, or $\delta \ge 1$ (that is, $\boldsymbol{X}$ is "thin"). In the supercritical regime, Su et al. (2017) proved that the highest achievable TPP of the Lasso, denoted $u^\star_{\text{DT}}$, takes the form

$$u^\star_{\text{DT}}(\delta, \epsilon) := 1 - \frac{(1-\delta)(\epsilon - \epsilon^\star)}{\epsilon(1 - \epsilon^\star)} < 1. \tag{2.4}$$

Throughout the paper, $u^\star_{\text{DT}}$ is referred to as the *DT power limit*. For completeness, in the subcritical regime the Lasso can achieve any power level. As such, we formally set $u^\star_{\text{DT}}(\delta, \epsilon) = 1$ when $\delta < 1, \epsilon \le \epsilon^\star(\delta)$, or $\delta \ge 1$.

This existing result, in conjunction with Theorem 2, immediately gives the following contrasting result concerning the Lasso and SLOPE. We use $\text{TPP}_{\text{Lasso}}(\boldsymbol{\beta}, \lambda)$ and $\text{FDP}_{\text{Lasso}}(\boldsymbol{\beta}, \lambda)$ to denote, respectively, the TPP and FDP of the Lasso with penalty parameter λ. Likewise, we use $\text{TPP}_{\text{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ and $\text{FDP}_{\text{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ to denote those of SLOPE as $\xi \to 0$.

COROLLARY 2.2 (SLOPE breaks the DT power limit). *In the supercritical regime, the following conclusions hold under the working assumptions:*

(a) *The power of the Lasso satisfies* $\text{TPP}_{\text{Lasso}}(\boldsymbol{\beta}, \lambda) < u^\star_{\text{DT}}$ *with probability tending to one.*
(b) *For any* $0 \le u < 1$*, there exists a SLOPE regularization prior* Λ *and a signal prior* Π *such that* $\text{TPP}_{\text{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) > u$ *with probability tending to one.*

For illustration, Figure 1 in the introduction reflects this distinction between SLOPE and the Lasso with $u^\star_{\text{DT}}(0.4, 0.7) = 0.4401$ in the right plot. Another illustration is the left plot of Figure 1 and Figure 4, which is vertically truncated at $u^\star_{\text{DT}}(0.3, 0.2) = 0.5676$. Note that SLOPE breaks the DT power limit, i.e. there are $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ pairs for which $u^\star_{\text{DT}} < \text{TPP}_{\text{SLOPE}} < 1$, while, as shown in the proof of Corollary 2.2, still preserving non-trivial FDP, i.e. $\text{FDP}_{\text{SLOPE}} < 1 - \epsilon$, where $1 - \epsilon$ would be the FDP associated with the trivial procedure that selects all predictors.

Corollary 2.2 highlights the benefit of using sorted ℓ_1 regularization over the less flexible ℓ_1 regularization in terms of power. As confirmed by Proposition 2.3 below, this sharp distinction persists no matter how large the effect sizes are and, therefore, it must be attributed to the flexibility of the SLOPE regularization sequence. As is well-known, the Lasso selects no more than n variables. Worse, a significant proportion of false variables are always interspersed on the Lasso path in the linear sparsity regime and, therefore, even though the Lasso can select up to $n > k$ variables, it would always miss a fraction of true variables, thereby imposing a limit on the power. SLOPE, like Lasso, has a significant proportion of false variables interspersed with discoveries. However, unlike Lasso, SLOPE does not bear the constraint that $\|\widehat{\boldsymbol{\beta}}\|_0 \le n$ owing to the flexibility of its regularization sequence. In fact, the corresponding constraint for SLOPE is that the number of *unique* non-zero entries is no more than n (Su and Candès, 2016). This flexibility allows SLOPE to have arbitrarily high power regardless of the regime that (δ, ϵ) belongs to.

[3] In the compressed sensing literature, $\epsilon^\star$ corresponds to the sparsity level where the Donoho–Tanner phase transition occurs (Donoho and Tanner, 2009b,a).

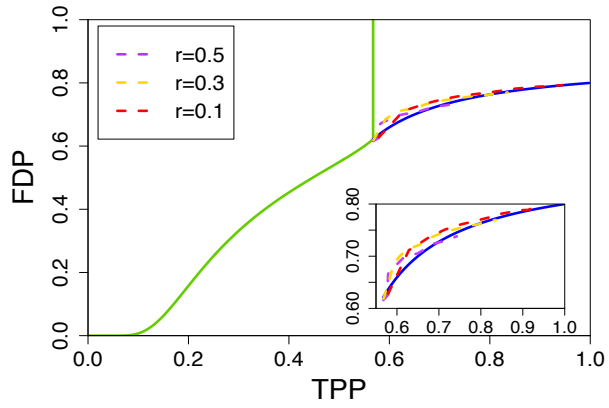

Fig 4: The Möbius part of the SLOPE trade-off upper bound $q^\star$. The solid curve denotes the upper bound specified by $(\delta, \epsilon) = (0.3, 0.2)$. The green line is the Lasso part of $q^\star$ and the blue line is the Möbius part. The numerical pairs of the TPP and FDP are obtained from experiments that are specified by the following parameters: $n = 300, p = 1000, \sigma^2 = 0$, signal prior $\Pi_M(\epsilon^\star/\epsilon)$ with $M = 10000$ in (2.6) (note that $\epsilon^\star(0.3) = 0.087$), and regularization prior $\boldsymbol{\lambda}_{\sqrt{M}, r\sqrt{M}, w}$ in (2.5) with varying w. Each pair is averaged over 50 independent trials.

Moving forward, we ask which regularization prior Λ and signal prior Π are "flexible" enough to enable SLOPE to break the DT power limit. To achieve desired flexibility, interestingly, it only requires a simple two-level regularization sequence for SLOPE. Consider the following *two-level* SLOPE regularization prior: given constants $a > b \geq 0$ and $0 < w < 1$, let $\Lambda_{a,b,w} = a$ with probability w and otherwise $\Lambda_{a,b,w} = b$. The SLOPE regularization sequence drawn from this prior takes the form

$$\boldsymbol{\lambda}_{a,b,w} := \big(\underbrace{a, a, \cdots, a}_{\text{around } wp}, \underbrace{b, b, \cdots, b}_{\text{around } (1-w)p} \big). \tag{2.5}$$

Next, for any $M > 0$ and $0 \leq \epsilon' \leq 1$, define the following signal prior:

$$\Pi_M(\epsilon') := \begin{cases} M, & \text{w.p.} \quad \epsilon\epsilon', \\ M^{-1}, & \text{w.p.} \quad \epsilon - \epsilon\epsilon', \\ 0, & \text{w.p.} \quad 1 - \epsilon. \end{cases} \tag{2.6}$$

Henceforth in this paper, denote by $\boldsymbol{\beta}_M(\epsilon')$ the regression coefficients sampled from $\Pi_M(\epsilon')$. In the following result, we take $M \to \infty$, rendering the nonzero entries of $\boldsymbol{\beta}_M(\epsilon')$ either very large or small.

Now we are ready to state the following result, which shows that SLOPE with the two-level regularization sequence can approach any point on the Möbius transformation (2.7) arbitrarily closely. This result also specifies the upper bound $q^\star$ in Theorem 2 in the supercritical regime:

$$q^\star(u; \delta, \epsilon) = \frac{\epsilon(1-\epsilon)u - \epsilon^\star(1-\epsilon)}{\epsilon(1-\epsilon^\star)u - \epsilon^\star(1-\epsilon)}, \tag{2.7}$$

for $u^\star_{\text{DT}} \leq u \leq 1$ (above the DT power limit). Note that this function takes the form of a *Möbius transformation*. Notably, taking $u = 1$ gives $q^\star(1; \delta, \epsilon) = \dfrac{(\epsilon - \epsilon^\star)(1-\epsilon)}{\epsilon(1-\epsilon^\star) - \epsilon^\star(1-\epsilon)} = 1 - \epsilon$, which is the FDP achieved by the trivial procedure that simply selects all predictors.

PROPOSITION 2.3. *For any $u^\star_{\mathrm{DT}} \le u \le 1$ in the supercritical regime, there exists $w \in (0,1)$ such that $\boldsymbol{\lambda}_{a,b,w}$ and $\boldsymbol{\beta}_M(\epsilon^\star/\epsilon)$ (defined via the prior in (2.6)) make SLOPE approach the point $(u, q^\star(u))$ in the sense that*

$$\lim_{M\to\infty} \lim_{\xi\to 0} \lim_{n,p\to\infty} (\mathrm{TPP}_\xi(\boldsymbol{\beta}_M(\epsilon^\star/\epsilon), \boldsymbol{\lambda}_{a,b,w}), \mathrm{FDP}_\xi(\boldsymbol{\beta}_M(\epsilon^\star/\epsilon), \boldsymbol{\lambda}_{a,b,w})) \to (u, q^\star(u)),$$

where $a = \sqrt{M}, b = r\sqrt{M}$ for a certain value $0 \le r \le 1$.

Figure 4 provides a numerical example that corroborates this proposition.

This result in fact implies Theorem 2 for $u^\star_{\mathrm{DT}} \le u \le 1$ in the supercritical regime. Note that the first limit $\lim_{n,p\to\infty}$ is taken in the sense of convergence in probability. See more details in its proof in Section 5.1. It is worthwhile to mention that the three-component mixture (2.6) is considered in Su et al. (2017) for the construction of favorable priors under sparsity constraint (see a generalization in Wang et al. (2020a)). This mixture prior is used to ensure that the effect sizes are either very strong or very weak. In particular, Proposition 2.3 remains true if M and $1/M$ are replaced by any value diverging to infinity and any value converging to 0, respectively.

2.3. *Below the Donoho–Tanner power limit.* Next, we continue to interpret Theorem 1 and Theorem 2, but with a focus on the regime below the DT power limit.

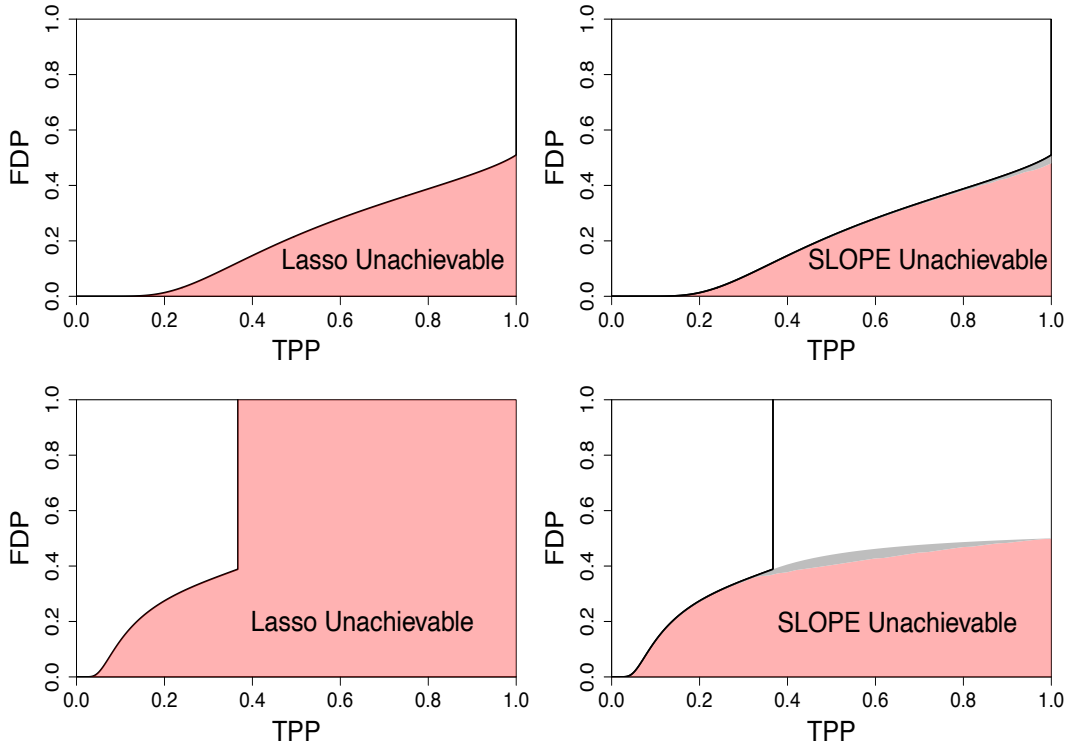

Fig 5: Examples of the TPP–FDP trade-off curve, with $(\delta, \epsilon) = (0.3, 0.2)$ on the top panel and $(0.3, 0.5)$ on the bottom. The left plot is the Lasso trade-off curve and the right plot describes the SLOPE trade-off gain. Neither the Lasso nor SLOPE can approach the red regions. The gray regions are sandwiched by the upper and lower bounds on the SLOPE trade-off.

First of all, the two right plots of Figure 5 show that the lower bound and the upper bound for q_{SLOPE} are very close to each other when $0 \le \mathrm{TPP} \le u^\star_{\mathrm{DT}}$ (recall that $u^\star_{\mathrm{DT}} = 1$ in the subcritical

regime). As a matter of fact, the upper bound in this regime is given by Su et al. (2017), which showed that, under the working assumptions, there exists a function $q^{\star}_{\text{Lasso}}(\cdot;\delta,\epsilon)$ such that

$$\text{FDP}_{\text{Lasso}}(\boldsymbol{\beta},\lambda) \geq q^{\star}_{\text{Lasso}}(\text{TPP}_{\text{Lasso}}(\boldsymbol{\beta},\lambda);\delta,\epsilon) - 0.0001,$$

holds with probability tending to one as $n,p \to \infty$. Here 0.0001 can be replaced by any arbitrarily small positive constant. Moreover, $q^{\star}_{\text{Lasso}}$ is tight in the sense that the Lasso can come arbitrarily close to any point on this curve by specifying a prior and a penalty parameter (see refined results in Wang et al. (2020a)). Recognizing that the Lasso is an instance of SLOPE, the tightness of $q^{\star}_{\text{Lasso}}$ allows us to set $q^{\star}(u) = q^{\star}_{\text{Lasso}}(u)$ for $0 \leq u \leq u^{\star}_{\text{DT}}$. To be more precise, letting $t^{\star}(u)$ be the largest positive root of the equation

$$\frac{2(1-\epsilon)\left[(1+x^2)\Phi(-x) - x\phi(x)\right] + \epsilon(1+x^2) - \delta}{\epsilon\left[(1+x^2)(1-2\Phi(-x)) + 2x\phi(x)\right]} = \frac{1-u}{1-2\Phi(-x)}, \tag{2.8}$$

we have

$$q^{\star}(u;\delta,\epsilon) = \begin{cases} q^{\star}_{\text{Lasso}}(u;\delta,\epsilon) = \dfrac{2(1-\epsilon)\Phi(-t^{\star}(u))}{2(1-\epsilon)\Phi(-t^{\star}(u)) + \epsilon u}, & \text{if } u \leq u^{\star}_{\text{DT}}(\delta,\epsilon), \\ \dfrac{\epsilon(1-\epsilon)u - \epsilon^{\star}(1-\epsilon)}{\epsilon(1-\epsilon^{\star})u - \epsilon^{\star}(1-\epsilon)}, & \text{if } u > u^{\star}_{\text{DT}}(\delta,\epsilon). \end{cases} \tag{2.9}$$

In the above expressions, $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the standard normal distribution, respectively.

Returning to the lower bound, in stark contrast, the situation becomes much more challenging. To be sure, to obtain a lower bound requires a good understanding of the superiority of sorted ℓ_1 regularization over its usual ℓ_1 counterpart. From a theoretical viewpoint, a major difficulty in the analysis of SLOPE arises from the *non-separability* of sorted ℓ_1 regularization. Note that the non-separability results from the sorting operation in the penalty term $\sum_{i=1}^{p} \lambda_i |b|_{(i)}$ in the SLOPE optimization program (1.2). To tackle this technical issue, in this paper we formulate the SLOPE trade-off as a calculus of variations problem and further cast it into infinite-dimensional convex optimization problems (see more details in Section 4).

In a nutshell, the flexibility of the SLOPE regularization sequence seems to only bring up limited improvement on the trade-off between the TPP and FDP below the DT power limit. However, the two right plots of Figure 5 present a noticeable departure between the two bounds when the TPP is slightly below $u^{\star}_{\text{DT}}$. This departure is not an artifact of our analysis. Indeed, in Section 5.3 we provide a problem instance whose asymptotic TPP and FDP trade-off falls strictly between the upper bound and the lower bound:

$$q_{\star}(u) + 0.0001 < \text{FDP} < q^{\star}(u) - 0.0001,$$

and $\text{TPP} \approx u < u^{\star}_{\text{DT}}$ with probability tending to one.

2.4. *On model selection and estimation.* An important but less-emphasized point is that the above-mentioned comparison between the two methods is over the *lower envelope* of all the instance-specific problems. In this regard, it would be too quick to conclude that the flexibility of the penalty sequence does not gain any benefits for SLOPE, even at points where $q_{\star}(u)$ may be very close to $q^{\star}_{\text{Lasso}}(u)$. Under the working hypotheses, indeed, we can formally prove that SLOPE is superior to the Lasso in the sense that we can always find a SLOPE regularization prior that strictly improves the Lasso on the same linear regression problem in terms of both model selection and estimation. Below, we let $\widehat{\boldsymbol{\beta}}$ denote the SLOPE or the Lasso estimate, and use the subscript to distinguish between the two methods.

THEOREM 3. *Under the working assumptions, namely (A1), (A2), and (A3), given any bounded signal prior* Π *and any Lasso regularization parameter* $\lambda > 0$*, there exists a SLOPE regularization* Λ *such that the following inequalities hold simultaneously with probability tending to one:*

(a) $\mathrm{TPP}_{\mathrm{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) > \mathrm{TPP}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda)$;
(b) $\mathrm{FDP}_{\mathrm{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) < \mathrm{FDP}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda)$;
(c) $\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \boldsymbol{\beta}\|^2 < \|\widehat{\boldsymbol{\beta}}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda) - \boldsymbol{\beta}\|^2$.

This theorem shows that SLOPE can outperform the Lasso from both the model selection and estimation viewpoints. We stress, however, that the result is *non-constructive* in that it does not provide the actual SLOPE penalty vector $\boldsymbol{\lambda}$ giving the good performance—it only claims that one *exists*. In practice, one would likely want to find a SLOPE sequence to optimize performance along one attribute only, depending on the goal (i.e., by considering model selection or estimation separately). The task of finding optimally performing SLOPE penalty sequences for any given fixed prior is an important open question, which we leave for future work.

The proof strategy of Theorem 3 leverages a simple form of SLOPE regularization sequences that admits two distinct values (see (2.5)). Due to space constraints, we relegate the proof of this theorem to Appendix A. It is somewhat surprising that such a simple two-level sequence can already exploit the benefits of using SLOPE over the Lasso.

As an aside, we remark that SLOPE has been shown to achieve the asymptotically exact minimax estimation when the sparsity level is much lower than considered in the present paper, largely owing to the adaptivity of sorted ℓ_1 regularization (Su and Candès, 2016). When it comes to the Lasso, however, cross validation is needed to select a penalty parameter that enables the Lasso to achieve similar estimation performance, which however is not exact as the constant is not sharp (Bellec et al., 2018).

3. Preliminaries for Proofs. In this section, we collect some preliminary results about SLOPE and AMP theory that allow us to get analytic expressions of the TPP and FDP asymptotically. Informally speaking, the AMP theory given in Bu et al. (2020, Theorem 3) characterizes the *asymptotic* joint distribution of the SLOPE estimator $\widehat{\boldsymbol{\beta}}$ and the true regression coefficients $\boldsymbol{\beta}$ (similar results are given in Hu and Lu (2019, Theorem 1) using the convex Gaussian minimax theory (CGMT) instead of AMP). Notably, since $\widehat{\boldsymbol{\beta}}$ depends on $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, when studying asymptotic properties of $\widehat{\boldsymbol{\beta}}$, we will work with their asymptotic distributions (Π, Λ). In this way, we drop the dependence on finite-sample quantities like n, p and the sparsity level $|\{j : \beta_j \neq 0\}|$ and instead work with asymptotic quantities such as (δ, ϵ) henceforth.

To be specific, under pseudo-Lipschitz functions (see Bu et al. (2020, Definition 3.1)) on $(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, the asymptotic distribution of the SLOPE (including the Lasso) estimator $\widehat{\boldsymbol{\beta}}$, which we denote as $\widehat{\Pi}$, can be described as

$$\widehat{\Pi} \overset{\mathcal{D}}{=} \eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z), \tag{3.1}$$

where Z is an independent standard normal and the superscript $\mathcal{D}$ means "in distribution". We will refer to η (to be introduced in (3.5)) as the *limiting scalar function* in Hu and Lu (2019), and (τ, A) is the unique solution to the *state evolution* and the *calibration* equations

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left(\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z) - \Pi\right)^2, \tag{3.2}$$

$$\Lambda \overset{\mathcal{D}}{=} \mathrm{A}\tau\left(1 - \frac{1}{\delta}\mathbb{E}\left(\eta'_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z)\right)\right). \tag{3.3}$$

In order to discuss properties of the limiting scalar function η, we first introduce the SLOPE proximal operator on $(\boldsymbol{y}, \boldsymbol{\theta}) \in \mathbb{R}^p \times \mathbb{R}^p$, where $\boldsymbol{\theta}$ is proportional to $\boldsymbol{\lambda}$ and $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p \geq 0$ with at least one inequality. We define the proximal operator as

$$\text{prox}_J(\boldsymbol{y}; \boldsymbol{\theta}) := \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{b}\|^2 + J_{\boldsymbol{\theta}}(\boldsymbol{b}) \right\}, \tag{3.4}$$

where $J_{\boldsymbol{\theta}}(\boldsymbol{b}) = \sum_{i=1}^p \theta_i |b|_{(i)}$. In the Lasso case when the penalty parameter is a constant, the proximal operator reduces to the soft-thresholding function:

$$\text{prox}_J(\boldsymbol{y}; \theta) = \eta_{\text{soft}}(\boldsymbol{y}; \theta) := \text{sign}(\boldsymbol{y}) \cdot \max\{|\boldsymbol{y}| - \theta, 0\}.$$

Generally speaking, the SLOPE proximal operator in (3.4) is *adaptive* and *non-separable*, in the sense that an element of the output generally will depend on all elements of the input. As a concrete example, we obtain via Algorithm 1 that the proximal operator for SLOPE is given by

$$\text{prox}_J([20, 13, 10, 6, 4]; [12, 10, 5, 5, 5]) = \eta_{\text{soft}}([20, 13, 10, 6, 4]; [12, 9, 6, 5, 5]) = [8, 4, 4, 1, 0].$$

On the one hand, the adaptivity arises from the fact that larger penalties are applied to larger elements of the input. On the other hand, for example, two elements of input $[13, 10]$ are not directly thresholded by the penalty $[10, 5]$, but rather an averaging step is triggered by the existence of the other inputs, which gives an effective threshold of $[9, 6]$. This is illustrated in Figure 6.

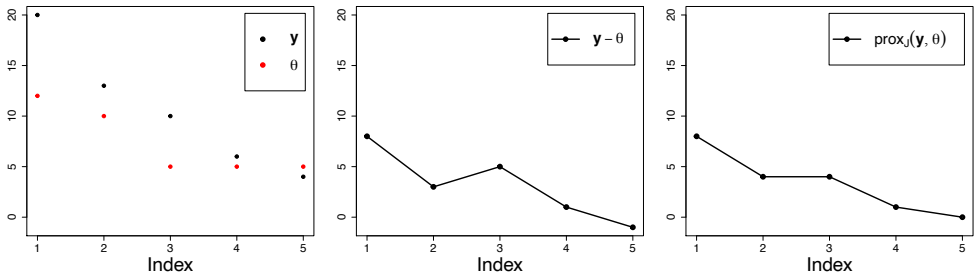

Fig 6: Illustration of how the SLOPE proximal operator can be interpreted as using an effective threshold. The leftmost figure plots two vectors $\boldsymbol{y}$ and $\boldsymbol{\theta}$. The middle image plots their difference $\boldsymbol{y} - \boldsymbol{\theta}$ and the rightmost image plots the output of the proximal operator $\text{prox}_J(\boldsymbol{y}; \boldsymbol{\theta})$.

Although the SLOPE proximal operator given in (3.4) is non-separable, nevertheless, as introduced in Hu and Lu (2019, Proposition 1), the SLOPE proximal operator is *asymptotically separable*: for sequences $\{\boldsymbol{\theta}(p)\}$ and $\{\boldsymbol{v}(p)\}$ growing in p with empirical distributions that weakly converge to distributions Θ and V, respectively, there exists a limiting scalar function η (determined by Θ and V) such that as $p \to \infty$,

$$\frac{1}{p} \|\text{prox}_J(\boldsymbol{v}(p); \boldsymbol{\theta}(p)) - \eta_{V,\Theta}(\boldsymbol{v}(p))\|^2 \to 0. \tag{3.5}$$

The work in Hu and Lu (2019) discusses many properties of this limiting scalar function, η. Indeed, it is shown to be odd, increasing, Lipschitz continuous with constant 1 and applied coordinate-wise to $\boldsymbol{v}(p)$ (hence it is separable; see Hu and Lu (2019, Proposition 2)). In more details, $\eta_{V,\Theta}(x)$ takes a scalar input, x, and performs soft-thresholding with a penalty

adaptive to x in a way that depends on V and Θ, meaning there is an input-dependent penalty $\lambda_{V,\Theta}(x)$ such that $\eta_{V,\Theta}(x) = \eta_{\text{soft}}(x; \lambda_{V,\Theta}(x))$. More details on the adaptive penalty function that relates the SLOPE proximal operator to the soft-thresholding function can be found in Appendix C.

We now discuss in more detail the so-called state evolution and calibration equations given in (3.2) and (3.3). We refer to A, which is defined implicitly via (3.3), as the *normalized penalty* distribution. Notice that A only differs from the original penalty distribution Λ by a constant factor. In fact, there exists a one-to-one mapping between A and Λ by Bu et al. (2020, Proposition 2.6), allowing one to analyze in either regime flexibly. Moreover, for a fixed Π, the quantity $\tau(\mathrm{A})$ can be uniquely derived from (3.2) and, as shown in Bu et al. (2020, Corollary 3.4), it can be used to characterize the estimation error via $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p \to \delta(\tau^2 - \sigma^2)$. In this work, we will use $\tau := \tau(\Pi, \Lambda)$ as a factor to define the *normalized prior*,

$$\pi(\Pi, \mathrm{A}) := \Pi/\tau(\Pi, \Lambda),$$

and, in particular, when it is clear from the context, we will use (Π, Λ) and (π, A) interchangeably since there exists a bijective calibration between the original problem instance and the normalized one provided by the fixed point recursion for the state evolution and the calibration mappings, (3.2) and (3.3). We refer the interested readers to Appendix B for a discussion of many nice properties of this fixed point recursion, such as the explicit form of the divergence η'.

Under the characterization of the asymptotic SLOPE distribution given in (3.1), we define $\mathrm{FDP}^{\infty}(\Pi, \Lambda)$ and $\mathrm{TPP}^{\infty}(\Pi, \Lambda)$ as the large system limits of FDP and TPP. The proof of convergence in probability is given in the next lemma. We will eventually let $\xi \to 0$, and in order for the FDP and TPP to converge, we consider FDP_ξ and TPP_ξ in (2.2) with ξ in the set

$$\Xi := \{\xi : \mathbb{P}(\widehat{\Pi}(\Pi, \Lambda) = \xi) = 0\}, \tag{3.6}$$

where $\widehat{\Pi}$ is the limiting distribution of $\widehat{\beta}_j$, defined in (3.1).

LEMMA 3.1. *Under the working assumptions, namely (A1), (A2), and (A3), for $\xi \in \Xi$ in* (3.6)*, the SLOPE estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ with the penalty sequence $\boldsymbol{\lambda}$ satisfies*

$$\mathrm{FDP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j = 0\}|}{|\{j : |\widehat{\beta}_j| > \xi\}|} \xrightarrow{P} \mathrm{FDP}^{\infty}_\xi(\Pi, \Lambda) := \frac{(1-\epsilon)\,\mathbb{P}\left(\left|\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\tau Z)\right| > \xi\right)}{\mathbb{P}\left(\left|\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z)\right| > \xi\right)},$$

$$\mathrm{TPP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j \neq 0\}|}{|\{j : \beta_j \neq 0\}|} \xrightarrow{P} \mathrm{TPP}^{\infty}_\xi(\Pi, \Lambda) := \mathbb{P}\left(\left|\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi^\star + \tau Z)\right| > \xi\right),$$

where superscript P denotes convergence in probability, Z is a standard normal independent of Π, and (τ, A) is the unique solution to the state evolution (3.2) *and calibration* (3.3)*. Furthermore, $\Pi^\star := (\Pi | \Pi \neq 0)$ is the signal prior distribution of the non-zero elements.*

By the continuity of the probability measure, we obtain

$$\begin{aligned} \lim_{\xi \to 0} \mathrm{FDP}^{\infty}_\xi(\Pi, \Lambda) &= \mathrm{FDP}^{\infty}(\Pi, \Lambda) := \frac{(1-\epsilon)\,\mathbb{P}\left(\eta_{\pi+Z, \mathrm{A}}(Z) \neq 0\right)}{\mathbb{P}\left(\eta_{\pi+Z, \mathrm{A}}(\pi + Z) \neq 0\right)}, \\ \lim_{\xi \to 0} \mathrm{TPP}^{\infty}_\xi(\Pi, \Lambda) &= \mathrm{TPP}^{\infty}(\Pi, \Lambda) := \mathbb{P}\left(\eta_{\pi+Z, \mathrm{A}}(\pi^\star + Z) \neq 0\right). \end{aligned} \tag{3.7}$$

Here, $\pi = \Pi/\tau$ is the normalized prior distribution and $\pi^\star := \Pi^\star/\tau$. We give the proof of Lemma 3.1 in Appendix D.1 by extending Bogdan et al. (2013, Theorem B.1).

Following the notions of FDP^∞ and TPP^∞ given in Lemma 3.1, we mathematically define the SLOPE trade-off curve as the envelope of all achievable SLOPE $(\text{TPP}^\infty, \text{FDP}^\infty)$ pairs:

$$q_{\text{SLOPE}}(u; \delta, \epsilon) := \inf_{(\Pi,\Lambda):\, \text{TPP}^\infty(\Pi,\Lambda)=u} \text{FDP}^\infty(\Pi, \Lambda).$$

To study the SLOPE trade-off, we will make use of a critical concept, the *zero-threshold* $\alpha(\Pi, \Lambda)$, which will be defined in Definition 4.1. Using the zero threshold, the limiting values in (3.7) can be simplified to

$$\begin{aligned} \text{TPP}^\infty(\Pi, \Lambda) &= \mathbb{P}(|\pi^\star + Z| > \alpha(\Pi, \Lambda)), \\ \text{FDP}^\infty(\Pi, \Lambda) &= \frac{2(1-\epsilon)\Phi(-\alpha(\Pi, \Lambda))}{2(1-\epsilon)\Phi(-\alpha(\Pi, \Lambda)) + \epsilon \cdot \text{TPP}^\infty(\Pi, \Lambda)}. \end{aligned} \tag{3.8}$$

Note from the equations above that for fixed $\text{TPP}^\infty = u$, the formula of FDP^∞ is decreasing in α. Therefore we consider the maximum of feasible zero-thresholds,

$$\alpha^\star(u) := \sup_{(\Pi,\Lambda):\, \text{TPP}^\infty=u} \alpha(\Pi, \Lambda),$$

in order to derive the minimum FDP^∞ on the SLOPE trade-off

$$q_{\text{SLOPE}}(u; \delta, \epsilon) := \frac{2(1-\epsilon)\Phi(-\alpha^\star(u))}{2(1-\epsilon)\Phi(-\alpha^\star(u)) + \epsilon u}. \tag{3.9}$$

4. Lower bound of SLOPE trade-off. The main purpose of this section is to provide a lower bound $q_\star$ on q_{SLOPE}. We accomplish this by (equivalently) giving an upper bound for $\alpha^\star(u)$ for *fixed* u, which we denote as $t_\star(u)$. As we shall see, in contrast to Lasso, our derivation for SLOPE requires non-standard tools from the calculus of variations and quadratic optimization programming. The optimization problem is a constrained one involving the SLOPE penalty and the probability density function of the normalized prior π as the decision variables, subject to the fixed $\text{TPP} = u$ and the monotonicity of the penalty.

To construct the upper bound $t_\star(u)$, we examine the state evolution (3.2), which gives

$$\tau^2 \geq \frac{1}{\delta}\,\mathbb{E}\left(\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z) - \Pi\right)^2 = \frac{\tau^2}{\delta}\,\mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) - \pi\right)^2.$$

Rearranging the above inequality yields the state evolution condition

$$E(\Pi, \Lambda) := \mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) - \pi\right)^2 \leq \delta. \tag{4.1}$$

Here the quantity $E(\Pi, \Lambda)$ can be viewed as the asymptotic mean squared error between the SLOPE estimator and the truth, scaled by $1/\tau^2$, since $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/p \to \tau^2 E(\Pi, \Lambda)$ in probability by Bu et al. (2020, Corollary 3.4).

Before we proceed, we first introduce an important (scalar) quantity that governs the sparsity, the TPP, and the FDP of the SLOPE estimator and will be used throughout the paper.

DEFINITION 4.1. *Let* (Π, Λ) *be a pair of prior and penalty distributions (or, equivalently, the normalized* (π, Λ)*) and suppose* $\alpha(\Pi, \Lambda)$ *is a positive number such that* $\eta_{\pi+Z,\mathrm{A}}(x) = 0$ *if and only if* $|x| \leq \alpha(\Pi, \Lambda)$*. Then we say that* $\alpha = \alpha(\Pi, \Lambda)$ *is the* zero-threshold.

Intuitively, the zero-threshold is a positive threshold, below which, the input is mapped to zero. Note that the necessary condition (4.1) sets the feasible domain of (π, A) pairs and thus prescribes limits to the zero-threshold α. In the Lasso case, the zero-threshold is indeed equivalent to the normalized penalty scalar A; but in SLOPE, it is a quantity derived from the normalized penalty distribution A in a highly nontrivial manner (see Proposition C.5 for details).

Next, we state another useful definition. Recall from Section 3 that the limiting scalar function η of SLOPE is separable and assigns a different penalty to different inputs. We therefore define the *effective penalty function* accordingly.

DEFINITION 4.2. *Given a normalized pair of prior and penalty (π, A), the effective penalty function $\widehat{\mathrm{A}}_{\text{eff}} : \mathbb{R} \to \mathbb{R}_+$ is a function such that*

$$\eta_{\text{soft}}(x; \widehat{\mathrm{A}}_{\text{eff}}(x)) = \eta_{\pi+Z,\mathrm{A}}(x).$$

It is not hard to show that $\widehat{\mathrm{A}}_{\text{eff}}$ is well-defined. In fact, given $\eta_{\pi+Z,\mathrm{A}}$, we can represent $\widehat{\mathrm{A}}_{\text{eff}}$ via the zero-threshold from Definition 4.1, namely,

$$\widehat{\mathrm{A}}_{\text{eff}}(x) = \begin{cases} x - \eta_{\pi+Z,\mathrm{A}}(x) & \text{if } x > \alpha(\pi, \mathrm{A}), \\ -x + \eta_{\pi+Z,\mathrm{A}}(x) & \text{if } x < -\alpha(\pi, \mathrm{A}), \\ \alpha(\pi, \mathrm{A}) & \text{if } |x| < \alpha(\pi, \mathrm{A}). \end{cases}$$

Equipped with this effective penalty function, we can rewrite the state evolution condition (4.1) as

$$F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}] := \mathbb{E}\left(\eta_{\text{soft}}(\pi + Z; \widehat{\mathrm{A}}_{\text{eff}}(\pi + Z)) - \pi\right)^2 \le \delta,$$

in which the functional objective F_α is defined on the effective penalty function $\widehat{\mathrm{A}}_{\text{eff}}$ as well as the probability density function of $\pi^\star$. Note here that $\pi^\star$ and π determine each other uniquely since $\pi^\star := \pi | \pi \neq 0$. We provide an explicit expression for $F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}]$ in (G.1).

Since the constraint (3.2) remains the same if π is replaced by $|\pi|$, we assume $\pi \ge 0$ without loss of generality. We minimize $F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}]$ over the functional space of $(\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star})$ through a relaxed variational problem:

$$\begin{aligned} &\min_{\mathrm{A}_{\text{eff}}, \rho \ge 0} \quad F_\alpha[\mathrm{A}_{\text{eff}}, \rho] \\ &\text{s.t.} \quad \mathrm{A}_{\text{eff}}(\alpha) \ge \alpha, \mathrm{A}'_{\text{eff}}(z) \ge 0 \text{ for all } z \ge \alpha, \\ &\int_0^\infty \rho(t)dt = 1, \int_0^\infty [\Phi(t - \alpha) + \Phi(-t - \alpha)]\rho(t)dt = u. \end{aligned} \tag{4.2}$$

Here the function A_{eff} is implicitly defined on $[\alpha, \infty)$ as $\mathrm{A}_{\text{eff}}(z) = \alpha$ for $0 \le z < \alpha$ and ρ is a probability measure defined on $[0, \infty)$. We remark that the constraints for A_{eff} in problem (4.2) are derived from the properties of $\widehat{\mathrm{A}}_{\text{eff}}$ in Appendix C, i.e. $\mathrm{A}'_{\text{eff}} \ge 0$ comes from Fact C.3 and the boundary condition $\mathrm{A}_{\text{eff}}(\alpha) \ge \alpha$ comes from Proposition C.5. Because some additional properties of $\widehat{\mathrm{A}}_{\text{eff}}$ may have been excluded in the relaxation, these constraints are only necessary and may not be sufficient. Therefore,

$$\min_{(\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star})} F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}] \ge \min_{(\mathrm{A}_{\text{eff}}, \rho)} F_\alpha[\mathrm{A}_{\text{eff}}, \rho],$$

with the inequality possibly being strict, provided the left optimization problem above is solved subject to (i) $\widehat{\mathrm{A}}_{\text{eff}}$ corresponds to the effective penalty in the limiting scalar function; and (ii) $p_{\pi^\star}$ is a probability density function such that $\mathrm{TPP}^\infty = \mathbb{P}(|\pi^\star + Z| > \alpha) = u$.

Leveraging the above relaxation (4.2), in order to lower bound q_{SLOPE} in (3.9), we can analogously define the maximum feasible zero-threshold $\alpha^\star(u)$ and upper bound it with $t_\star(u)$ as follows:

(4.3)
$$\alpha^\star(u) := \sup\left\{\alpha : \min_{(\Pi,\Lambda)} F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}] \le \delta\right\} \le t_\star(u) := \sup\left\{\alpha : \min_{(\mathrm{A}_{\text{eff}},\rho)} F_\alpha[\mathrm{A}_{\text{eff}}, \rho] \le \delta\right\}.$$

With these definitions in place, we are now in a position to describe the procedure to find the optimal prior and the optimal penalty in problem (4.2), given $\mathrm{TPP}^\infty = u$ and $\alpha(\Pi, \Lambda) = \alpha$.

4.1. *Optimal prior is three-point prior.* To solve problem (4.2), we must search over all possible distributions $\pi^\star$, which is generally infeasible. To overcome this obstacle, we use the concept of extreme points (i.e. points that do not lie on the line connecting any other two points of the same set) to show that the optimal $\pi^\star$ for problem (4.2) is a two-point distribution, having probability mass at only two non-negative (and possibly infinite) values (t_1, t_2). In doing so, we significantly reduce the search domain, from infinite dimensional to two-dimensional. Because π has an additional point mass at 0, the optimal prior π (that can achieve minimum FDP when accompanied with the properly chosen penalty) is a three-point prior taking values at $(0, t_1, t_2)$. We recall that the two-point $\pi^\star$ is consistent to the Lasso result in Su et al. (2017, Section 2.5), where the optimal $\pi^\star$ is the infinity-or-nothing distribution with $t_1 = 0^+, t_2 = \infty$.

To see that $\pi^\star$ admits a two-point form, suppose that $(\mathrm{A}^*_{\text{eff}}, \rho^*)$ is the global minimum of problem (4.2). Then clearly ρ^* is also the global minimum of the following linear problem (4.4) with linear constraints.

(4.4)
$$\begin{aligned} \min_{\rho \ge 0} \quad & F_\alpha[\mathrm{A}^*_{\text{eff}}, \rho] \\ \text{s.t.} \quad & \int_0^\infty \rho(t)dt = 1, \int_0^\infty [\Phi(t-\alpha) + \Phi(-t-\alpha)]\rho(t)dt = u. \end{aligned}$$

Intuitively, since there are two constraints, we need two parameters (which will be t_1, t_2) to characterize the minimum. We formalize this intuition in the next lemma (proved in Appendix G) and show that ρ^* indeed takes the form of a sum of two Dirac delta functions.

LEMMA 4.3. *If ρ^* is a global minimum of problem* (4.4)*, then*
$$\rho^*(t) = p_1\delta(t - t_1) + p_2\delta(t - t_2)$$
for some constants p_1, p_2, t_1, t_2, and $p_1 + p_2 = 1$, $p_1, p_2 \ge 0$.

The above specific form of the optimal ρ^* allows us to search over all (t_1, t_2), each pair of which uniquely corresponds to either a single-point prior $\rho(t; t_1, t_2) = \delta(t - t_1)$ if $t_1 = t_2$, or a two-point prior by

(4.5)
$$\begin{aligned} \rho(t; t_1, t_2) &= p_1\delta(t - t_1) + p_2\delta(t - t_2), \\ p_1(t_1, t_2) &= \frac{u - [\Phi(t_2 - \alpha) + \Phi(-t_2 - \alpha)]}{[\Phi(t_1 - \alpha) + \Phi(-t_1 - \alpha)] - [\Phi(t_2 - \alpha) + \Phi(-t_2 - \alpha)]}, \\ p_2(t_1, t_2) &= 1 - p_1(t_1, t_2), \end{aligned}$$

where the last two equations come from the constraints in problem (4.4).

In light of Lemma 4.3, each pair (t_1, t_2) forms a different instantiation of problem (4.2), which will be problem (4.6) and whose optimal penalty is denoted by $\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2)$ so as to be explicitly dependent on (t_1, t_2). Before we proceed to optimize the penalty $\mathrm{A}_{\text{eff}}(\cdot; t_1, t_2)$, we assure the skeptical reader that our procedure – doing a grid search on (t_1, t_2) and considering the minimal value of all programs (4.6) parameterized by (t_1, t_2) to be equivalent to the minimal value of problem (4.2) – is indeed a valid approach. This claim is theoretically grounded by noting that $F_\alpha[\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2), \rho(\cdot; t_1, t_2)]$ is continuous in (t_1, t_2). Continuity can be seen from a perturbation analysis of the optimal value in problem (4.6). In our case, the perturbation analysis is not hard since the constraint is independent of (t_1, t_2) and F_α depends on $\mathrm{A}^*_{\text{eff}}$ in a strongly-convex manner: a small perturbation in (t_1, t_2) only results in a small perturbation in $\mathrm{A}^*_{\text{eff}}$ and thus in $F_\alpha[\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2), \rho(\cdot; t_1, t_2)]$. We refer the curious reader to a line of perturbation analysis for such optimization problems in Bonnans and Shapiro (2013); Shapiro (1992); Bonnans and Shapiro (1998).

4.2. *Characterizing the optimal penalty analytically.* By Lemma 4.3, we reduce the multivariate non-convex problem (4.2) to a set of univariate convex problems (4.6) over A_{eff}. In this section, we describe the optimal penalty function $\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2)$, which is the solution to the problem below:

$$\begin{aligned} \min_{\mathrm{A}_{\text{eff}}} \quad & F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)] \\ \text{s.t.} \quad & \mathrm{A}_{\text{eff}}(\alpha) \geq \alpha, \quad \mathrm{A}'_{\text{eff}}(z) \geq 0 \text{ for all } z \geq \alpha. \end{aligned} \tag{4.6}$$

This is a quadratic problem with a non-holonomic constraint. To see this, we can expand the objective functional F_α from (G.1) and split it into a functional integral that involves A_{eff} and other terms which do not, i.e.

$$\begin{aligned} F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)] = \int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz + \epsilon p_1 t_1^2 \Big[\Phi(\alpha - t_1) - \Phi(-\alpha - t_1)\Big] \\ + \epsilon p_2 t_2^2 \Big[\Phi(\alpha - t_2) - \Phi(-\alpha - t_2)\Big]. \end{aligned}$$

This split changes our objective functional from $F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)]$ to the new functional $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ with

$$\begin{aligned} L(z, \mathrm{A}_{\text{eff}}) := {} & 2(1-\epsilon)(z - \mathrm{A}_{\text{eff}}(z))^2 \phi(z) \\ & + \epsilon p_1 \Big(\big(z - t_1 - \mathrm{A}_{\text{eff}}(z)\big)^2 \phi(z - t_1) + \big(-z - t_1 + \mathrm{A}_{\text{eff}}(z)\big)^2 \phi(-z - t_1) \Big) \\ & + \epsilon p_2 \Big(\big(z - t_2 - \mathrm{A}_{\text{eff}}(z)\big)^2 \phi(z - t_2) + \big(-z - t_2 + \mathrm{A}_{\text{eff}}(z)\big)^2 \phi(-z - t_2) \Big). \end{aligned} \tag{4.7}$$

We will numerically optimize the functional $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ together with the constraints in problem (4.6). In addition, although we cannot derive the analytic form of $\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2)$ from problem (4.6), we can still analytically characterize it at points z where the monotonicity constraint is non-binding (that is, when $\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2)$ is strictly increasing in a neighborhood of z), as shown in Appendix E.1.

4.3. *Searching over the optimal penalty numerically.* To solve the functional optimization problem (4.6), we approximate it by a discrete optimization problem via Euler's finite difference method. Specifically, we approximate the function $L(z, \mathrm{A}_{\text{eff}})$ (and hence F_α) on a

discretized uniform grid of z and solve the resulting quadratic programming problem with linear constraints.

To this end, we denote vectors $\boldsymbol{z} = [\alpha, \alpha + \Delta z, \alpha + 2\Delta z, \cdots, \alpha + m\Delta z]$ and $\mathbf{A} = [\mathrm{A}_{\text{eff}}(\alpha), \mathrm{A}_{\text{eff}}(\alpha + \Delta z), \cdots, \mathrm{A}_{\text{eff}}(\alpha + m\Delta z)]$ for some small Δz and large m. Then problem (4.6) is discretized into the convex quadratic program

$$\begin{aligned} \min_{\mathbf{A}_{\text{eff}}} \quad & \bar{F}_\alpha(\mathbf{A}; t_1, t_2) \\ \text{s.t.} \quad & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \mathbf{A} \geq \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \end{aligned} \tag{4.8}$$

in which the new objective $\bar{F}_\alpha(\mathbf{A}; t_1, t_2)$ (derived in (G.2) and also presented below) is the discretized objective of $F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)]$ from problem (4.6).

As $\Delta z \to 0$ and $m \to \infty$, problem (4.8) recovers problem (4.6) by well-known convergence theory for Euler's finite difference method. To simplify the exposition, we write the objective of problem (4.8) in matrix and vector notation as follows:

$$\mathbf{Q} = \mathrm{diag}\left(2(1-\epsilon)\phi(\boldsymbol{z}) + \epsilon \sum_{j=1,2} p_j \Big[\phi(\boldsymbol{z} - t_j) + \phi(-\boldsymbol{z} - t_j) \Big] \right),$$

$$\mathbf{d} = 2(1-\epsilon)\boldsymbol{z}\phi(\boldsymbol{z}) + \epsilon \sum_{j=1,2} p_j \Big[(\boldsymbol{z} - t_j)\phi(\boldsymbol{z} - t_j) + (\boldsymbol{z} + t_j)\phi(\boldsymbol{z} + t_j) \Big],$$

and observe that

$$\bar{F}_\alpha(\mathbf{A}; t_1, t_2) = (\mathbf{A}^\top \mathbf{Q} \mathbf{A} - 2\mathbf{A}^\top \mathbf{d})\Delta z + \epsilon \sum_{j=1,2} p_j t_j^2 \Big[\Phi(\alpha - t_j) - \Phi(-\alpha - t_j) \Big].$$

The discretized problem (4.8) is equivalent to a standard quadratic programming problem, whose objective is the discrete version of $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ in (4.7),

$$\begin{aligned} \min_{\mathbf{A}} \quad & \frac{1}{2}\mathbf{A}^\top \mathbf{Q} \mathbf{A} - \mathbf{A}^\top \mathbf{d} \\ \text{s.t.} \quad & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \mathbf{A} \geq \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \end{aligned} \tag{4.9}$$

4.4. *Solving the quadratic program.* Here we briefly discuss our numerical approach to solving the quadratic program (4.9). Generally speaking, quadratic programming problems do not admit closed-form solutions. However, they can be efficiently solved by classical numerical methods, including the interior point method (Dikin, 1967; Sra et al., 2012), active set method (Murty and Yu, 1988; Ferreau et al., 2014) and other dual methods (Goldfarb and Idnani, 1983; Frank and Wolfe, 1956). In this work, we use the dual method in Goldfarb and Idnani (1983), as implemented in the R library `quadprog`, to solve (4.9).

We remark that problem (4.9) is not the only way to discretize problem (4.6) and we now mention other approaches, which can result in better discretization accuracy. The discretization of problem (4.6) contains two parts: (i) a numerical integration to approximate the objective and (ii) a numerical differentiation to approximate the constraints.

When formulating the quadratic programming problem (4.9), we chose to apply the left endpoint rule to approximate the objective integral $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}})dz$ in (4.7) by $(\mathbf{A}^\top \mathbf{Q}\mathbf{A} - 2\mathbf{A}^\top \mathbf{d})\Delta z$, as well as the backward finite difference (with first-order accuracy) to describe the constraint $\mathrm{A}'_{\text{eff}}(z) \geq 0$. Alternatively, one can use different numerical quadratures to approximate the integral $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}})dz$ or use a change of variable to approximate a different integral. We can also apply different finite differences to discretize the monotonicity constraint in problem (4.6).

4.4.1. *Numerical integration to approximate the objective.* More specifically, for the approximation of the objective in problem (4.6), we can alternatively apply numerical quadratures such as the trapezoid rule, Simpson's rule, or Gauss-Laguerre quadrature (Salzer and Zucker, 1949) to improve the numerical integration for $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}})dz$. On the other hand, we may use a change of variable $z = \frac{x}{1-x} + \alpha$ to transform the integral $\int_\alpha^\infty L(z)dz$ over an infinite interval $[\alpha, \infty)$ to the integral $\int_0^1 L\left(\frac{x}{1-x} + \alpha\right)\frac{dx}{(1-x)^2}$ over a finite interval $[0, 1]$. This new integral can then be approximated by the same left endpoint rule (or other rules) but with different $\mathbf{Q}$ and $\mathbf{d}$.

4.4.2. *Numerical differentiation to approximate the constraints.* As for the monotonicity constraint $\mathrm{A}'_{\text{eff}}(z) \geq 0$, we may alternatively use other difference methods, e.g. the central difference, or higher-order accuracies. Doing so will result in a different matrix that left-multiplies $\mathbf{A}$ in the constraint of (4.9).

In conclusion, different numerical integration and differentiation schemes will lead to other formulations of the quadratic programming that are different from (4.9). We do not pursue these additional numerical aspects in the present work.

4.5. *Summary.* To summarize everything so far, the procedure of finding the lower bound $q_\star(u)$ involves the following steps: fixing $\text{TPP}^\infty = u$, we search over a line of zero-thresholds $\{\alpha\}$; for each α, we search over a two-dimensional finite grid of (t_1, t_2), each pair defining a standard quadratic programming problem (4.9); we then solve the quadratic problem and reject (t_1, t_2) if the minimal value of the equivalent problem (4.8) is larger than δ; if all (t_1, t_2) are rejected, then the current zero-threshold α is too large to be valid. We set the largest valid zero-threshold as $t_\star(u)$ in (4.3) and write the lower bound of the FDP^∞ as $q_\star(u) = \frac{2(1-\epsilon)\Phi(-t_\star(u))}{2(1-\epsilon)\Phi(-t_\star(u))+\epsilon u}$. Note that $q_\star(u) > 0$ for any possible $t_\star(u)$.

We finally mention that, in addition to minimizing FDP at a fixed TPP over all penalty-prior pairs, our quadratic programming approach also works when the prior Π is fixed. The fixed prior scenario has been extensively studied in Hu and Lu (2019), who optimize over the limiting scalar function η while we are optimizing over the penalty function A_{eff}. Our approach adds a new angle that can be algorithmically more efficient. We defer the details of the procedure to Appendix J.

4.6. *Differences between SLOPE and Lasso.* We end this section by discussing why deriving the SLOPE trade-off is fundamentally more complicated than the Lasso case. We highlight that the variational problem (4.2) is *non-convex*, even though it is convex with respect to each variable A_{eff} and ρ (i.e. it is bi-convex but non-convex). Generally speaking, approximate solutions to non-convex problems are not accompanied by theoretical guarantees, except for some special cases. Our bi-convex problem (4.2) cannot be solved by alternating descent, namely, fixing one variable, optimizing over the other and then alternating. Furthermore, our constraints only add another layer of complexity to the problem: in particular, the monotonicity constraint of A_{eff} is non-holonomic (i.e. the constraint $\mathrm{A}'_{\text{eff}} \geq 0$ does not depend explicitly on A_{eff}).

More precisely, the difficulty in directly solving the problem (4.2) is two-fold. The first difficulty lies in the search for the optimal penalty. For the Lasso case, the penalty distribution A and the penalty function $\widehat{\mathrm{A}}_{\text{eff}}$ are not adaptive to the input and hence they both equal the zero-threshold α. Therefore, we can perform a grid search on $\mathrm{A} \in \mathbb{R}$ and simply optimize over ρ. However, for SLOPE, the penalty $\widehat{\mathrm{A}}_{\text{eff}}$ is a *function* and hence it is intractable to search over the SLOPE penalty function space. The functional form of the penalty is the reason we must rely on the calculus of variations to study the associated optimization problem.

To demonstrate the second difficulty, we again consider the convex problem (4.4), which is over the probability density function ρ, assuming the optimal penalty $\mathrm{A}^*_{\text{eff}}$ has been obtained. In the Lasso case, it was shown in Su et al. (2017, Equation (C.2)) that the optimal $\pi^\star$ is the infinity-or-nothing distribution: $\mathbb{P}(\pi^\star = 0) = 1 - \epsilon'$ and $\mathbb{P}(\pi^\star = \infty) = \epsilon'$. In other words, given $\mathrm{A}^*_{\text{eff}}$, we can easily derive the optimal ρ. However, a key concavity result in Su et al. (2017, Lemma C.1), which holds for Lasso and determines the optimal $\pi^\star$, unfortunately breaks in SLOPE. Therefore, the optimal form of $\pi^\star$ is inaccessible for SLOPE with existing tools, even if the optimal penalty $\mathrm{A}^*_{\text{eff}}$ is known.

5. Upper bound of SLOPE trade-off. In this section, we rigorously analyze the SLOPE trade-off upper boundary curve $q^\star$ (defined in (2.9)). As stated in Theorem 2, $q^\star$ takes two forms: below the DT power limit, i.e. when $\text{TPP}^\infty < u^\star_{\text{DT}}$ for $u^\star_{\text{DT}}$ defined in (2.4), we have $q^\star = q^\star_{\text{Lasso}}$, and beyond the DT power limit, $q^\star$ is a Möbius curve.

We start by giving some intuition for why the domain of $q^\star$ is the entire interval $[0, 1]$, whereas, the Lasso trade-off curve is only defined on $[0, u^\star_{\text{DT}})$. Intuitively, SLOPE is capable of overcoming the DT power limit and achieving 100% TPP since it is possible for SLOPE estimators to select all p features, hence, by the definition of TPP (see Section 2.1), one can find a completely dense SLOPE estimator whose TPP is automatically 1. This is not true for the Lasso, since it can select *at most* n out of p features. The corresponding constraint for the SLOPE estimator follows from the AMP calibration in (3.3) (discussed in detail in Appendix B), namely it says that the number of *unique absolute values* in the entries of the SLOPE estimator is at most n out of p. However, this does not directly constrain the sparsity of SLOPE estimator, and thus it can still be dense. In other words, the SLOPE estimator always satisfies the following:

$$\text{the number of unique non-zero magnitude } |\widehat{\beta}_i| \text{ in } \widehat{\boldsymbol{\beta}}(p) \le n. \tag{5.1}$$

Notice that, in the Lasso sub-case, the above implies a direct sparsity constraint $|\{i : \widehat{\beta}_i \neq 0\}| \le n$ as just discussed, since all non-zero entries in Lasso have unique magnitudes. We also remark that the asymptotically (5.1) is a necessary and sufficient condition to satisfy the constraint (3.3).

With this intuition, we are prepared to prove Theorem 2 and show that $q^\star$ indeed serves as an upper bound of q_{SLOPE}. Following Proposition 2.3, we have the tightness of $q^\star$ when $u \ge u^\star_{\text{DT}}$. We will further discuss the proof of Proposition 2.3 in Section 5.1, but leave the full details for Appendix D.4. The tightness of $q^\star$ when $u < u^\star_{\text{DT}}$ follows from the existing tightness result on the Lasso trade-off (see Su et al. (2017, Section 2.5)), since the Lasso is a sub-case of SLOPE and $q^\star$ matches the Lasso trade-off curve for $u < u^\star_{\text{DT}}$. Hence, we have the corollary below.

COROLLARY 5.1. *For any $0 \le u \le 1$, there exists an $\epsilon' \in [0, \epsilon^\star/\epsilon]$, and values $r(u) \in [0, 1]$ and $w(u) \in [0, 1]$, both depending on u, such that the penalty $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\sqrt{M}, r(u)\sqrt{M}, w(u)}$ (defined in (2.5)) and the prior $\boldsymbol{\beta}_M(\epsilon')$ (defined in (2.6)) make SLOPE approach the point $(u, q^\star(u))$ in the sense*

$$\lim_{M\to\infty}\lim_{\xi\to 0}\lim_{n,p\to\infty}\left(\mathrm{TPP}_\xi(\boldsymbol{\beta}_M(\epsilon'),\boldsymbol{\lambda}),\,\mathrm{FDP}_\xi(\boldsymbol{\beta}_M(\epsilon'),\boldsymbol{\lambda})\right)\to(u,q^\star(u)).$$

Moreover, when $u<u^\star_{\mathrm{DT}}$, we can set $r(u)=1$, without specifying $w(u)$, and $\epsilon'=\epsilon'(u)$ will also depend on u. When $u\ge u^\star_{\mathrm{DT}}$, we fix $\epsilon'=\epsilon^\star/\epsilon$ and set $r(u)$ via (5.2) *and $w(u)$ via* (5.3) *below.*

An interesting aspect of this result is that there are two different strategies for attaining $q^\star(u)$, depending on whether $\mathrm{TPP}^\infty=u$ is above or below the DT power limit. In both cases, we use a two-level penalty $\boldsymbol{\lambda}_{\sqrt{M},r(u)\sqrt{M},w(u)}$ and a sparse prior (see (2.6)) with very small and very large non-zeros. However, when $\mathrm{TPP}^\infty=u<u^\star_{\mathrm{DT}}$, the strategy for attaining $q^\star(u)$ is to vary the proportion of strong signals (which equals $\epsilon\epsilon'$ and ϵ' varies with u), but when $u\ge u^\star_{\mathrm{DT}}$, sharpness in the Möbius part of $q^\star$ is attained by keeping the sequence of priors fixed and instead tuning the ratio between strong and weak penalties.

The sharpness result of Corollary 5.1 shows that over the entire domain, $q^\star$ is arbitrarily closely achievable, thus, $q^\star(u)$ must serve as the upper bound of the minimum FDP^∞, $q_{\mathrm{SLOPE}}(u)$, hence we have completed the proof of Theorem 2.

5.1. *Möbius upper bound is achievable.* In this section, we will sketch the proof of Proposition 2.3, which is used to prove Corollary 5.1 in the regime $u\ge u^\star_{\mathrm{DT}}$. To complement Proposition 2.3 and Corollary 5.1, for concreteness, we give a specific prior and penalty pair in (2.7) that approaches $q^\star(u)$ when $u\ge u^\star_{\mathrm{DT}}$. The fully rigorous proof of Proposition 2.3, together with the derivation of (r,w), is given in Appendix D.4.

Before we sketch the proof, we will provide some intuition for what makes the specific choice of priors and penalties behave effectively in terms of reducing the FDP^∞ while still driving TPP^∞ to 1, in order that we are able to approach $q^\star(u)$ for all $u\ge u^\star_{\mathrm{DT}}$. We remind the reader that, because there is a one-to-one correspondence between original instance (Π,Λ) and the normalized (π,A), we will use the two notations interchangeably.

First, for fixed $\mathrm{TPP}^\infty=u$, we can reduce the FDP^∞ through a smart use of the priors defined in (2.6), where many elements equal 0 exactly, while some non-zero elements are small (equal to $1/M$) and others large (equal to M) with M tending to ∞. This is the same strategy as was used for demonstrating the achievability of the Lasso curve in Su et al. (2017), and the intuition that we present here is based on this analysis. At a high level, extremely strong signals are unlikely to be missed, and thus the TPP^∞ can be high at the cost of rendering the constraint (4.1) tight. On the other hand, weak signals help reduce the FDP because they are not counted toward the number of false positives and have little influence on (4.1). Mathematically speaking, for the Lasso, Su et al. (2017, Lemma C.1) revealed a concave relationship in Π between the normalized estimation error $E(\Pi,\Lambda)=\mathbb{E}(\eta_{\pi+Z,\mathrm{A}}(\pi+Z)-\pi)^2$ in (4.1) and the sparsity $\mathbb{P}(\eta_{\pi+Z,\mathrm{A}}(\pi+Z)\neq 0)$, which also depends on the pair (Π,Λ). We remind the reader that, because there is a one-to-one correspondence between original instance (Π,Λ) and the normalized (π,A), we will use the two notations interchangeably. The idea is that minimizing FDP^∞ corresponds to minimizing the sparsity (this can be seen, for example, by the relationship in (5.11) where $\kappa(\Pi,\Lambda)$ denotes the sparsity). Therefore, to find a prior Π that satisfies the state evolution condition (4.1), while minimizing the sparsity, the optimal (normalized) distribution for the non-zero elements, $\pi^\star$, for the Lasso case has probability masses concentrated at the endpoints of the domain, namely 0^+ and ∞. In this way, the form of the signal prior Π contributes to reducing the FDP^∞ by mixing the weak effects β_i with the zero effects.

Combining the priors discussed above, with a special subset of the possible penalties, namely the two-level penalties defined in (2.5), we are able to reduce the FDP$^\infty$ while still increasing the TPP$^\infty$ to its maximum value of 1, hence attaining $q^\star(u)$ for all $u \geq u^\star_{\text{DT}}$. Interestingly, the fact that SLOPE can do this, is through its penalty, which mixes the weak predictors $\widehat{\beta}_i$ and the zero predictors (see Figure 4). This mix-up is in fact triggered by the averaging step in the SLOPE proximal operator (see Algorithm 1; the averaging is determined by the sorted ℓ_1 norm in the SLOPE problem), which creates non-zero magnitudes that are shared by some predictors and hence maintains the quota of unique magnitudes in (5.1). As a consequence, the SLOPE estimator can overcome the DT power limit (and reach higher TPP$^\infty$) without violating the uniqueness constraint (5.1) on its magnitudes.

When constructing the two-level penalties just discussed, we must choose a pair (r, w) that, respectively, defines the downweighting of the $\sqrt{M}$ used for the smaller penalty and the proportion of penalties getting each value. Concretely speaking, in Proposition 2.3 and Corollary 5.1, we set

$$r(u) = \Phi^{-1}\left(\frac{2\epsilon - \epsilon^\star - \epsilon u}{2(\epsilon - \epsilon^\star)}\right) / t^\star(u^\star_{\text{DT}}). \tag{5.2}$$

where $\epsilon^\star$ and $u^\star_{\text{DT}}$ define the DT power limit and are given in (2.3)-(2.4) and $t^\star$ is defined in (2.8). Moreover,

$$w(u) = \epsilon^\star + \frac{2(1-\epsilon^\star)}{1-r}\left[\Phi(-t^\star(u^\star_{\text{DT}})) - r\Phi(-rt^\star(u^\star_{\text{DT}})) - \frac{\phi(-t^\star(u^\star_{\text{DT}})) - \phi(-rt^\star(u^\star_{\text{DT}}))}{t^\star(u^\star_{\text{DT}})}\right], \tag{5.3}$$

where r in the above is shorthand for the $r(u)$ from (5.2).

Without going into details, the key reason for choosing such pair (r, w) is so that the sequence of two-level penalties have two different penalization effects: for one, the SLOPE estimator $\eta_{\pi+Z,A}(\pi + Z)$ is equivalent to a Lasso estimator $\eta_{\text{soft}}(\pi + Z; t^\star(u^\star_{\text{DT}}))$ in the sense of (5.4); for the other, the SLOPE estimator is equivalent to a different Lasso estimator $\eta_{\text{soft}}(\pi + Z; rt^\star(u^\star_{\text{DT}}))$ in the sense of (5.5).

To be precise, it can be shown that

$$\eta_{\pi+Z,\text{A}}(\pi + Z) \overset{P}{=} \eta_{\text{soft}}(\pi + Z; t^\star(u^\star_{\text{DT}})),$$

and

$$\mathbb{E}(\eta_{\pi+Z,\text{A}}(\pi + Z) - \pi)^2 = \mathbb{E}(\eta_{\text{soft}}(\pi + Z; t^\star(u^\star_{\text{DT}})) - \pi)^2, \tag{5.4}$$

so when considering the asymptotic magnitude of the elements of the SLOPE estimator, or its asymptotic estimation error (4.1), we can analyze the limiting scalar function instead using a soft-thresholding function with threshold given by $t^\star(u^\star_{\text{DT}})$. Moreover, this implies that SLOPE satisfies the state evolution constraint (4.1) in a similar way to how the Lasso satisfies its corresponding state evolution constraint.

However, analysis of the asymptotic sparsity of the SLOPE estimator or of its asymptotic TPP and FDP, relies on the fact that one can prove

$$\mathbb{P}(\eta_{\pi+Z,\text{A}}(\pi + Z) \neq 0) = \mathbb{P}(\eta_{\text{soft}}(\pi + Z; rt^\star(u^\star_{\text{DT}})) \neq 0), \tag{5.5}$$

Hence, again, instead of analyzing the limiting scalar function one can analyze a soft-thresholding function, but now with a smaller threshold given by $rt^\star(u^\star_{\text{DT}})$ for some $0 \leq r \leq 1$ defined in (5.2). Reducing the threshold in this way functions to improve the attainable TPP–FDP over the comparable Lasso problem by allowing more elements in the estimate with non-zero values. We visualize the above claims in Figure 4(d).

Essentially, the state evolution condition (4.1) must always hold, but it uses the larger pseudo zero-threshold $t^\star(u^\star_{\text{DT}})$, while inference is conducted on the true, but smaller, zero-threshold $rt^\star(u^\star_{\text{DT}})$. In this way, we can extend attainability of $q^\star_{\text{Lasso}}$ to attainability $q^\star$, while still working within the state evolution constraint (4.1).

5.2. *Infinity-or-nothing prior has FDP above upper bound.* The goal of this section is to provide some intuition for the Möbius form of the curve $q^\star(u)$ when u is larger than the DT power limit. This will be done by demonstrating that, in the case of infinity-or-nothing priors, with a special subset of penalties, the SLOPE FDP$^\infty$ is always above $q^\star$ in Proposition 5.2. This also motivates the achievability results of Section 5.1, as the proof given in Section 5.1 essentially tries to construct prior penalty pairs such that the inequality in Proposition 5.2 becomes an equality. While we only consider infinity-or-nothing priors here, we remark that in the Lasso case these are actually the *optimal* priors (ses Su et al. (2017, Section 2.5)), meaning that they achieve the minimum FDP$^\infty$ given TPP$^\infty$.

PROPOSITION 5.2. *Under the working assumptions, namely (A1), (A2), and (A3), for $\xi \in \Xi$ in* (3.6)*, assuming that $\boldsymbol{\beta}$ is sampled i.i.d. from* (2.6) *for any $\epsilon' \in [0,1]$, $M \to \infty$, and that $\boldsymbol{\lambda}$ is the order statistics of i.i.d. realization of a non-negative Λ with $\mathbb{P}(\Lambda = \max \Lambda) \geq \epsilon\epsilon'$, the following inequality holds with probability tending to one:*

$$\text{FDP}_\xi(\boldsymbol{\beta}_M(\epsilon'), \boldsymbol{\lambda}) \geq q^\star\left(\text{TPP}_\xi(\boldsymbol{\beta}_M(\epsilon'), \boldsymbol{\lambda}); \delta, \epsilon\right) - c_\xi(\Pi_M(\epsilon'), \Lambda),$$

for some positive constant c_ξ which tends to 0 as $\xi \to 0$.

PROOF OF PROPOSITION 5.2. As in Section 4, we assume $\pi \geq 0$ without loss of generality since the analysis holds if we replace π by $|\pi|$. Consider a *subset of priors*, namely the infinity-or-nothing priors: for some $\epsilon' \in [0,1]$,

$$\pi_\infty(\epsilon') = \begin{cases} \infty & \text{w.p. } \epsilon\epsilon', \\ 0 & \text{w.p. } 1 - \epsilon\epsilon'. \end{cases} \tag{5.6}$$

Although the infinity-or-nothing prior in (5.6) does not satisfy the assumption (A2) that $\mathbb{P}(\Pi \neq 0) = \mathbb{P}(\pi \neq 0) = \epsilon$, this does not affect our discussion[4].

In fact, as demonstrated by Lemma 5.3 below, for infinity-or-nothing priors, the state evolution constraint (4.1) guarantees that $\epsilon' \leq \epsilon^\star/\epsilon$. Since $\epsilon^\star$ is the same for the Lasso and SLOPE, this means that the maximum proportion of ∞ signals in the infinity-or-nothing prior is the same for both as well.

LEMMA 5.3. *Under assumptions in Proposition 5.2, we must have $\epsilon' \in [0, \epsilon^\star/\epsilon]$.*

The proof of Lemma 5.3 is given in Appendix D.3. It turns out that the DT threshold $\epsilon^\star$ plays an important role in understanding the relationship between the sparsity and TPP$^\infty$. Before illustrating this relationship, we introduce the concept of *sparsity*. In a finite dimension, the sparsity of SLOPE estimator is $|\{j : \widehat{\beta}_j \neq 0\}|$. However, as $p \to \infty$, the count of non-zeros will also go to infinity, meaning a quantity like $\lim_p |\{j : \widehat{\beta}_j \neq 0\}|$ is not well-defined. Therefore we introduce the *asymptotic sparsity* of the SLOPE estimator via the distributional characterization in (3.1), denoting the limit in probability by plim,

$$\kappa(\Pi, \Lambda) := \mathbb{P}\left(\eta_{\pi+Z,\text{A}}(\pi + Z) \neq 0\right) = \mathbb{P}\left(\widehat{\Pi} \neq 0\right) = \text{plim}\, |\{j : \widehat{\beta}_j \neq 0\}|/p. \tag{5.7}$$

[4]The infinity-or-nothing prior can be approximated arbitrarily closely by a sequence of priors that satisfy the assumption. For example, let $M \to \infty$ and consider $\pi_M(\epsilon')$ defined in (2.6).

Making use of the DT threshold $\epsilon^\star(\delta)$, we show in Lemma 5.4 that the sparsity $\kappa(\Pi,\Lambda)$ sets an upper bound on achievable TPP^∞.

LEMMA 5.4. *Consider SLOPE based on the pair* (Π,Λ) *with* Π *from* (2.6) *and set* $M\to\infty$. *Then with the asymptotic sparsity* $0\le\kappa(\Pi,\Lambda)\le 1$[5], *we have* $\text{TPP}^\infty(\Pi,\Lambda)\le u^\star(\kappa(\Pi,\Lambda);\epsilon,\delta)$ *where*

$$u^\star(\kappa;\epsilon,\delta):=\begin{cases}1-\frac{(1-\kappa)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}, & \text{if } \delta<1 \text{ and } \epsilon>\epsilon^\star(\delta),\\ 1, & \text{otherwise.}\end{cases} \tag{5.8}$$

PROOF OF LEMMA 5.4. We will only prove $\text{TPP}^\infty(\Pi,\Lambda)\le 1-\frac{(1-\kappa)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}$ when $\delta<1$ and $\epsilon>\epsilon^\star(\delta)$. We note that the bound on $u^\star$ given in (5.8) when $\delta\ge 1$ or $\epsilon\le\epsilon^\star(\delta)$ is trivial since, by definition, $\text{TPP}^\infty(\Pi,\Lambda)\le 1$.

As $M\to\infty$ in (2.6), the prior π converges to the infinity-or-nothing priors $\pi_\infty(\epsilon')$ in (5.6). In addition, $\pi^\star=\pi_\infty(\epsilon'/\epsilon)$. By the intermediate value theorem, there must exist some $\epsilon'\in[0,1]$ such that

$$\begin{aligned}\text{TPP}^\infty(\Pi,\Lambda)=\mathbb{P}(|\pi^\star+Z|>\alpha)&=(1-\epsilon')\,\mathbb{P}(|Z|>\alpha)+\epsilon'\,\mathbb{P}(|\infty+Z|>\alpha)\\&=2(1-\epsilon')\Phi(-\alpha)+\epsilon'.\end{aligned}$$

Here the first equality is given by (3.8) and $\alpha\equiv\alpha(\Pi,\Lambda)$ is the zero-threshold in Definition 4.1. The second equality follows from substituting the infinity-or-nothing $\pi^\star$. Therefore, the asymptotic sparsity in (5.7) is

$$\kappa(\Pi,\Lambda)=\mathbb{P}(|\pi+Z|>\alpha)=(1-\epsilon)\,\mathbb{P}(|Z|>\alpha)+\epsilon\,\text{TPP}^\infty=2(1-\epsilon\epsilon')\Phi(-\alpha)+\epsilon\epsilon',$$

where the first equality follows by the definition of the zero-threshold in Definition 4.1, the second uses that $\text{TPP}^\infty(\Pi,\Lambda)=\mathbb{P}(|\pi^\star+Z|>\alpha)$, and the third is the result from the previous equation.

Some rearrangement gives

$$\Phi(-\alpha)=\frac{\kappa(\Pi,\Lambda)-\epsilon\epsilon'}{2(1-\epsilon\epsilon')},\quad\text{and}\quad\text{TPP}^\infty(\Pi,\Lambda)=\frac{(1-\epsilon')(\kappa(\Pi,\Lambda)-\epsilon\epsilon')}{1-\epsilon\epsilon'}+\epsilon'. \tag{5.9}$$

Simple calculus shows that the $\text{TPP}^\infty(\Pi,\Lambda)$ in (5.9) is an increasing function of ϵ'. To see this, notice that the derivative is $\frac{(1-\epsilon)(1-\kappa)}{(1-\epsilon\epsilon')^2}\ge 0$. Given that $\epsilon'\le\epsilon^\star/\epsilon$ by Lemma 5.3, we have

$$\text{TPP}^\infty(\Pi,\Lambda)\le\frac{(1-\frac{\epsilon^\star}{\epsilon})(\kappa(\Pi,\Lambda)-\epsilon\cdot\frac{\epsilon^\star}{\epsilon})}{1-\epsilon\cdot\frac{\epsilon^\star}{\epsilon}}+\frac{\epsilon^\star}{\epsilon}=1-\frac{(1-\kappa)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}.$$

□

In fact, Lemma 5.4 is an extension of Su et al. (2017, Lemma C.2) (restated in Corollary 2.2(a)), which claims that, in the Lasso case, for all priors including those are not infinity-or-nothing, $\text{TPP}^\infty\le u^\star(\delta;\epsilon,\delta)$. In particular, we remark that $u^\star(\delta;\epsilon,\delta)$ is equivalent to $u^\star_{\text{DT}}(\delta,\epsilon)$, since any Lasso estimator has an asymptotic sparsity no larger than δ.

As an immediate consequence of Lemma 5.4, we can reversely set a lower bound on the sparsity $\kappa(\Pi,\Lambda)$ given $\text{TPP}^\infty(\Pi,\Lambda)$. This is achieved by inverting the mapping in (5.8) and setting $u^\star=\text{TPP}^\infty$:

$$\kappa(\Pi,\Lambda)\ge 1-\frac{\epsilon(1-\text{TPP}^\infty(\Pi,\Lambda))(1-\epsilon^\star)}{\epsilon-\epsilon^\star}. \tag{5.10}$$

[5]To distinguish from the Lasso, we note that SLOPE can reach $\kappa=1$ and thus gives a dense solution whose TPP is 1.

Finally, leveraging the lower bound on the sparsity, we can minimize the FDP^∞ by minimizing the sparsity $\kappa(\Pi, \Lambda)$, since by definition

$$\mathrm{FDP}^\infty(\Pi, \Lambda) = 1 - \frac{\epsilon \cdot \mathrm{TPP}^\infty(\Pi, \Lambda)}{\kappa(\Pi, \Lambda)}. \tag{5.11}$$

Plugging (5.10) into (5.11), we finish the proof that $\mathrm{FDP}^\infty \geq q^\star(\mathrm{TPP}^\infty)$ for the SLOPE when we restrict the priors to be infinity-or-nothing: with $\mathrm{TPP}^\infty = u$,

$$\mathrm{FDP}^\infty(\Pi, \Lambda) \geq q^\star(u; \delta, \epsilon) := 1 - \frac{\epsilon u}{1 - \frac{\epsilon(1-u)(1-\epsilon^\star)}{\epsilon - \epsilon^\star}} = \frac{\epsilon u(1-\epsilon) - \epsilon^\star(1-\epsilon)}{\epsilon u(1-\epsilon^\star) - \epsilon^\star(1-\epsilon)}.$$

□

5.3. *Gap between upper and lower bounds.* Considering Figure 2, we observe that the upper and lower boundary curves, $q_\star$ and $q^\star$, can be visually and numerically close to each other, especially when $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}$. One may wonder whether these boundaries actually coincide below the DT power limit. We answer this question in the negative and show analytically that there may exist pairs of $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ with the FDP^∞ strictly below $q^\star(\mathrm{TPP}^\infty)$ when $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}$. In other words, there are instances where $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ points lie between the boundary curves $q_\star$ and $q^\star$.

PROPOSITION 5.5. *For some* (δ, ϵ)*, there exists* $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}(\delta, \epsilon)$ *defined in* (2.4) *such that*

$$q_\star(\mathrm{TPP}^\infty) < \mathrm{FDP}^\infty < q^\star(\mathrm{TPP}^\infty).$$

In the following, we prove Proposition 5.5 by constructing a specific problem instance (Π, Λ) which has FDP^∞ falling between the bounds. By showing that the gap between $q^\star(u)$ and $q_\star(u)$ indeed exists, we rigorously demonstrate a gap between $q^\star(u)$ and the unknown SLOPE trade-off q_{SLOPE}.

We note that, for the Lasso trade-off at $(u, q^\star(u))$, the zero-threshold $\alpha(\Pi, \lambda) = t^\star(u)$ (defined in (2.8)) exactly and the state evolution constraint (4.1) is binding, i.e. $E(\Pi, \lambda) = \delta$ (see Su et al. (2017, Lemma C.4, Lemma C.5)).

Fixing $\mathrm{TPP}^\infty = u$, our strategy (detailed in Appendix E) is to construct (π, A) for SLOPE such that $\alpha(\pi, \mathrm{A}) = t^\star(u)$ as well but the state evolution constraint (4.1) is not binding, i.e. $E(\Pi, \Lambda) < \delta$. If such a construction succeeds, we can use a strictly larger zero-threshold than $t^\star(u)$ that can increase until $E(\Pi, \Lambda) > \delta$. Then, by using a larger zero-threshold, the SLOPE FDP^∞ is guaranteed to be strictly smaller than $q^\star(\mathrm{TPP}^\infty)$ by (3.8). Thus we will complete the proof that $q_\star(u) < q^\star(u)$ for some $u < u^\star_{\mathrm{DT}}$.

To construct (π, A) satisfying $\alpha(\pi, \mathrm{A}) = t^\star(u)$ with $E(\Pi, \Lambda) < \delta$, we leverage our empirical observation that the optimal priors $\pi^\star$, in the sense of problem (4.2), which achieves the lower bound $q_\star$, are oftentimes either infinity-or-nothing or constant. This motivates us to consider constant priors $\pi^\star = t_1$, for some constant t_1 (i.e. $p_1 = 1, t_1 = t_2$ in (4.5)), and hence

$$\pi = \begin{cases} t_1 & \text{w.p. } \epsilon, \\ 0 & \text{w.p. } 1 - \epsilon. \end{cases}$$

In fact, conditioning on $\alpha(\Pi, \Lambda) = t^\star$ and $\mathrm{TPP}^\infty = u$, the constant $t_1(u)$ is uniquely determined by (3.8):

$$\mathbb{P}\left(|t_1 + Z| > t^\star(u)\right) = u,$$

where Z is a standard normal.

Next, we use a common tool in the calculus of variations, known as the Euler-Lagrange equation (detailed in Appendix E.2), to construct an effective penalty function $\mathrm{A}_{\text{eff}}(z)$ analytically on the interval $[0, \infty)$. The explicit form of $\mathrm{A}_{\text{eff}}(z)$ is defined in (E.1) with $\alpha = t^\star$. We emphasize that the constructed A_{eff} may not be a feasible SLOPE penalty function in the sense that it may violate the constraints in problem (4.6); however, if A_{eff} is increasing, then the optimal SLOPE effective penalty must be A_{eff}, as it is the minimizer of the unconstrained version of problem (4.6) and clearly satisfies the constraints. In the case that A_{eff} is feasible, we compare $E(\Pi, \Lambda) = F_{t^\star(u)}[\mathrm{A}_{\text{eff}}, p_{t_1}]$ with δ to determine whether $q^\star(u) > q_\star(u)$.

We now give a concrete example, which is elaborated in Appendix E.3. When $\delta = 0.3, \epsilon = 0.2, \Pi^\star = 4.9006, \text{TPP}^\infty = u^\star_{\text{DT}} = 0.5676$, the maximum Lasso zero-threshold $t^\star(u^\star_{\text{DT}}) = 1.1924$ and the minimum Lasso $\text{FDP}^\infty = 0.6216$. We can construct the SLOPE penalty A_{eff} that has the same zero-threshold and achieves $E(\Pi, \Lambda) = 0.2773 < \delta$. We can further construct the SLOPE penalty with larger zero-threshold, up to 1.2567, eventually have the SLOPE $\text{FDP}^\infty = 0.5954$, which is much smaller than the minimum Lasso FDP^∞. In fact, our method can construct SLOPE penalty that outperforms the Lasso trade-off for any $\text{TPP}^\infty \in (0.5283, 1]$, as shown in Figure 8.

6. Discussion. In this paper, we have investigated the possible advantages of employing sorted ℓ_1 regularization in model selection instead of the usual ℓ_1 regularization. Focusing on SLOPE, which instantiates sorted ℓ_1 regularization, our main results are presented by lower and upper bounds on the trade-off between false and true positive rates. On the one hand, the two tight bounds together demonstrate that type I and type II errors cannot both be small simultaneously using the SLOPE method with any regularization sequences, no matter how large the effect sizes are. This is the same situation as the Lasso (Su et al., 2017), which instantiates ℓ_1 regularization. More importantly, our results on the other hand highlight several benefits of using sorted ℓ_1 regularization. First, SLOPE is shown to be capable of achieving arbitrarily high power, thereby breaking the DT power limit. For comparison, the Lasso cannot pass the DT power limit in the supercritical regime, no matter how strong the effect sizes are. Second, moving to the regime below the DT power limit, we provide a problem instance where the SLOPE TPP and FDP trade-off is strictly better than the Lasso. Third, we introduce a comparison theorem which shows that any solution along the Lasso path can be dominated by a certain SLOPE estimate in terms of both the TPP and FDP and the estimation risk. In other words, the flexibility of sorted ℓ_1 regularization can always improve on the usual ℓ_1 regularization in the instance-specific setting.

The assumptions underlying the above-mentioned results include the random designs that have independent Gaussian entries and linear sparsity. In the venerable literature on high-dimensional regression, however, a more common sparsity regime is sublinear regimes where k/p tends to zero. As such, it is crucial to keep in mind the distinction in the sparsity regime when interpreting the results in this paper. From a technical viewpoint, our assumptions here enable the use of tools from AMP theory and in particular a very recent technique for tackling non-separable penalties. To obtain the lower bound, moreover, we have introduced several novel elements that might be useful in establishing trade-offs for estimators using other penalties.

In closing, we propose several directions for future research. Perhaps the most pressing question is to obtain the exact optimal trade-off for SLOPE. Regarding this question, a closer look at Figure 3 and Figure 5 suggests that our lower and upper bounds seem to coincide exactly when the TPP is small. If so, part of the optimal trade-off would already be specified. Having shown the advantage of SLOPE over the Lasso, a question of practical importance is to develop an approach to selecting regularization sequences for SLOPE to realize these benefits. Next, we would welcome extensions of our results to other methods using sorted

ℓ_1 regularization, such as the group SLOPE (Brzyski et al., 2019). For this purpose, our optimization-based technique for the variational calculus problems would likely serve as an effective tool. Recognizing that we have made heavy use of the two-level regularization sequences in many of our results, one is tempted to examine the possible benefits of using multi-level sequences for SLOPE (Zhang and Bu, 2021). Finally, a challenging question is to investigate the SLOPE trade-off under correlated design matrices; the recent development by Celentano et al. (2020) can be a stepping stone for this highly desirable generalization.

Acknowledgments. Weijie Su was supported in part by NSF through CAREER DMS-1847415 and CCF-1934876, an Alfred Sloan Research Fellowship, and the Wharton Dean's Research Fund. Cynthia Rush was supported by NSF through CCF-1849883 and this work was done in part while the author was visiting the Simons Institute for the Theory of Computing. Jason M. Klusowski was supported in part by NSF through DMS-2054808 and HDR TRIPODS DATA-INSPIRE DCCF-1934924.

REFERENCES

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

M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

P. C. Bellec, G. Lecué, and A. B. Tsybakov. SLOPE meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

M. Bogdan, E. v. d. Berg, W. Su, and E. Candès. Statistical estimation and testing via the sorted l1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

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*, 9(3):1103, 2015.

J. F. Bonnans and A. Shapiro. Optimization problems with perturbations: A guided tour. *SIAM review*, 40(2): 228–264, 1998.

J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

D. Brzyski, A. Gossmann, W. Su, and M. Bogdan. Group SLOPE—Adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525):419–433, 2019.

Z. Bu, J. M. Klusowski, C. Rush, and W. J. Su. Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537, 2020.

Z. Bu, J. M. Klusowski, C. Rush, and W. J. Su. Supplement to "Characterizing the SLOPE Trade-off: A Variational Perspective and the Donoho–Tanner Limit". *The Annals of Statistics*, 2022.

M. Celentano, A. Montanari, and Y. Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

I. Dikin. Iterative solution of problems of linear and quadratic programming. In *Doklady Akademii Nauk*, volume 174, pages 747–748. Russian Academy of Sciences, 1967.

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

D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009b.

D. L. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. 2005.

D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete & Computational Geometry*, 35(4):617–652, 2006.

D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

H. J. Ferreau, C. Kirches, A. Potschka, H. G. Bock, and M. Diehl. qpoases: A parametric active-set algorithm for quadratic programming. *Mathematical Programming Computation*, 6(4):327–363, 2014.

M. Figueiredo and R. Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Artificial Intelligence and Statistics*, pages 930–938. PMLR, 2016.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical programming*, 27(1):1–33, 1983.

M. G. G'Sell, T. Hastie, and R. Tibshirani. False variable selection rates in regression. *arXiv preprint arXiv:1302.2303*, 2013.

H. Hu and Y. M. Lu. Asymptotics and optimal designs of SLOPE for sparse linear regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 375–379. IEEE, 2019.

M. Kos and M. Bogdan. On the asymptotic properties of SLOPE. *Sankhya A*, 82(2):499–532, 2020.

A. Mousavi, A. Maleki, and R. G. Baraniuk. Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148, 2018.

K. G. Murty and F.-T. Yu. *Linear complementarity, linear and nonlinear programming*, volume 3. Citeseer, 1988.

H. E. Salzer and R. Zucker. Table of the zeros and weight factors of the first fifteen laguerre polynomials. *Bulletin of the American Mathematical Society*, 55(10):1004–1012, 1949.

A. Shapiro. Perturbation analysis of optimization problems in banach spaces. *Numerical Functional Analysis and Optimization*, 13(1-2):97–116, 1992.

S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. MIT Press, 2012.

W. Su and E. J. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

W. Su, M. Bogdan, and E. J. Candès. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45 (5):2133–2150, 2017.

W. J. Su. When is the first spurious variable selected by sequential regression procedures? *Biometrika*, 105(3): 517–527, 2018.

P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558, 2019.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

H. Wang, Y. Yang, and W. J. Su. The price of competition: Effect size heterogeneity matters in high dimensions. *arXiv preprint arXiv:2007.00566*, 2020a.

S. Wang, H. Weng, and A. Maleki. Does SLOPE outperform bridge regression? *arXiv preprint arXiv:1909.09345*, 2019.

S. Wang, H. Weng, and A. Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791–2823, 2020b.

H. Weng, A. Maleki, and L. Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Annals of Statistics*, 46(6A):3099–3129, 2018.

Y. Zhang and Z. Bu. Efficient designs of slope penalty sequences in finite dimension. *The 24th International Conference on Artificial Intelligence and Statistics*, 2021.