The Price of Competition: Effect Size Heterogeneity Matters in High Dimensions

Hua Wang, Yachong Yang, and Weijie J. Su, *Member, IEEE*

***Abstract*— In high-dimensional sparse regression, would increasing the signal-to-noise ratio while fixing the sparsity level always lead to better model selection? For high-dimensional sparse regression problems, surprisingly, in this paper we answer this question in the *negative* in the regime of linear sparsity for the Lasso method, relying on a new concept we term *effect size heterogeneity*. Roughly speaking, a regression coefficient vector has high effect size heterogeneity if its nonzero entries have significantly different magnitudes. From the viewpoint of this new measure, we prove that the false and true positive rates achieve the optimal trade-off *uniformly* along the Lasso path when this measure is maximal in a certain sense, and the worst trade-off is achieved when it is minimal in the sense that all nonzero effect sizes are roughly equal. Moreover, we demonstrate that the first false selection occurs much earlier when effect size heterogeneity is minimal than when it is maximal. The underlying cause of these two phenomena is, metaphorically speaking, the "competition" among variables with effect sizes of the same magnitude in entering the model. Taken together, our findings suggest that effect size heterogeneity shall serve as an important complementary measure to the sparsity of regression coefficients in the analysis of high-dimensional regression problems. Our proofs use techniques from approximate message passing theory as well as a novel technique for estimating the rank of the first false variable.**

***Index Terms*— Approximate message passing, false discovery rate, high-dimensional sparse regression, model selection, signal-to-noise ratio.**

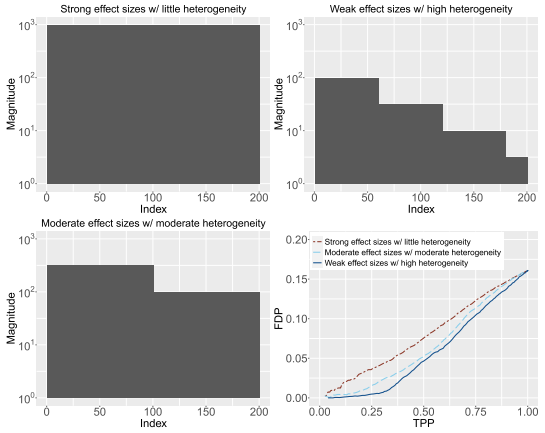

Fig. 1. The TPP–FDP trade-off along the entire Lasso path, with three different sets of regression coefficients. Note that the TPP–FDP trade-off is equivalent to the receiver operating characteristic curve. The sparsity of $\boldsymbol{\beta}$ is fixed to $k = 200$ (throughout this paper, we use k to denote the sparsity level) and the 200 true effects are plotted in the logarithmic scale in the three panels. For example, in the "Strong effect sizes w/ little heterogeneity" setting, $\beta_1 = \cdots = \beta_{200} = 10^3$, and $\beta_{201} = \cdots = \beta_{1000} = 0$. The design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has independent $\mathcal{N}(0, 1/n)$ entries, where $n = p = 1000$, and the noise term $\boldsymbol{z}$ has independent $\mathcal{N}(0, \sigma^2)$ entries with $\sigma = 0.01$. The bottom-right panel shows the plot of FDP as a function of TPP, averaged over 100 independent runs. For completeness, we remark that effect size heterogeneity influences model selection in a more complex manner at a higher noise level (see Figure 10).

I. Introduction

WE CONSIDER high-dimensional sparse regression problems in which we observe an n-dimensional response vector $\boldsymbol{y}$ that is generated by a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}, \tag{I.1}$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix of features, $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes an unknown vector of regression coefficients, and $\boldsymbol{z} \in \mathbb{R}^n$ is a noise term. In the big data era, this model has been increasingly applied to high-dimensional settings where the number of variables p is comparable to or even much larger than the number of observational units n. While this reality poses challenges to the regression problem, in many scientific problems there are good reasons to suspect that truly relevant variables account for a small fraction of all the observed variables or, equivalently, $\boldsymbol{\beta}$ is sparse in the sense that many of its components are zero or nearly zero. Indeed, a very impressive body of *theoretical* work shows that the difficulty of variable selection in the high-dimensional setting relies crucially on how sparse the regression coefficients are [1], [2].

This paper, however, asks whether there are other measures concerning the regression coefficients that have a *practical* impact on variable selection for the linear model (I.1). To address this question, we present a simulation study in Figure 1. Notably, the sparsity—or, put differently, the number of nonzero components—of the regression coefficients $\boldsymbol{\beta}$ is *fixed* to 200 across three experimental settings, but with vary-

Manuscript received 6 March 2021; revised 20 November 2021; accepted 5 March 2022. Date of publication 13 April 2022; date of current version 13 July 2022. This work was supported in part by NSF through CAREER under Grant DMS-1847415, Grant CCF-1763314, and Grant CCF-1934876; in part by the Wharton Dean's Research Fund; and in part by the Facebook Faculty Research Award. *(Hua Wang and Yachong Yang are co-first authors.) (Corresponding author: Weijie J. Su.)*

The authors are with the Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: suw@wharton.upenn.edu).

Communicated by A. Maleki, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3166720.

Digital Object Identifier 10.1109/TIT.2022.3166720

ing magnitudes of the 200 true effect sizes. The method we use for variable selection is the Lasso [3], which is perhaps the most popular model selector in the high-dimensional setting. Given a penalty parameter $\lambda > 0$, this method finds the solution to the convex optimization program

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_1, \tag{I.2}$$

where $\|\cdot\|$ and $\|\cdot\|_1$ denote the ℓ_2 and the ℓ_1 norms, respectively. A variable j is selected by this method at λ if $\widehat{\beta}_j(\lambda) \neq 0$, and a false selection occurs if it is a noise variable in the sense that $\beta_j = 0$. Formally, we use the false discovery proportion (FDP) and true positive proportion (TPP) as measures of the type I error and power, respectively, to assess the quality of the selected model $\{1 \leq j \leq p : \widehat{\beta}_j(\lambda) \neq 0\}$:

$$\text{FDP}_\lambda = \frac{\#\{j : \beta_j = 0 \text{ and } \widehat{\beta}_j(\lambda) \neq 0\}}{\#\{j : \widehat{\beta}_j(\lambda) \neq 0\}}, \tag{I.3}$$

$$\text{TPP}_\lambda = \frac{\#\{j : \beta_j \neq 0 \text{ and } \widehat{\beta}_j(\lambda) \neq 0\}}{\#\{j : \beta_j \neq 0\}}. \tag{I.4}$$

As is clear, we wish to select a model with a small FDP and large TPP.

Despite weaker effect sizes, strikingly, Figure 1 shows that the Lasso can achieve *better* model selection in terms of the TPP–FDP trade-off and, in particular, this counterintuitive behavior holds uniformly along the entire Lasso path or, equivalently, over all values of λ. Existing theory often analyzes how the worst-case performance of the Lasso and other related procedures depends on the regression coefficients through the sparsity of the regression coefficients (see, for example, [4]). However, the sparsity level is fixed across the experimental settings of Figure 1. In light of this, therefore, one would expect that the strong signals and weak signals would yield the best and worst model selection results, respectively. Figure 1 shows that this is not necessarily the case.

Thus, a *finer-grained* structural analysis of the effect sizes is needed to better understand the Lasso in some settings. In this paper, we address this important question by proposing a concept that we term *effect size heterogeneity* concerning the regression coefficients in high dimensions. Roughly speaking, a regression coefficient vector has higher effect size heterogeneity than another vector (of the same sparsity) if the nonzero entries of the former are more heterogeneous than those of the latter in terms of magnitude. As a complement to sparsity, effect size heterogeneity will be shown to have a significant impact on how the Lasso performs in terms of the false and true positive rates trade-off: while the sparsity level of the regression coefficients is fixed, the higher the effect size heterogeneity is, the better the Lasso performs. Turning back to Figure 1, we note that the strong effect sizes are the least heterogeneous in magnitude, and the weak effect sizes are the most heterogeneous. Therefore, the comparisons made in Figure 1 match well the implication of effect size heterogeneity.

Concretely, the main thrust of this paper lies in the development of two complementary perspectives to precisely quantify

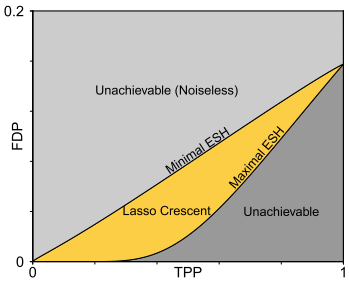

Fig. 2. The Lasso Crescent diagram specified by the parameters $n/p = 1$ and $k/p = 0.2$, following the setting in Figure 1. The lower/upper smooth curve is asymptotically achieved with maximal/minimal effect size heterogeneity (ESH) in the regime of infinite signal-to-noise ratio. The explicit expressions of the curves are given in Section II. Our Theorem 1 implies that nowhere on the Lasso path we can find any (TPP, FDP) pairs in the region below the Lasso Crescent (also see [5]). In the noiseless setting, moreover, this impossibility result continues to hold in the region above the Lasso Crescent (Theorem 2), which is labeled "Unachievable (Noiseless).".

the impact of effect size heterogeneity. First, following the setup of Figure 1, we consider the full possible range of the asymptotic trade-off between the TPP and FDP along the Lasso path, while varying the level of effect size heterogeneity. Assuming a random design with independent Gaussian entries and working in the regime of linear sparsity—meaning that the fraction of true effect sizes tends to a constant—we formally show that the full possible range is enclosed by two smooth curves in the (TPP, FDP) plane, which we referred to as the *Lasso Crescent*. Figure 2 presents an instance of the Lasso Crescent. More precisely, having excluded the impact of noise by taking $\boldsymbol{z} = \boldsymbol{0}$ in the linear model (I.1), the lower curve is asymptotically achieved when effect size heterogeneity is maximal in the sense that all true effect sizes are widely different from each other, while the upper curve is asymptotically achieved when the heterogeneity is minimal in the sense that all true effects are of the same size. In general, the (TPP, FDP) pairs computed from the entire Lasso path must be asymptotically sandwiched between the two curves in the noiseless setting or, equivalently, in the regime of the infinite signal-to-noise ratio. The gap between the two curves is fundamental in the sense that it persists no matter how strong the effects are.

While the TPP–FDP trade-off essentially examines the "bulk" of the Lasso solution path, the second perspective we take extends to the "edge": when does the first noise variable enter the model along the Lasso path? More precisely, we decrease λ from ∞ to 0 and find the first "time" a false selection occurs. To indicate the difficulty of consistent model selection, formally, we consider the rank of the first noise variable or, put concretely, one plus the number of the true variables before the Lasso selects the first noise variable. Intuitively, a large rank is desirable. As with the first perspective, assuming a Gaussian random design and regression coefficients with linear sparsity, we prove that the rank of the first false selection is bounded above by $(1+o(1))n/(2\log p)$. This upper bound, which approximately equals 72 in the setting of Figure 1, holds no matter how strong the effect sizes are. Interestingly, this upper bound is *exactly* achieved when effect

size heterogeneity is maximal and the noise level tends to zero. On the other hand, [6] has obtained a sharp prediction of the rank of the first false variable in the case of minimal effect size heterogeneity, which, together with our new result, shows that the first noise variable occurs much earlier with minimal effect size heterogeneity than with maximal effect size heterogeneity. Although not entirely related, the two perspectives consistently demonstrate that effect size heterogeneity is an important and useful concept for understanding the performance of the Lasso.

The fact that effect size heterogeneity matters, as shown above, is due to the bias introduced by the shrinkage nature of the Lasso. This bias in turn makes the residuals absorb many of the true effects that act as what we may want to call "shrinkage noise". Metaphorically, variables yet to be selected tend to "compete" with each other in entering the Lasso path and contribute to the shrinkage noise. The "competition" is particularly intensive among variables having about the same effect sizes, which is the case when effect size heterogeneity is low. As a price, the shrinkage noise gets inflated and some noise variables may be selected early due to their high correlations with the shrinkage noise. This is why false selections occur with a good chance and early. In contrast, when the heterogeneity is high, the largest true effect yet to be selected tends to have a significant correlation with the residuals, thereby having a better chance to be selected sooner. To appreciate this heuristic explanation, it is instructive to note that the least-squares estimator, if available, does not exhibit this price-of-competition phenomenon, as it is unbiased for the regression coefficients.[1] An alternative but less direct way to appreciate effect size heterogeneity is to relate it to the restricted eigenvalue condition [2]. Roughly speaking, this condition is concerned with a vector of regression coefficients such that its ℓ_1 norm is largely contributed by a few components, and such approximate sparse regression coefficients can be well estimated by the Lasso and the Dantzig selector under certain designs [2], [7]. From the viewpoint of this condition, therefore, regression coefficients with high effect size heterogeneity can be thought of as having a smaller effective sparsity level, which is favored by the Lasso.

As a final remark, the price-of-competition phenomenon does not appear if the sparsity is sub-linear in the ambient dimension p, which is often assumed in the copious body of literature on high-dimensional regression. In this regime of sparsity, effect size heterogeneity has a vanishing impact on the performance of the Lasso if the signal-to-noise ratio is sufficiently strong or the beta-min condition is satisfied. Our paper also departs from this line of literature from a technical standpoint. Indeed, the proofs of the results in this paper make heavy use of approximate message passing (AMP) theory [8]–[10], with nontrivial extensions.

A. Organization

The remainder of this paper is organized as follows. In Section II, we formalize the Lasso Crescent diagram by presenting our theoretical results that predict the TPP–FDP trade-off with respect to effect size heterogeneity. Section III extends the investigation of effect size heterogeneity to the problem of the first false variable along the Lasso path. Section IV is devoted to proving the results in Section II, whereas technical details of the proofs are deferred to the appendix. In Section V, we provide numerical studies to demonstrate the impact of effect size heterogeneity in general settings. We conclude the paper in Section VI with a few directions for future research.

[1] If the effect sizes are sufficiently strong, variable selection using the t values of the least-squares estimator can lead to full power without any type I errors.

II. The Lasso Crescent

In this section, we derive the full possible range of the asymptotic trade-off between the TPP and FDP along the Lasso path, with a focus on its dependence on effect size heterogeneity. Specifically, our results can be pictorially presented by the Lasso Crescent as in Figure 2, hence the title of this section. The proofs are deferred to Section IV.

Throughout this section, and indeed the entire paper, we assume the following *working hypotheses* to specify the linear model (I.1). For ease of reading, we use boldface letters to denote vectors and matrices.

Gaussian Design Matrix: We consider a sequence of designs $\boldsymbol{X} \in \mathbb{R}^{n_l \times p_l}$ consisting of i.i.d. $\mathcal{N}(0, 1/n_l)$ entries so that each column has an approximate unit ℓ_2 norm. As the index $l \to \infty$, we assume $p_l, n_l \to \infty$ with $n_l/p_l \to \delta$ for a constant $\delta > 0$. The index l is often omitted for the sake of simplicity.

Regression Coefficients: Let the regression coefficients $\beta_1, \ldots, \beta_p$ be i.i.d. copies of a random variable Π that satisfies $\mathbb{E}\,\Pi^2 < \infty$. Of particular interest to this paper is an ϵ-sparse prior Π in the sense that $\mathbb{P}(\Pi \neq 0) = \epsilon$ for a constant $0 < \epsilon < 1$. Thus, the realized $\beta_1, \ldots, \beta_p$ are in the linear sparsity regime since the sparsity is approximately equal to ϵp.

Noise: The noise term $\boldsymbol{z}$ consists of i.i.d. elements drawn from $\mathcal{N}(0, \sigma^2)$, where the noise level $\sigma \geq 0$ is fixed.

For completeness, $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{z}$ are jointly independent. These assumptions are used in the literature on AMP theory and its applications (see, for example, [9]–[13]) and, more recently, have been commonly made in the high-dimensional regression literature [14]–[17]. On top of that, we adopt some adjustments made by [5] that slightly simplify the assumptions on $\boldsymbol{\beta}$ and $\boldsymbol{z}$. Regarding the assumption on the noise, it is worth noting that we do not exclude the case $\sigma = 0$, which corresponds to noiseless observations. For some of the results in this section, the price-of-competition phenomenon manifests itself most clearly in the noiseless setting.

A. Most Heterogeneous Effect Sizes

Our first main theorem considers regression coefficients that are drawn from the following prior distribution:

Definition 2.1: For $M > 0$ and an integer $m > 0$, we call

$$\Pi^{\Delta} = \begin{cases} 0 & \text{w.p. } 1-\epsilon \\ M & \text{w.p. } \frac{\epsilon}{m} \\ M^2 & \text{w.p. } \frac{\epsilon}{m} \\ \cdots & \cdots \\ M^m & \text{w.p. } \frac{\epsilon}{m} \end{cases} \tag{II.1}$$

the (ϵ, m, M)-heterogeneous prior.

For notational convenience, we suppress the dependence of Π^{Δ} on ϵ, m, M. This prior is ϵ-sparse in the sense of the working hypotheses. As is clear, larger values of m, M would render the prior more heterogeneous. Indeed, this paper is primarily concerned with the case where both $M, m \to \infty$. This corresponds to the regime where the signal-to-noise ratio tends to infinity and, in addition, the true effect sizes are increasingly different. To be complete, the (ϵ, m, M)-heterogeneous prior is only a specific example that attains increasing heterogeneity. See Remark 2.2 for more examples.

Following (I.3), $\text{FDP}_\lambda(\Pi)$ and $\text{TPP}_\lambda(\Pi)$ denote the (random) false discovery proportion and true positive proportion, respectively, of the Lasso estimate at λ when the regression coefficients in (I.1) are i.i.d. draws from a prior Π. For ease of reading, we say a pair (TPP, FDP) *outperforms* another pair $(\text{TPP}', \text{FDP}')$ if $\text{TPP} > \text{TPP}'$ and $\text{FDP} < \text{FDP}'$. As noted earlier, all theoretical results in this paper are obtained under the working hypotheses. For conciseness, the statements of our theorems shall not mention this fact anymore.

Theorem 1: Let $C > c > 0$ be fixed. For any ϵ-sparse prior Π, if both m and M are sufficiently large in the (ϵ, m, M)-heterogeneous prior Π^{Δ}, then the following conclusions are true:

(a) The event

$$\bigcup_{c<\lambda,\lambda'<C} \Big\{ (\text{TPP}_{\lambda'}(\Pi), \text{FDP}_{\lambda'}(\Pi)) \text{ outperforms } (\text{TPP}_{\lambda}(\Pi^{\Delta}), \text{FDP}_{\lambda}(\Pi^{\Delta})) \Big\}$$

happens with probability tending to **zero** as $n, p \to \infty$.

(b) For any constant $\nu > 0$, no matter how we choose $\widehat{\lambda}'(\boldsymbol{y}, \boldsymbol{X}) \geq c$ adaptively as long as it always satisfies $\text{TPP}_{\widehat{\lambda}'}(\Pi) > \nu$, with probability approaching **one** there exists $\widehat{\lambda} > 0$ such that $\big(\text{TPP}_{\widehat{\lambda}}(\Pi^{\Delta}), \text{FDP}_{\widehat{\lambda}}(\Pi^{\Delta})\big)$ outperforms $\big(\text{TPP}_{\widehat{\lambda}'}(\Pi), \text{FDP}_{\widehat{\lambda}'}(\Pi)\big)$.

Remark 2.2: The priors for which the theorem holds can be extended in the following way. Consider a sequence of priors Π^{Δ} satisfying $\Pi^{\Delta} = 0$ with probability $1 - \epsilon$ and $\Pi^{\Delta} = M_i \neq 0$ with probability $\epsilon\gamma_i$ for $i = 1, \ldots, m$ such that $\gamma_1 + \cdots + \gamma_m = 1, \max_i \gamma_i \to 0$, and $\min_{1 \leq i \leq m} |M_i / M_{i-1}| \to \infty$ (set $M_0 = 1$). Alternatively, the nonzero component of the prior can be drawn from a continuous random variable with cumulative distribution function of form $\frac{\log_M x}{m}$ for $1 \leq x \leq M^m$. While the theorem statement is restricted to (ϵ, m, M)-heterogeneous priors for brevity, its proof considers the general case.

This theorem demonstrates the optimality of heterogeneous and strong effects in terms of the trade-off between the TPP and FDP. Importantly, this optimality is *uniform* in the sense that it holds along the entire Lasso path, no matter how strong the true effects coming from Π are. To be sure, the event as a union in (a) is taken over any (TPP, FDP) pair from the prior Π and any pair from the prior Π^{Δ}. Although each conclusion alone is not a consequence of the other, as we will see from the proof in Section IV, the two conclusions are built on top of the fact that the pairs $(\text{TPP}_\lambda, \text{FDP}_\lambda)$ with varying λ converge uniformly to a deterministic smooth curve for both Π and Π^{Δ}. This fact allows us to obtain the following byproduct:

Proposition 2.3: Under the assumptions of Theorem 1, for any sufficiently small constant $\nu > 0$, the following statement holds with probability tending to one: for any $\lambda, \lambda' > c$ such that $\big|\text{TPP}_\lambda(\Pi^{\Delta}) - \text{TPP}_{\lambda'}(\Pi)\big| < \nu$ and $\text{TPP}_{\lambda'}(\Pi) > 0.001$, we have

$$\text{FDP}_\lambda(\Pi^{\Delta}) < \text{FDP}_{\lambda'}(\Pi).$$

This result makes it self-evident why the prior Π^{Δ} is a most favorable for the entire Lasso path, though literally, we should interpret this favorability in the limit $m, M \to \infty$. More precisely, this result implies that given a required power level, the smallest possible FDP is achieved when the effects are increasingly heterogeneous and strong. Of note, the number 0.001 above can be replaced by any small positive constant, and it does not impede the interpretability of the theorem since we are generally not interested in a model that includes only a tiny fraction of true variables.

An interesting yet unaddressed question is to find an expression of the asymptotic minimum of FDP given $\text{TPP}_\lambda(\Pi^{\Delta}) = u$ in the limit $m, M \to \infty$. Call this function $q^{\Delta}(u; \delta, \epsilon)$. From our results, one can easily see that q^{Δ} is nothing but the lower *envelope* of instance-specific TPP–FDP trade-off curves over all ϵ-sparse priors. To see this, first note that one can prove that as $n, p \to \infty$, the pairs $(\text{TPP}_\lambda(\Pi), \text{FDP}_\lambda(\Pi))$ over all λ converge to a smooth curve, which is denoted by $q^{\Pi}(u)$ (see Section IV). Recognizing that Π^{Δ} is also ϵ-sparse and assuming $\lim_{m,M\to\infty} q^{\Pi^{\Delta}}$ exists, we must have

$$q^{\Delta}(u) := \lim_{m,M\to\infty} q^{\Pi^{\Delta}}(u) \geq \inf_{\Pi:\epsilon\text{-sparse}} q^{\Pi}(u). \tag{II.2}$$

On the other hand, it follows from Theorem 1 and in particular Proposition 2.3 that

$$q^{\Pi}(u) \geq \lim_{m,M\to\infty} \Big(q^{\Pi^{\Delta}}(u) + o(1) \Big) = q^{\Delta}(u)$$

for any ϵ-sparse prior Π. This display, together with (II.2), gives

$$q^{\Delta}(u) = \inf_{\Pi:\epsilon\text{-sparse}} q^{\Pi}(u). \tag{II.3}$$

Interestingly, the right-hand side of (II.3) has been tackled in [5], leading to a precise expression. To describe this expression, let $t^{\Delta}(u)$ be the largest positive root of the following equation in t,

$$\frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]} = \frac{1-u}{1-2\Phi(-t)}, \tag{II.4}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively. Theorem 2.1 in [5] shows that

$$\inf_{\Pi:\epsilon\text{-sparse}} q^{\Pi}(u) = \frac{2(1-\epsilon)\Phi(-t^{\Delta}(u))}{2(1-\epsilon)\Phi(-t^{\Delta}(u)) + \epsilon u}. \tag{II.5}$$

Taken together, (II.3) and (II.5) yield

$$q^{\Delta}(u) = \frac{2(1-\epsilon)\Phi(-t^{\Delta}(u))}{2(1-\epsilon)\Phi(-t^{\Delta}(u)) + \epsilon u}. \tag{II.6}$$

Remark 2.4: If $u = 0$, treat ∞ as a root of the equation and set $0/0 = 0$ in (II.5). As such, q^{Δ} satisfies $q^{\Delta}(0) = 0$. If $\delta < 1$ and ϵ is larger than a threshold determined by δ, the function q^{Δ} is defined only for u between 0 and a certain number strictly smaller than 1. This is where the celebrated Donoho–Tanner phase transition occurs [18] (also see Section B.2). Throughout this paper, however, we focus on the regime that is below the Donoho–Tanner phase transition—that is, the case where $\delta \geq 1$, or $\delta < 1$ and ϵ is small so that the range of u is the unit interval $[0, 1]$. In contrast, above the phase transition, the mapping from the TPP to FDP might not be unique (see Figure 3.1 in [19] and [20]).

In summary, we have the following corollary, which addresses the aforementioned question.

Corollary 2.5: Under the assumptions of Theorem 1, we have

$$\lim_{m,M\to\infty} \lim_{n,p\to\infty} \sup_{\lambda>c} \left|\mathrm{FDP}_{\lambda}(\Pi^{\Delta}) - q^{\Delta}\left(\mathrm{TPP}_{\lambda}(\Pi^{\Delta})\right)\right| = 0,$$

where $\lim_{n,p\to\infty}$ is taken in probability. Moreover, for any ϵ-sparse prior Π, we have

$$\mathrm{FDP}_{\lambda}(\Pi) \geq q^{\Delta}\left(\mathrm{TPP}_{\lambda}(\Pi)\right) - 0.001$$

for all $\lambda > c$ with probability tending to one.

Remark 2.6: As $\lambda \to \infty$, both $\mathrm{TPP}_{\lambda}(\Pi^{\Delta})$ and $\mathrm{FDP}_{\lambda}(\Pi^{\Delta})$ tend to 0. Hence, there is no need to impose an upper bound on λ when taking the supremum $\sup_{\lambda>c}$. The second conclusion of Corollary 2.5 follows from Proposition 2.3 in conjunction with the continuity of q^{Δ}. As earlier, 0.001 can be replaced by any positive constant.

The second conclusion of Corollary 2.5 is part of Theorem 2.1 in [5] and demonstrates that true variables and irrelevant variables are always interspersed along the Lasso path. In particular, this is true when the regularization parameter λ tends to 0. In this case, indeed, the Lasso would select a significant fraction of false variables with vanishing but nonzero estimated coefficients. This fact necessitates a form of calibration of the Lasso estimates for variable selection [19], [20].

The significance of Theorem 1 and Corollary 2.5, however, extends beyond earlier results. Precisely, [5] derived the expression (II.5) by constructing a different signal prior Π for each power level u. Indeed, the nonzero component of the prior constructed in [5] has two different magnitudes with weights depending on u, as opposed to an increasing number of different magnitudes as in the (ϵ, m, M)-heterogeneous prior. The increasing level of heterogeneity allows us to give a one-shot construction of most heterogeneous priors at all power levels.

B. Least Heterogeneous Effect Sizes

We now turn to the opposite question: which effect sizes lead to the worst trade-off between the TPP and FDP along the Lasso path? Inspired by the interpretation of effect size heterogeneity, it is natural to consider the following signal prior as a candidate:

Definition 2.7: For $M > 0$, we call

$$\Pi^{\nabla} = \begin{cases} 0 & \text{w.p. } 1-\epsilon \\ M & \text{w.p. } \epsilon \end{cases} \tag{II.7}$$

the (ϵ, M)-homogeneous prior.

This prior would render all true effect sizes equal, thereby being least heterogeneous or most homogeneous among all ϵ-sparse priors. The following theorem confirms our intuition that this homogeneous prior is least favorable for the Lasso as the resulting effect sizes give the least optimal trade-off between false positives and power.

Theorem 2: Let $C > c > 0$ be fixed. In the noiseless setting — that is, $z = 0$ — for any ϵ-sparse prior Π that is non-constant conditional on $\Pi \neq 0$, the following conclusions are true for the (ϵ, M)-homogeneous prior Π^{∇}:

(a) The event

$$\bigcup_{c<\lambda,\lambda'<C} \Big\{ \left(\mathrm{TPP}_{\lambda}(\Pi^{\nabla}), \mathrm{FDP}_{\lambda}(\Pi^{\nabla})\right) \text{ outperforms } \left(\mathrm{TPP}_{\lambda'}(\Pi), \mathrm{FDP}_{\lambda'}(\Pi)\right) \Big\}$$

happens with probability tending to **zero** as $n, p \to \infty$.

(b) For any constant $\nu > 0$, no matter how we choose $\widehat{\lambda'}(\boldsymbol{y}, \boldsymbol{X}) \geq c$ adaptively as long as it always satisfies $\mathrm{TPP}_{\widehat{\lambda'}}(\Pi) > \nu$, with probability tending to **one** there exists $\widehat{\lambda} > 0$ such that $\left(\mathrm{TPP}_{\widehat{\lambda'}}(\Pi), \mathrm{FDP}_{\widehat{\lambda'}}(\Pi)\right)$ outperforms $\left(\mathrm{TPP}_{\widehat{\lambda}}(\Pi^{\nabla}), \mathrm{FDP}_{\widehat{\lambda}}(\Pi^{\nabla})\right)$.

This theorem is similar, but in the opposite sense, to Theorem 1. One distinction between the two theorems is that Theorem 2 assumes the noiseless setting, as opposed to the noisy setting considered in Theorem 1. The noiseless setting is equivalent to an infinite value of the signal-to-noise ratio, which allows us to better isolate the impact of effect size heterogeneity from that of the noise term. That said, this theorem remains true in the presence of noise by setting a sufficiently large magnitude M for the true effect sizes.

Just as Proposition 2.3 does, the following result follows from the proof of Theorem 2 presented in Section IV.

Proposition 2.8: Under the assumptions of Theorem 2, for any sufficiently small constant $\nu > 0$, the following statement holds with probability tending to one: if $\lambda, \lambda' > c$ satisfy $\mathrm{TPP}_{\lambda'}(\Pi) > 0.001$ and $\left|\mathrm{TPP}_{\lambda}(\Pi^{\nabla}) - \mathrm{TPP}_{\lambda'}(\Pi)\right| < \nu$, then

$$\mathrm{FDP}_{\lambda}(\Pi^{\nabla}) > \mathrm{FDP}_{\lambda'}(\Pi).$$

As is clear, this result demonstrates that the prior Π^{∇} is least favorable for the entire Lasso path in the noiseless case. Roughly speaking, this proposition shows that if the two Lasso problems agree on the value of the TPP along their paths, then the (ϵ, M)-homogeneous prior Π^{∇} must yield a higher FDP. As with Proposition 2.3, 0.001 can be replaced by any positive constant. On a related note, the prior (II.7) is known to be least favorable for certain estimation problems both in the noiseless and noisy settings [21] (see also Lemma 4.4.1 and Corollary 4.4.3 in [22]). An interesting direction for future research is to study the relationship between estimation and variable selection with regard to the least favorability of the prior distribution.

The sharp distinction between Propositions 2.3 and 2.8 must be attributed to the priors Π^{Δ} and Π^{∇}. The cause is, loosely speaking, due to the "competition" among variables with about the same effect sizes in entering the Lasso model. However, we find it easier to elucidate the underlying cause when studying the rank of the first false variable and thus defer the detailed discussion to Section III.

We now proceed to specify the curve on which $(\mathrm{TPP}_\lambda(\Pi^{\nabla}), \mathrm{FDP}_\lambda(\Pi^{\nabla}))$ lies in the limit. For a fixed α, let $\varsigma = \varsigma(\alpha)$ denote the largest root of the equation

$$\begin{aligned}\delta = {} & 2(1-\epsilon)[(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha)] - \epsilon(2\alpha+\varsigma)\phi(\varsigma) \\ & + \epsilon\varsigma\phi(2\alpha+\varsigma) + \epsilon(1+\alpha^2)[\Phi(\varsigma) + \Phi(-2\alpha-\varsigma)] \\ & + \epsilon(\varsigma+\alpha)^2[\Phi(-\varsigma) + \Phi(-2\alpha-\varsigma)],\end{aligned}$$

and let $t^{\nabla} = t^{\nabla}(u;\delta,\epsilon)$ be the largest root of the following equation in α:

$$\Phi(\varsigma(\alpha)) + \Phi(-2\alpha - \varsigma(\alpha)) = u.$$

With all of these in place, define

$$q^{\nabla}(u;\delta,\epsilon) = \frac{2(1-\epsilon)\Phi(-t^{\nabla}(u))}{2(1-\epsilon)\Phi(-t^{\nabla}(u)) + \epsilon u}. \tag{II.8}$$

The derivation of the expression is given in Lemma 1.16 in Section B.2. The following result shows that this function describes the limiting trade-off between the TPP and FDP in the case of minimal effect size heterogeneity:

Corollary 2.9: Under the assumptions of Theorem 2, we have

$$\lim_{n,p\to\infty} \sup_{\lambda > c} \left| \mathrm{FDP}_\lambda(\Pi^{\nabla}) - q^{\nabla}\left(\mathrm{TPP}_\lambda(\Pi^{\nabla})\right) \right| = 0.$$

Moreover, for any ϵ-sparse prior Π, we have

$$\mathrm{FDP}_\lambda(\Pi) \le q^{\nabla}\left(\mathrm{TPP}_\lambda(\Pi)\right) + 0.001$$

for all $\lambda > c$ with probability tending to one.

As with Theorem 2, Corollary 2.9 holds for any $M > 0$ because of the noiseless setting.

Taken together, Corollaries 2.5 and 2.9 give the following result:

Theorem 3: Let $c > 0$ be any small constant. In the noiseless setting, for any ϵ-sparse prior Π, we have

$$q^{\Delta}\left(\mathrm{TPP}_\lambda(\Pi)\right) - 0.001 \le \mathrm{FDP}_\lambda(\Pi) \le q^{\nabla}\left(\mathrm{TPP}_\lambda(\Pi)\right) + 0.001$$

for all $\lambda > c$ with probability tending to one.

The two curves q^{Δ} and q^{∇} enclose a crescent-shaped region, which we call the Lasso Crescent. This theorem shows any $(\mathrm{TPP}, \mathrm{FDP})$ pairs along the entire Lasso path would essentially lie in the corresponding Lasso Crescent that is specified by the shape n/p of the design and the sparsity ratio k/p of the effect sizes, and this region is tight. Figure 3 presents two instances of the Lasso Crescent, with simulations showing good agreement between the predicted and observed behaviors.[2]

[2]R and Matlab code to calculate q^{Δ} and q^{∇} is available at https://github.com/huawang-wharton/effectsizeheterogeneity.

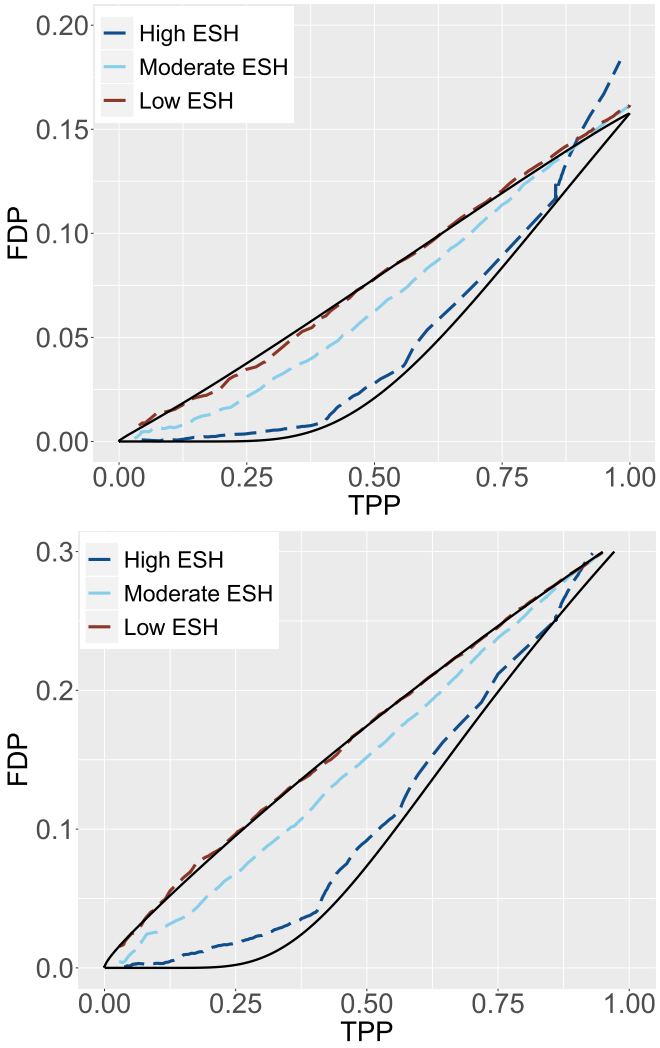

Fig. 3. Illustration of the interpretation of the Lasso Crescent via Theorem 3. The design of size $n \times p$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries and the noise level is set to 0. Specifically, we use $n = p = 1000$, and sparsity $k = 200$ in the left panel, and $n = 800, p = 1200$, $k = 200$ in the right panel. The "high effect size heterogeneity (ESH)" setting: the 200 coefficients take 4 different values; The "moderate ESH" setting: the first 100 coefficients are set to 100 and the second 100 coefficients are set to 50; The "low ESH" setting: the 200 coefficients are set to 100. The dashed lines are averaged over 200 independent runs of the Lasso path. The two boundaries q^{Δ} and q^{∇} are in solid black lines.

III. The First False Selection

In this section, we examine the impact of effect size heterogeneity on model selection by the Lasso from a different perspective: when is the first false variable selected? Intuitively, the later the first false variable occurs, the better the method performs. Using a mix of new and old results, this section will show that the first false variable occurs much earlier when effect size heterogeneity is minimal than when it is maximal.

Denote the rank of the first falsely selected variable by

$$T := \#\{j : \widehat{\beta}_j(\lambda^* - 0) \neq 0\} = \#\{j : \widehat{\beta}_j(\lambda^*) \neq 0\} + 1.$$

Above, λ^* is the first time along the Lasso path that a false variable is about to be selected:

$$\lambda^* = \sup\{\lambda : \text{there exists } 1 \le i \le p, \widehat{\beta}_i(\lambda) \neq 0, \beta_i = 0\},$$

and $\lambda^* - 0$ informally represents a value that is infinitesimally smaller than λ^*. In words, T is equal to one plus the number of true variables before the first false variable.

The problem of the rank of the first false selection has been considered by [6] in the case where all nonzero regression coefficients are equal. This corresponds to minimal effect size heterogeneity. While we continue employing the working hypotheses as earlier, in this section the regression coefficients $\boldsymbol{\beta}$ are assumed to be deterministic.

Proposition 3.1: [6, Theorem 2] Under the working hypotheses, let $\beta_j = M$ for $1 \le j \le k$ and $\beta_j = 0$ for $k+1 \le j \le p$, where $k/p \to \epsilon$ and $M \to \infty$ as $n, p \to \infty$. Then, the rank T of the first false variable selected by the Lasso satisfies

$$\log T = (1 + o_{\mathbb{P}}(1))\sqrt{\frac{2\delta \log p}{\epsilon}},$$

where $o_{\mathbb{P}}(1)$ tends to 0 in probability.

This result also applies to forward stepwise regression and least angle regression [23]. Note that this proposition considers the regime where the signal-to-noise ratio $M/\sigma \to \infty$ as σ is fixed. If M/σ is bounded, one has $\log T \le (1 + o_{\mathbb{P}}(1))\sqrt{\frac{2\delta \log p}{\epsilon}}$ [6, Theorem 1]. Indeed, the original theorem predicts that

$$\log T \le (1 + o_{\mathbb{P}}(1))\Big(\sqrt{2n(\log p)/k} - n/(2k) + \log\big(n/(2p\log p)\big)\Big),$$

which is reduced to the upper bound above since $n/p \to \delta$ and $k/p \to \epsilon$ under our working hypotheses.

Turning to most heterogeneous effect sizes, we have the result below.

Proposition 3.2: Under the working hypotheses, let $\beta_j = M^{k+1-j}$ for $1 \le j \le k$ and $\beta_j = 0$ for $k+1 \le j \le p$, where $k/p \to \epsilon$. If M is sufficiently large, then there exists λ depending on n such that

$$\#\left\{j : \widehat{\beta}_j(\lambda) \ne 0, \beta_j \ne 0\right\} = (1 + o_{\mathbb{P}}(1))\frac{n}{2\log p},$$
$$\text{and } \#\left\{j : \widehat{\beta}_j(\lambda) \ne 0, \beta_j = 0\right\} = 0$$

as $n, p \to \infty$.

The proof of this proposition is given in the appendix. Regarding how large M should be, precisely, this proposition holds if M satisfies $M \ge n^a$ as $n \to \infty$ for any constant $a > \frac{1}{2}$. It is also worth mentioning that the proof is adapted from the proof of Theorem 1 in [24]. The effect sizes in Proposition 3.2 are essentially the same as an (ϵ, m, M)-heterogeneous prior (II.1) with $m \to \infty$.

Proposition 3.2 asserts that all $(1 + o_{\mathbb{P}}(1))\frac{n}{2\log p}$ selected variables are true at some point along the Lasso path. If the Lasso does not kick out any selected variables before that point,[3] this result implies that $T \ge (1 + o_{\mathbb{P}}(1))\frac{n}{2\log p}$. Recognizing the fact that

$$\mathrm{e}^{(1+o_{\mathbb{P}}(1))\sqrt{\frac{2\delta\log p}{\epsilon}}} \ll (1 + o_{\mathbb{P}}(1))\frac{n}{2\log p},$$

most heterogeneous effect sizes are more *favorable* than least heterogeneous effect sizes for the Lasso not only in terms of the TPP–FDP trade-off as shown in the previous section, but also in terms of the rank of the first false variable.

Unlike Theorem 1 and Theorem 2, the two propositions here are silent on whether their bounds can be extended to general ϵp-sparse effect sizes. The following theorem gives a partial *affirmative* answer to this question, which broadly applies to all regression coefficients with sparsity no more than ϵp, as opposed to the exact sparsity level ϵp.

Theorem 4: Under the working hypotheses, for arbitrary regression coefficients $\boldsymbol{\beta}$ with sparsity satisfying $k \le \epsilon p$, the rank T of the first false variable selected by the Lasso satisfies

$$T \le (1 + o_{\mathbb{P}}(1))\frac{n}{2\log p}$$

as $n, p \to \infty$.

Together with Proposition 3.2, this theorem indicates that maximal effect size heterogeneity is most favorable for the Lasso in terms of the rank of the first false variable. Importantly, the sharp bound $(1 + o_{\mathbb{P}}(1))\frac{n}{2\log p}$ is the maximum number of true variables before a false selection for essentially all sparsity levels, no matter how strong and how heterogeneous the effect sizes are. This novel result is a contribution of independent interest to high-dimensional statistics. The proof is given in Section III-A and does not involve any elements from AMP theory.

In regard to Proposition 3.1, however, it is tempting to ask whether minimal effect size heterogeneity is least favorable from the same standpoint; that is, whether or not

$$\log T \ge (1 + o_{\mathbb{P}}(1))\sqrt{\frac{2\delta \log p}{\epsilon}}$$

for any ϵp-sparse regression coefficients in the noiseless case. We leave this question for future work.

In passing, we briefly explain how and why effect size heterogeneity has a significant impact on model selection by the Lasso, shedding light on the price-of-competition phenomenon. To ease the elaboration, we assume the noiseless setting ($\boldsymbol{z} = \boldsymbol{0}$) and denote by S the set of all true variables. Consider the Lasso solution $\widehat{\boldsymbol{\beta}}(\lambda)$ at some λ where no false selection occurs (that is, the support $\widehat{S}$ of $\widehat{\boldsymbol{\beta}}$ is a subset of S). Our explanation relies crucially on the fact that a variable $\boldsymbol{X}_j$ ($j \notin \widehat{S}$) is likely to be the next selected variable if its inner product with the residual, $\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, is the largest in magnitude. Note that (denote by $\boldsymbol{X}_Q$ the matrix that is formed by the columns corresponding to Q for a subset Q of $\{1, \ldots, p\}$)

$$\begin{aligned}\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) &= \boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_{\widehat{S}}\widehat{\boldsymbol{\beta}}_{\widehat{S}}) \\ &= \boldsymbol{X}_j^\top \boldsymbol{X}_{S\setminus\widehat{S}}\boldsymbol{\beta}_{S\setminus\widehat{S}} + \boldsymbol{X}_j^\top \boldsymbol{X}_{\widehat{S}}(\boldsymbol{\beta}_{\widehat{S}} - \widehat{\boldsymbol{\beta}}_{\widehat{S}}).\end{aligned}$$

Now, we argue that the largest $\boldsymbol{X}_j^\top \boldsymbol{X}_{S\setminus\widehat{S}}\boldsymbol{\beta}_{S\setminus\widehat{S}}$ in absolute value in the case of high effect size heterogeneity is likely to be from a true variable $\boldsymbol{X}_j$ ($j \in S \setminus \widehat{S}$) and, conversely,

[3] It is well-known that along the Lasso path, a selected variable can be dropped out as λ decreases [23]. However, we did not observe this phenomenon before the first $(1 + o_{\mathbb{P}}(1))\frac{n}{2\log p}$ variables are selected in all of our simulations.

it is likely to be from an irrelevant variable $\boldsymbol{X}_j$ ($j \notin S$) if effect size heterogeneity is low. Informally, regarding $\widehat{S}$ as deterministic, then $\boldsymbol{X}_j^\top \boldsymbol{X}_{S\setminus\widehat{S}}\boldsymbol{\beta}_{S\setminus\widehat{S}}$ is approximately normally distributed with variance $\|\boldsymbol{X}_{S\setminus\widehat{S}}\boldsymbol{\beta}_{S\setminus\widehat{S}}\|^2/n \approx \|\boldsymbol{\beta}_{S\setminus\widehat{S}}\|^2/n$ and mean

$$\begin{cases} 0 & \text{if } j \notin S \\ \beta_j & \text{if } j \in S \setminus \widehat{S}. \end{cases}$$

In the setting of Proposition 3.2 where true effect sizes are widely different from each other, the standard deviation $\|\boldsymbol{\beta}_{S\setminus\widehat{S}}\|/\sqrt{n}$ is much smaller than $\sup_{j\in S\setminus\widehat{S}}\beta_j$. Consequently, the unselected variable with the largest effect size $\sup_{j\in S\setminus\widehat{S}}\beta_j$ tends to stand out, with essentially no "competition" among all unselected variables, thereby being the next selected variable. In the setting of Proposition 3.1, however, the standard deviation $\|\boldsymbol{\beta}_{S\setminus\widehat{S}}\|/\sqrt{n}$ is comparable to the largest unselected effect sizes, which are in fact of the same size. Another way to put this is that the overall effect is evenly distributed across true variables, and the resulted competition renders any variable dwarfed by the noise. Accordingly, some noise variable $\boldsymbol{X}_j$ is very likely to have a larger inner product $\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ in magnitude than any unselected true variable does. As such, a false selection is likely to occur very early when effect size heterogeneity is low.

A. Proof of Theorem 4

Let $\nu > 0$ be any small constant and denote by $\mathcal{A}_\nu$ the event that the rank of the first false variable

$$T \geq (1+\nu)\frac{n}{2\log p}.$$

The proof follows if one can show $\mathbb{P}(\mathcal{A}_\nu) \to 0$ for all $\nu > 0$ as $n, p \to \infty$. Recall that S denotes the support $\mathrm{supp}(\boldsymbol{\beta})$. If the sparsity $|S| = k < (1+\nu)\frac{n}{2\log p} - 1 = (1+\nu+o(1))\frac{n}{2\log p}$, the event $\mathcal{A}_\nu$ is an empty set because T is always no greater than $|S|+1 < (1+\nu)\frac{n}{2\log p}$, leading to $\mathbb{P}(\mathcal{A}_\nu) = 0$.

Now, we turn to the more challenging case where $k \geq (1+\nu)\frac{n}{2\log p} - 1$ and the remainder of the proof aims to show $\mathbb{P}(\mathcal{A}_\nu) \to 0$. Consider the solution $\widetilde{\boldsymbol{\beta}}(\lambda)$ to the restricted Lasso problem

$$\widetilde{\boldsymbol{\beta}}(\lambda) := \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^k} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_S\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_1. \tag{III.1}$$

Let

$$\overline{\lambda} = \sup\left\{\lambda : \|\widetilde{\boldsymbol{\beta}}(\lambda)\|_0 \geq (1+\nu)\frac{n}{2\log p} - 1\right\}$$

be the first time that the restricted Lasso selects $(1+\nu)\frac{n}{2\log p} - 1$ variables and denote by $\widehat{S}$ the support of $\widetilde{\boldsymbol{\beta}}(\overline{\lambda}-0)$ (here $\overline{\lambda}-0$ is infinitesimally smaller than $\overline{\lambda}$). In particular, this set must satisfy

$$(1+\nu)\frac{n}{2\log p} - 1 \leq |\widehat{S}| \leq (1+\nu)\frac{n}{2\log p}. \tag{III.2}$$

On the event $\mathcal{A}_\nu$, the support of the full Lasso solution is a subset of S. Therefore, $\widetilde{\boldsymbol{\beta}}(\overline{\lambda})$ defined in (III.1) is also the solution to the full Lasso problem at $\overline{\lambda}$:

$$\widetilde{\boldsymbol{\beta}}(\overline{\lambda}) = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \overline{\lambda}\|\boldsymbol{b}\|_1.$$

Note that $\widetilde{\boldsymbol{\beta}}(\overline{\lambda})$ may be k-dimensional as in (III.1) or p-dimensional by setting the remaining $p-k$ entries to zero, depending on the context. Writing $\widetilde{\boldsymbol{\beta}}$ as a shorthand for $\widetilde{\boldsymbol{\beta}}(\overline{\lambda})$, as a consequence, we have $\left|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\right| \leq \overline{\lambda}$ for all $j \notin S$ on the event $\mathcal{A}_\nu$, thereby certifying

$$\mathbb{P}(\mathcal{A}_\nu) \leq \mathbb{P}\left(\left|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\right| \leq \overline{\lambda} \text{ for all } j \notin S\right).$$

To prove $\mathbb{P}(\mathcal{A}_\nu) \to 0$, therefore, it suffices to show that

$$\max_{j\notin S}\left|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\right| > \overline{\lambda} \tag{III.3}$$

with probability tending to one. Making use of the independence between $\boldsymbol{X}_j$ and $\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}}$, $\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})$'s are $p-k$ i.i.d. normal random variables with mean 0 and variance $\|\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}}\|^2/n$, conditional on $\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}}$. This gives

$$\begin{aligned}\max_{j\notin S}\left|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\right| &= (1+o_{\mathbb{P}}(1))\frac{\|\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}}\|}{\sqrt{n}}\sqrt{2\log(p-k)} \\ &\geq (1+o_{\mathbb{P}}(1))\frac{\|\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\|}{\sqrt{n}}\sqrt{2\log(p-k)},\end{aligned} \tag{III.4}$$

where the inequality follows since $\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top$ is a projection. For the moment, take the inequality

$$\|\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\| \geq (1+c)\overline{\lambda}\sqrt{\frac{n}{2\log p}} \tag{III.5}$$

as given for some constant $c > 0$ possibly depending on ν, with probability tending to one. Combining (III.4) and (III.5) yields

$$\begin{aligned}\max_{j\notin S}\left|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\right| &\geq (1+o_{\mathbb{P}}(1))\sqrt{2\log(p-k)}\frac{(1+c)\overline{\lambda}}{\sqrt{2\log p}} \\ &= (1+c+o_{\mathbb{P}}(1))\overline{\lambda}\sqrt{\frac{\log(p-k)}{\log p}} \\ &\geq (1+c+o_{\mathbb{P}}(1))\overline{\lambda}\sqrt{\frac{\log(p-\epsilon p)}{\log p}} \\ &= (1+c+o_{\mathbb{P}}(1))\overline{\lambda}\end{aligned}$$

with probability tending to one, which ensures (III.3).

We proceed to complete the proof of this theorem by verifying (III.5). The Karush–Kuhn–Tucker condition for the Lasso asserts that

$$\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}}) = \overline{\lambda}\,\mathrm{sgn}(\widetilde{\boldsymbol{\beta}}_{\widehat{S}}) \in \overline{\lambda}\{1,-1\}^{|\widehat{S}|},$$

from which we get

$$\left\|\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y} - \boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\right\| = \overline{\lambda}\sqrt{|\widehat{S}|}.$$

A classical result in random matrix theory (see Lemma 1.1 in the appendix) shows that the singular values of

$\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}$ are all bounded below by $\frac{1}{\sqrt{1+\theta}}$ with probability $1-1/p^2$, where

$$\theta = C\sqrt{\frac{(1+\nu)\frac{n}{2\log p}\log(p/((1+\nu)\frac{n}{2\log p}))}{n}} \asymp \sqrt{\frac{\log\log p}{\log p}} \tag{III.6}$$

for an absolute constant C. This allows us to get

$$\begin{aligned}\|\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y}-\boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\| &\geq \frac{1}{\sqrt{1+\theta}}\|\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y}-\boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\| \\ &= \sqrt{\frac{|\widehat{S}|}{1+\theta}}\cdot\overline{\lambda}\end{aligned} \tag{III.7}$$

with probability tending to one. Recognizing that $\theta < \nu/2$ for sufficiently large p and plugging (III.6) and (III.2) into (III.7), we obtain

$$\begin{aligned}\|\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y}-\boldsymbol{X}_S\widetilde{\boldsymbol{\beta}})\| &\geq \sqrt{\frac{(1+\nu)\frac{n}{2\log p}-1}{1+\nu/2}}\cdot\overline{\lambda} \\ &= (1+c)\overline{\lambda}\sqrt{\frac{n}{2\log p}}\end{aligned}$$

with probability approaching one, where $c = \sqrt{\frac{1+\nu}{1+\nu/2}} - 1 > 0$. This proves (III.5), thereby completing the proof of Theorem 4.

IV. Proofs for the Lasso Crescent

To prove Theorems 1 and 2, we start by introducing AMP theory at a *minimal* level. In the case of the Lasso, loosely speaking, tools from AMP theory enable the characterization of the asymptotic joint distribution of the Lasso estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ and the true regression coefficients $\boldsymbol{\beta}$ under the working hypotheses [8]–[10]. The distribution is determined by several parameters that can be solved from two equations (see (IV.1) below). It is important to note, however, that this body of literature only allows for the analysis of the Lasso at a *fixed* value of λ. As such, these tools are not directly applicable to the full Lasso path that this paper deals with.

To overcome this difficulty, we leverage a recent extension on AMP theory that allows us to work on the Lasso problem uniformly over its penalty parameter [5]. Under the working hypotheses, let $\tau > 0$ and $\alpha > \alpha_0$ be the unique solution to

$$\begin{aligned}\tau^2 &= \sigma^2 + \frac{1}{\delta}\,\mathbb{E}(\eta_{\alpha\tau}(\Pi+\tau W)-\Pi)^2 \\ \lambda &= \left(1-\frac{1}{\delta}\,\mathbb{P}(|\Pi+\tau W|>\alpha\tau)\right)\alpha\tau,\end{aligned} \tag{IV.1}$$

where $\eta_c(x) := \mathrm{sgn}(x)\cdot\max\{|x|-c,0\}$ is the soft-thresholding function, W is a standard normal random variable that is independent of Π, and $\alpha_0 = 0$ if $\delta > 1$ and otherwise is the unique root of $(1+t^2)\Phi(-t) - t\phi(t) = \frac{\delta}{2}$ in $t \geq 0$. Let $\Pi^\star$ be distributed the same as Π conditional on $\Pi \neq 0$, and define the two deterministic functions

$$\begin{aligned}\mathrm{tpp}_\lambda^\infty(\Pi) &= \mathbb{P}(|\Pi^\star+\tau W|>\alpha\tau) \\ \mathrm{fdp}_\lambda^\infty(\Pi) &= \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha)+\epsilon\mathbb{P}(|\Pi^\star+\tau W|>\alpha\tau)}.\end{aligned} \tag{IV.2}$$

Above, $\Pi^\star$ remains independent of W. For convenience, we use $\xrightarrow{\mathbb{P}}$ to denote convergence in probability. With the notation in place, now we state a lemma that our proofs rely on.

Lemma 4.1 ([25, Lemma A.2]): Fix $0 < c < C$. Under the working hypotheses, we have

$$\begin{aligned}\sup_{c<\lambda<C}|\mathrm{TPP}_\lambda(\Pi)-\mathrm{tpp}_\lambda^\infty(\Pi)| &\xrightarrow{\mathbb{P}} 0, \\ \text{and}\quad \sup_{c<\lambda<C}|\mathrm{FDP}_\lambda(\Pi)-\mathrm{fdp}_\lambda^\infty(\Pi)| &\xrightarrow{\mathbb{P}} 0.\end{aligned} \tag{IV.3}$$

Lemma 4.1 offers all the elements the present paper needs from AMP theory. In addition to the use of this lemma, notably, our proofs of Theorems 1 and 2 involve several technical novelties that we shall highlight in Sections IV-A and IV-B. Relating to the literature, the convergence of $\mathrm{TPP}_\lambda(\Pi)$ and $\mathrm{FDP}_\lambda(\Pi)$ for a single λ has been established earlier in [10], [26].

We use $q^\Pi(\cdot)$ to represent the λ-parameterized curve $(\mathrm{tpp}_\lambda^\infty, \mathrm{fdp}_\lambda^\infty)$ in the sense that

$$\mathrm{fdp}_\lambda^\infty(\Pi) = q^\Pi(\mathrm{tpp}_\lambda^\infty(\Pi)).$$

Formally, Lemma 1.11 in Section B.1 demonstrates that the instance-specific trade-off curve q^Π is continuously differentiable and strictly increasing. Relating to Section II, Corollary 2.5 implies that, taking the (ϵ, m, M)-heterogeneous prior Π^Δ, q^{Π^Δ} converges to q^Δ as $m, M \to \infty$. Likewise, from Corollary 2.9 it follows that $q^{\Pi^\nabla}(\cdot)$ is identical to $q^\nabla(\cdot)$ in the noiseless setting.

A. The Upper Boundary

Our first aim is to prove Theorem 2 along with Proposition 2.8. The proof is built on top of the following important lemma, which considers a non-constant $\Pi^\star$.

Lemma 4.2: Let Π be any ϵ-sparse prior that is non-constant conditional on $\Pi \neq 0$. In the noiseless setting $\sigma = 0$, we have

$$q^\Pi(u) < q^\nabla(u)$$

for all $0 < u < 1$.

Taking this lemma as given for the moment, we prove part (a) of Theorem 2.

Proof of Theorem 2(a): We start by pointing out the following fact: there exists a constant $\upsilon > 0$ such that for all $c < \lambda, \lambda' < C$, the two inequalities

$$\mathrm{tpp}_\lambda^\infty(\Pi^\nabla) > \mathrm{tpp}_{\lambda'}^\infty(\Pi) - \upsilon \text{ and } \mathrm{fdp}_\lambda^\infty(\Pi^\nabla) < \mathrm{fdp}_{\lambda'}^\infty(\Pi) + \upsilon \tag{IV.4}$$

cannot hold simultaneously.

Assuming this fact for the moment, it is a stone's throw away to prove Theorem 2. Lemma 4.1 ensures that,

with probability tending to one as $n, p \to \infty$, the four terms $\left|\text{TPP}_\lambda(\Pi^\nabla) - \text{tpp}^\infty_\lambda(\Pi^\nabla)\right|$, $\left|\text{FDP}_\lambda(\Pi^\nabla) - \text{fdp}^\infty_\lambda(\Pi^\nabla)\right|$, $|\text{TPP}_{\lambda'}(\Pi) - \text{tpp}^\infty_{\lambda'}(\Pi)|$, and $|\text{FDP}_{\lambda'}(\Pi) - \text{fdp}^\infty_{\lambda'}(\Pi)|$ are all smaller than $\upsilon/2$ for all $c < \lambda, \lambda' < C$. In this event,

$$\text{TPP}_\lambda(\Pi^\nabla) > \text{TPP}_{\lambda'}(\Pi)$$

implies

$$\text{tpp}^\infty_\lambda(\Pi^\nabla) > \text{tpp}^\infty_{\lambda'}(\Pi) - \upsilon$$

and, likewise, $\text{FDP}_\lambda(\Pi^\nabla) < \text{FDP}_{\lambda'}(\Pi)$ implies $\text{fdp}^\infty_\lambda(\Pi^\nabla) < \text{fdp}^\infty_{\lambda'}(\Pi) + \upsilon$. Recognizing that the two inequalities in (IV.4) cannot both hold, therefore, in this event the following inequalities

$$\text{TPP}_\lambda(\Pi^\nabla) > \text{TPP}_{\lambda'}(\Pi) \quad \text{and} \quad \text{FDP}_\lambda(\Pi^\nabla) < \text{FDP}_{\lambda'}(\Pi)$$

cannot hold simultaneously for all $c < \lambda, \lambda' < C$. In words, $\big(\text{TPP}_\lambda(\Pi^\nabla), \text{FDP}_\lambda(\Pi^\nabla)\big)$ does not outperform $(\text{TPP}_{\lambda'}(\Pi), \text{FDP}_{\lambda'}(\Pi))$, and this applies to all $c < \lambda, \lambda' < C$ with probability tending to one.

We conclude the proof by verifying (IV.4). To this end, first find $0 < u_1 < u_2 < 1$ such that the asymptotic powers $\text{tpp}^\infty_\lambda(\Pi^\nabla), \text{tpp}^\infty_{\lambda'}(\Pi)$ are always between u_1 and u_2 for $c < \lambda$, $\lambda' < C$. Next, set

$$\upsilon' := \inf_{u_1 \le u \le u_2} \left(q^\nabla(u) - q^\Pi(u)\right). \tag{IV.5}$$

From Lemma 4.2, we must have $\upsilon' > 0$. Since q^∇ is a continuous function on the closed interval $[0, 1]$, its uniform continuity gives

$$\left|q^\nabla(u) - q^\nabla(u')\right| < \frac{\upsilon'}{2} \tag{IV.6}$$

as long as $|u - u'| \le \upsilon''$ for some $\upsilon'' > 0$.

As the final step, we show that (IV.4) cannot hold simultaneously by taking $\upsilon = \min\{\upsilon'/2, \upsilon''\}$. To see this, suppose we already have $\text{tpp}^\infty_\lambda(\Pi^\nabla) > \text{tpp}^\infty_{\lambda'}(\Pi) - \upsilon$, from which we get

$$\begin{aligned}\text{fdp}^\infty_\lambda(\Pi^\nabla) &= q^\nabla(\text{tpp}^\infty_\lambda(\Pi^\nabla)) \\ &\ge q^\nabla\left(\text{tpp}^\infty_\lambda(\Pi^\nabla) + \upsilon\right) - \frac{\upsilon'}{2} \\ &> q^\nabla\left(\text{tpp}^\infty_{\lambda'}(\Pi)\right) - \frac{\upsilon'}{2}.\end{aligned}$$

Above, the first inequality follows from (IV.6). We proceed by leveraging (IV.5) and obtain

$$\begin{aligned}\text{fdp}^\infty_\lambda(\Pi^\nabla) &> q^\nabla\left(\text{tpp}^\infty_{\lambda'}(\Pi)\right) - \frac{\upsilon'}{2} \\ &\ge q^\Pi\left(\text{tpp}^\infty_{\lambda'}(\Pi)\right) + \upsilon' - \frac{\upsilon'}{2} \\ &= q^\Pi\left(\text{tpp}^\infty_{\lambda'}(\Pi)\right) + \frac{\upsilon'}{2}.\end{aligned}$$

Finally, note that

$$q^\Pi\left(\text{tpp}^\infty_{\lambda'}(\Pi)\right) + \frac{\upsilon'}{2} \ge q^\Pi\left(\text{tpp}^\infty_{\lambda'}(\Pi)\right) + \upsilon = \text{fdp}^\infty_{\lambda'}(\Pi) + \upsilon.$$

Taken together, these calculations reveal that the condition $\text{tpp}^\infty_\lambda(\Pi^\nabla) > \text{tpp}^\infty_{\lambda'}(\Pi) - \upsilon$ implies $\text{fdp}^\infty_\lambda(\Pi^\nabla) > \text{fdp}^\infty_{\lambda'}(\Pi) + \upsilon$. As such, the two inequalities in (IV.4) cannot hold at the same time. This completes the proof. □

The same reasoning in the proof above can be used to prove part (b) of Theorem 2 and Proposition 2.8. More precisely, the first step is to establish the desired result for the deterministic functions $\text{tpp}^\infty_\lambda$ and $\text{fdp}^\infty_\lambda$ using Lemma 4.2, followed by the second step that shows the uniform convergence using Lemma 4.1. In particular, part (b) of Theorem 2 relies on the strictly increasing property of q^∇. Moreover, note that a lower bound on $\text{TPP}_{\lambda'}(\Pi)$ can be translated into an upper bound on λ' [25, Lemma D.1].

Before turning to the proof of Lemma 4.2, we propose the following preparatory lemma.

Lemma 4.3 ([5, Lemma C.1]): For any fixed $\alpha > 0$, define a function $y = f(x)$ in the parametric form

$$\begin{aligned}x(t) &= \mathbb{P}(|t + W > \alpha|) \\ y(t) &= \mathbb{E}(\eta_\alpha(t + W) - t)^2\end{aligned}$$

for $t \ge 0$, where W is a standard normal random variable. Then f is strictly concave.

Proof of Lemma 4.2: We parameterize the curve $(\text{tpp}^\infty_\lambda, \text{fdp}^\infty_\lambda)$ using $\alpha > \alpha_0$. Explicitly, treating α as the free parameter instead of λ, we can solve τ from the AMP equation (IV.1). Define

$$\begin{aligned}\text{fd}^\infty_\alpha(\Pi) &= 2(1 - \epsilon)\Phi(-\alpha) \\ \text{td}^\infty_\alpha(\Pi) &= \epsilon\,\mathbb{P}(|\Pi^\star + \tau W| > \alpha\tau).\end{aligned}$$

This allows us to express the asymptotic power and FDP as functions of α:

$$\begin{aligned}\text{tpp}^\infty_\alpha(\Pi) &= \frac{\text{td}^\infty_\alpha(\Pi)}{\epsilon} \\ \text{fdp}^\infty_\alpha(\Pi) &= \frac{\text{fd}^\infty_\alpha(\Pi)}{\text{fd}^\infty_\alpha(\Pi) + \text{td}^\infty_\alpha(\Pi)}.\end{aligned}$$

To prove Lemma 4.2, for each $\alpha > \alpha_0$, it is sufficient to find a certain value of M such that

$$\text{fd}^\infty_\alpha(\Pi) = \text{fd}^\infty_\alpha(\Pi^\nabla) \quad \text{and} \quad \text{td}^\infty_\alpha(\Pi) > \text{td}^\infty_\alpha(\Pi^\nabla), \tag{IV.7}$$

where Π^∇ is the (ϵ, M)-homogeneous prior (II.7). To see this fact, suppose on the contrary that

$$q^\Pi(u) \ge q^\nabla(u) \tag{IV.8}$$

for some $0 < u < 1$. Let α satisfy $\text{tpp}^\infty_\alpha(\Pi) = u$. From (IV.7) we obtain

$$u = \text{tpp}^\infty_\alpha(\Pi) = \frac{\text{td}^\infty_\alpha(\Pi)}{\epsilon} > \frac{\text{td}^\infty_\alpha(\Pi^\nabla)}{\epsilon} = \text{tpp}^\infty_\alpha(\Pi^\nabla) := u^\nabla \tag{IV.9}$$

and

$$\begin{aligned}\text{fdp}^\infty_\alpha(\Pi) = \frac{\text{fd}^\infty_\alpha(\Pi)}{\text{fd}^\infty_\alpha(\Pi) + \text{td}^\infty_\alpha(\Pi)} &< \frac{\text{fd}^\infty_\alpha(\Pi^\nabla)}{\text{fd}^\infty_\alpha(\Pi^\nabla) + \text{td}^\infty_\alpha(\Pi^\nabla)} \\ &= \text{fdp}^\infty_\alpha(\Pi^\nabla),\end{aligned}$$

which gives

$$q^\Pi(u) = \text{fdp}^\infty_\alpha(\Pi) < \text{fdp}^\infty_\alpha(\Pi^\nabla) = q^\nabla(u^\nabla).$$

This inequality combined with (IV.8) gives

$$q^{\nabla}(u) < q^{\nabla}(u^{\nabla}),$$

which, together with the fact that q^{∇} is an increasing function, leads to $u < u^{\nabla}$. This is a contradiction to (IV.9). Therefore, (IV.8) cannot hold for any $0 < u < 1$.

The remainder of the proof aims to establish (IV.7) by constructing a certain prior Π^{∇}. Explicitly, it suffices to show

$$\mathrm{td}_{\alpha}^{\infty}(\Pi) > \mathrm{td}_{\alpha}^{\infty}(\Pi^{\nabla}) \tag{IV.10}$$

because the equality in (IV.7) holds regardless of the choice of Π^{∇}. To construct Π^{∇}, we first write $\mathrm{td}_{\alpha}^{\infty}(\Pi)$ as

$$\mathrm{td}_{\alpha}^{\infty}(\Pi) = \epsilon \int \mathbb{P}(|t+W| > \alpha) \mathrm{d}\pi(t), \tag{IV.11}$$

where $\mathrm{d}\pi(t)$ denotes the measure of $\Pi^{\star}/\tau$. Since $\mathbb{P}(|t+W| > \alpha)$ is a strictly increasing function of t, there must exist $t' > 0$ such that

$$\mathrm{td}_{\alpha}^{\infty}(\Pi) = \epsilon\, \mathbb{P}(|t'+W| > \alpha). \tag{IV.12}$$

Following (II.7), we let $\Pi^{\nabla} = t'\tau$ with probability ϵ and $\Pi^{\nabla} = 0$ otherwise.

Now, let τ^{∇} denote the solution to (IV.1) given α and Π^{∇}. That is (note that $\sigma = 0$),

$$(1-\epsilon)\, \mathbb{E}\, \eta_{\alpha}(W)^2 + \epsilon\, \mathbb{E}\left(\eta_{\alpha}\left(\frac{t'\tau}{\tau^{\nabla}} + W\right) - \frac{t'\tau}{\tau^{\nabla}}\right)^2 = \delta.$$

Our next step is to show

$$\tau^{\nabla} > \tau.$$

To this end, we invoke Lemma 4.3 and the strict concavity of f gives

$$\begin{aligned}
\mathbb{E}\left(\eta_{\alpha}\Big(\frac{t'\tau}{\tau} + W\Big) - \frac{t'\tau}{\tau}\right)^2 &\equiv f\big(\mathbb{P}(|t'+W| > \alpha)\big) \\
&= f\left(\int \mathbb{P}(|t+W| > \alpha) \mathrm{d}\pi(t)\right) \\
&> \int f\left(\mathbb{P}(|t+W| > \alpha)\right) \mathrm{d}\pi(t) \\
&= \int \mathbb{E}\left(\eta_{\alpha}\left(t+W\right) - t\right)^2 \mathrm{d}\pi(t) \\
&= \mathbb{E}\left(\eta_{\alpha}\left(\frac{\Pi^{\star}}{\tau} + W\right) - \frac{\Pi^{\star}}{\tau}\right)^2,
\end{aligned} \tag{IV.13}$$

where the second equality follows from the definition of t' in (IV.11) and (IV.12), and the inequality is strict because $\Pi^{\star}$ is not constant. Together with the AMP equation for Π

$$(1-\epsilon)\, \mathbb{E}\, \eta_{\alpha}(W)^2 + \epsilon\, \mathbb{E}\left(\eta_{\alpha}\left(\frac{\Pi^{\star}}{\tau} + W\right) - \frac{\Pi^{\star}}{\tau}\right)^2 = \delta,$$

(IV.13) implies

$$(1-\epsilon)\, \mathbb{E}\, \eta_{\alpha}(W)^2 + \epsilon\, \mathbb{E}\left(\eta_{\alpha}\left(\frac{t'\tau}{\tau} + W\right) - \frac{t'\tau}{\tau}\right)^2 > \delta$$

or, equivalently,

$$(1-\epsilon)\, \mathbb{E}\, \eta_{\alpha}(W)^2 + \mathbb{E}\left[\left(\eta_{\alpha}\left(\frac{\Pi^{\nabla}}{\tau} + W\right) - \frac{\Pi^{\nabla}}{\tau}\right)^2 ; \Pi^{\nabla} \neq 0\right] > \delta. \tag{IV.14}$$

By definition, however, τ^{∇} must satisfy

$$(1-\epsilon)\, \mathbb{E}\, \eta_{\alpha}(W)^2 + \mathbb{E}\left[\left(\eta_{\alpha}\left(\frac{\Pi^{\nabla}}{\tau^{\nabla}} + W\right) - \frac{\Pi^{\nabla}}{\tau^{\nabla}}\right)^2 ; \Pi^{\nabla} \neq 0\right] = \delta. \tag{IV.15}$$

A comparison between (IV.14) and (IV.15) immediately gives $\tau^{\nabla} > \tau$.

Having shown $\tau^{\nabla} > \tau$, we complete the proof by noting

$$\begin{aligned}
\mathrm{td}_{\alpha}^{\infty}(\Pi) &= \epsilon\, \mathbb{P}(|t'+W| > \alpha) \\
&> \epsilon\, \mathbb{P}\left(\left|\frac{t'\tau}{\tau^{\nabla}} + W\right| > \alpha\right) \\
&= \epsilon\, \mathbb{P}\left(\left|\Pi^{\nabla} + \tau^{\nabla} W\right| > \alpha\tau^{\nabla} \middle| \Pi^{\nabla} \neq 0\right) \\
&= \mathrm{td}_{\alpha}^{\infty}(\Pi^{\nabla}).
\end{aligned}$$

This verifies (IV.10).

□

B. The Lower Boundary

Now we turn to the proof of Theorem 1. As with the architecture of the proof of Theorem 2, our strategy is to first prove the theorem for the deterministic functions $\mathrm{tpp}_{\lambda}^{\infty}$ and $\mathrm{fdp}_{\lambda}^{\infty}$, and then apply Lemma 4.1 to carry over the results to the random functions TPP_{λ} and FDP_{λ}. Having said this, it is important to note that the proof presents a novel element to the literature. Below, we shall highlight the novel part of the proof of Theorem 1 and leave the rest to the appendix.

As shown in [5], the trade-off curve q^{Π} of any ϵ-sparse prior Π obeys

$$q^{\Pi}(u) > q^{\Delta}(u)$$

for $0 < u < 1$ in both the noiseless and noisy settings, where the curve q^{Δ} is defined in (II.6). If the $(\mathrm{TPP}, \mathrm{FDP})$ pairs from the (ϵ, m, M)-heterogeneous prior Π^{Δ} form the curve q^{Δ} asymptotically as $n, p \to \infty$, the proof of Theorem 1 would follow immediate, just as Theorem 2. For any values of m and M, however, the λ-parameterized curve $(\mathrm{tpp}_{\lambda}^{\infty}(\Pi^{\Delta}), \mathrm{fdp}_{\lambda}^{\infty}(\Pi^{\Delta}))$ does not agree with q^{Δ}. This is in contrast to the proof of Theorem 2, where the $(\mathrm{TPP}, \mathrm{FDP})$ pairs from the (ϵ, M)-homogeneous prior (II.7) converge to the curve q^{∇} for any value of $M \neq 0$, thanks to the assumed noiseless setting.

To tackle this challenge, our strategy is to *uniformly* approximate q^{Δ} using a more general prior for effect sizes that takes the form

$$\Pi^{\Delta}(\boldsymbol{M}, \boldsymbol{\gamma}) = \begin{cases} 0 & \text{w.p. } 1-\epsilon \\ M_1 & \text{w.p. } \epsilon\gamma_1 \\ M_2 & \text{w.p. } \epsilon\gamma_2 \\ M_3 & \text{w.p. } \epsilon\gamma_3 \\ \cdots & \cdots \\ M_m & \text{w.p. } \epsilon\gamma_m, \end{cases} \tag{IV.16}$$

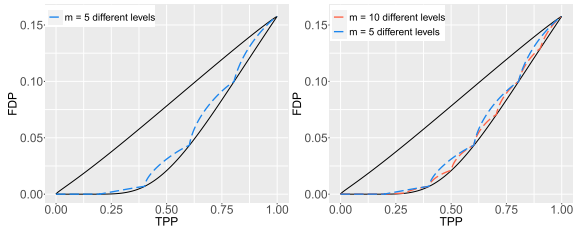

Fig. 4. Illustration of Lemma 4.4, showing the convergence to the lower curve q^{Δ}. Left: $m = 5$ different levels in the prior (IV.16) with $\gamma_1 = \cdots = \gamma_5 = 0.2$, and the associated trade-off curve touches the lower boundary at 4 points; Right: the case $m = 10$ and $\gamma_1 = \cdots = \gamma_{10} = 0.1$ is added as a comparison to the left case.

where $0 < M_1 < M_2 < \cdots < M_m$ and $\gamma_1 + \cdots + \gamma_m = 1$ with $\gamma_i > 0$. Fixing $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ while letting $M_1 \to \infty$ and $M_{i+1}/M_i \to \infty$ for all i, we have the following lemma:

Lemma 4.4: The curve $q^{\Pi^{\Delta}(\boldsymbol{M},\boldsymbol{\gamma})}$ converges to a function that agrees with q^{Δ} at $m-1$ points on $(0, 1)$.

For convenience, denote by $q^{\Delta(\boldsymbol{\gamma})}$ the limiting curve of $q^{\Pi^{\Delta}(\boldsymbol{M},\boldsymbol{\gamma})}$ as $M_1 \to \infty$ and $M_{i+1}/M_i \to \infty$. Figure 4 provides an illustration of this limiting curve. To see why $q^{\Delta(\boldsymbol{\gamma})}$ is close to q^{Δ}, note that Lemma 4.4 ensures that there exist $0 < u_1 < u_2 < \cdots < u_{m-1} < 1$ such that

$$q^{\Delta(\boldsymbol{\gamma})}(u_i) = q^{\Delta}(u_i)$$

for $i = 1, \ldots, m-1$. In fact, the two functions also agree at $u_0 := 0$ and $u_m := 1$. Recognizing that both functions are increasing, for any $u_i \le u \le u_{i+1}$ we get

$$\begin{aligned} 0 \le q^{\Delta(\boldsymbol{\gamma})}(u) - q^{\Delta}(u) &\le q^{\Delta(\boldsymbol{\gamma})}(u_{i+1}) - q^{\Delta}(u_i) \\ &= q^{\Delta}(u_{i+1}) - q^{\Delta}(u_i). \end{aligned}$$

Making use of the uniform continuity of q^{Δ}, the desired conclusion follows if we show that the gaps $u_{i+1} - u_i$ are small for all $i = 0, \ldots, m-1$. The proof of Lemma 4.4, indeed, reveals that this is true if $\max \gamma_i$ is sufficiently small. See the proof of this lemma and the remaining details in Section B.

In passing, we remark that (IV.16) in the special case $m = 2$ has been considered in [5]. Explicitly, the lower boundary q^{Δ} is formed as the lower envelope of the instance-specific trade-off curves induced by the ϵ-sparse priors. See the discussion following (II.3) in Section II.

V. Illustrations

In this section, we present simulation studies to illustrate the impact of effect size heterogeneity beyond the working hypotheses, with a focus on how the impact depends on the design matrix and the noise level.

A. Design Matrix

We perform four simulation studies to examine the impact of effect size heterogeneity on the Lasso method under various synthetic design matrices. Overall, the simulation results show that effect size heterogeneity remains an influential factor in determining the performance of the Lasso far beyond Gaussian designs.

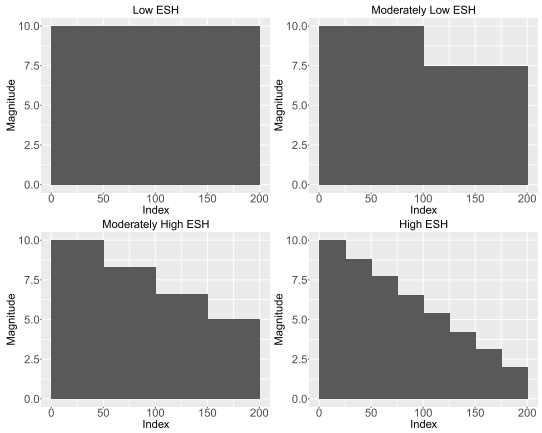

Fig. 5. Four sets of effect sizes ranked in increasing order of their effect size heterogeneity. The corresponding regression coefficients in $\mathbb{R}^{1000}$ with sparsity 200 are used in the experiments of Figures 6, 7, and 10.

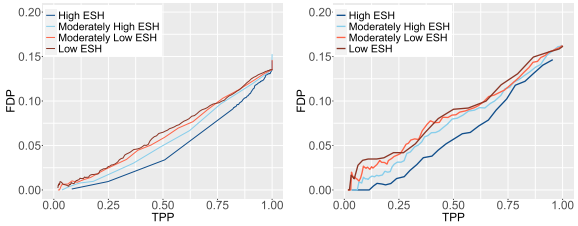

Fig. 6. The TPP–FDP trade-off along the Lasso path under a correlated Gaussian design and a Bernoulli design (Study 1). We set $n = p = 1000$, $k = 200$ and $\sigma = 0$ in both simulations. Left: Gaussian design matrix, each row having covariance $\boldsymbol{\Sigma}$ taking the form $\Sigma_{ij} = 0.5^{|i-j|}$. Right: design matrix with i.i.d. Bernoulli entries taking the value $1/\sqrt{1000}$ or $-1/\sqrt{1000}$ with equal probability. The four sets of regression coefficients are described in Figure 5. The mean FDP is obtained by averaging over 200 replicates.

Study 1. We consider a design matrix of size 1000×1000 that has each row independently drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}/1000$ and another design matrix of the same size that has independent Bernoulli entries, which take the value $1/\sqrt{1000}$ with probability half and otherwise $-1/\sqrt{1000}$. The sparsity is fixed to $k = 200$ while we consider four scenarios of the 200 true effects corresponding to low, moderately low, moderately high, and high effect size heterogeneity (see Figure 5). The results on the TPP–FDP trade-off are presented in Figure 6

Study 2. In this study, we use a dataset of size 1000×892 that is simulated from the admixture of the African-American and European populations, based on the HapMap genotype data [27] (see more details in [28], [29]). The variables can only take $0, 1$, or 2 according to the genotype of a genetic marker. To improve the conditioning of the design matrix, we add i.i.d. $\mathcal{N}(0, 1/1000)$ perturbations to all the entries. Each column is further standardized to have mean 0 and unit norm. We use the effect sizes described in Figure 5 to generate a synthetic response $\boldsymbol{y}$ following the linear model (I.1), with noise $\boldsymbol{z} = \mathbf{0}$. The results are plotted in Figure 7.

Study 3. Working under Gaussian and Bernoulli designs, we now empirically examine the rank of the first false variable. This study considers a varying sparsity level k and sets the

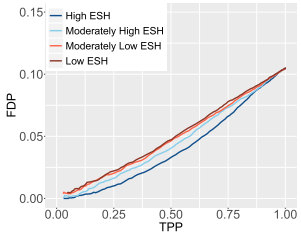

Fig. 7. The TPP–FDP trade-off for the genotype dataset (Study 2). The four curves correspond to the four sets of effect sizes described in Figure 5. The noise term is set to be $\mathbf{0}$. The results are obtained by averaging over 200 replicates.

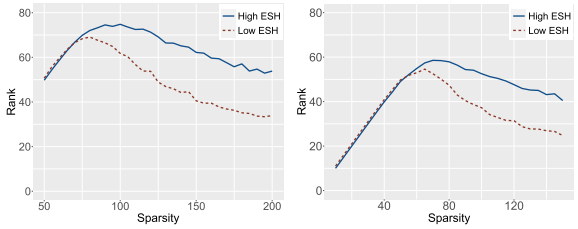

Fig. 8. The rank of the first spurious variable with varying sparsity (Study 3). Left: design matrix of size 1000×1000 consists of i.i.d. $\mathcal{N}(0, \frac{1}{1000})$ entries. Right: design matrix of size 800×1200, with i.i.d. Bernoulli entries that take the value $1/\sqrt{500}$ with probability $1/2$ and value $-1/\sqrt{500}$ otherwise. Each curve is averaged over 200 independent replicates.

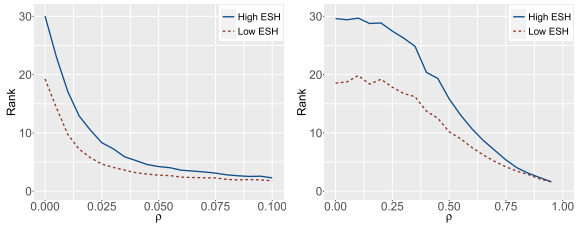

Fig. 9. The rank of the first spurious variable (Study 4). Left: Gaussian design with an equi-correlation covariance matrix, with the non-diagonal correlation ρ varying from 0 to 0.1. Right: Gaussian design with covariance $\mathbf{\Sigma}$ taking the form $\Sigma_{ij} = \rho^{|i-j|}/1000$. Each curve is averaged over 200 independent replicates.

effect sizes to $\beta_j = 100$ for $j = 1, \ldots, k$ (low effect size heterogeneity) or $\beta_j = j$ for $j = 1, \ldots, k$ (high effect size heterogeneity). Each noise component z_i follows $\mathcal{N}(0, 1)$ independently. Figure 8 shows the results under an independent Gaussian random design and an independent Bernoulli design.

Study 4. This scenario uses 500×1000 design matrices that have each row drawn independently from $\mathcal{N}(0, \mathbf{\Sigma})$. In the left panel of Figure 9, the 1000×1000 covariance matrix $\mathbf{\Sigma}$ is set to $\Sigma_{ij} = \rho/1000$ if $i \neq j$ and $\Sigma_{jj} = 1/1000$. In the right panel, the covariance satisfies $\Sigma_{ij} = \rho^{|i-j|}/1000$ for all i, j, with ρ varying from 0 to 0.95. The effect sizes are set to $\beta_j = 100\sqrt{2 \log p}$ for $j \leqslant k$ (low effect size heterogeneity) or, in the low effect size heterogeneity case, the true effect sizes are set to a decreasing sequence from $100\sqrt{2 \log p}$ to 0. The noise $\boldsymbol{z}$ consists of independent standard normal variables. As is clear, both Figure 8 and Figure 9 show that the rank of the first false variable is larger when effect size heterogeneity is high, aligning with our analysis in Section III.

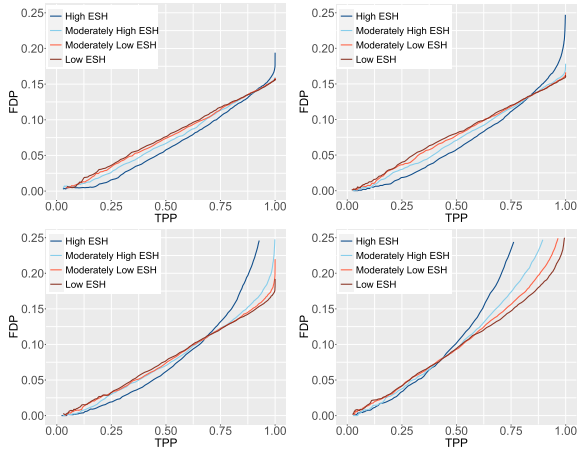

Fig. 10. The TPP–FDP trade-off plot with varying noise levels. The design matrix is specified by $n = p = 1000$, with i.i.d. Gaussian entries. The regression coefficients are from Figure 5, and the noise vector has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, where σ is set to $0.1, 0.2, 0.5$ and 1 in the top-left, top-right, bottom-left, and bottom-right panels, respectively. The mean FDP is obtained by averaging over 100 replicates.

B. Noise Level

While Theorem 2 concerning the regime of low effect size heterogeneity only applies to the noiseless case, we make an attempt to show the impact of effect size heterogeneity in the noisy setting via simulations. Under an independent Gaussian random design of size 1000×1000, we set the nonzero regression coefficients to the four sets of effect sizes as depicted in Figure 5. The noise term $\boldsymbol{z}$ consists of independent $\mathcal{N}(0, \sigma^2)$ entries with $\sigma = 0.1, 0.2, 0.5, 1.0$. The results are displayed in Figure 10.

As with our previous simulation results, higher effect size heterogeneity tends to give rise to a better trade-off between the TPP and FDP from the beginning of the Lasso path. Interestingly, we observe a crossing point in each of the four panels of Figure 10 where higher heterogeneity undergoes a transition from giving a better trade-off down to a worse trade-off. In particular, the crossing point occurs earlier as the noise level σ goes up. While it requires further research to understand this transition in a concrete manner, our observation is that the unselected effect sizes in the late stage of the Lasso path tend to be relatively small compared to the noise level, especially the effect sizes depicted in the bottom-right panel of Figure 10, which have relatively high effect size heterogeneity. Intuitively, this crossing point is where signal-to-noise ratio becomes the dominant factor in place of effect size heterogeneity.

VI. Discussion

In this paper, we have proposed a concept termed effect size heterogeneity for measuring how diverse the nonzero regression coefficients are. Working under Gaussian random designs, we demonstrate that effect size heterogeneity has

a significant impact on model selection consistency of the Lasso when the sparsity is linear in the ambient dimension. In short, we prove that the Lasso attains the optimal trade-off between true and false positives uniformly along its path when the effect sizes are strong and heterogeneous, and attains the worst trade-off when the effects are about the same size in the noiseless case. We also identify similar dependence of the rank of the first noise variable on effect size heterogeneity. While the two theoretical results are proved under certain assumptions, our simulations show that effect size heterogeneity has a significant impact on the Lasso estimate in a much wider range of settings.

Moving forward, this paper opens up several directions for future research. First, it is important to develop methods that incorporate the level of effect size heterogeneity for solving high-dimensional regression problems. In particular, one would be tempted to improve on the Lasso when the level of effect size heterogeneity is low. Interestingly, the SLOPE method has inadvertently addressed this question as its sorted ℓ_1 penalty generally increases as the heterogeneity gets higher [30]–[32]. Another related method developed from a Bayesian angle is the spike-and-slab Lasso procedure [33], which enables the adaptation to a mixture of large and small effects. Nevertheless, it is highly desirable to have methods that leverage effect size heterogeneity more directly. Moreover, a pressing question is to give a quantitative and formal definition of effect size heterogeneity. From a practical standpoint, regression coefficients are seldom exactly zero and thus it might be more appropriate to consider the Type S error, which occurs when a nonzero effect is selected but with the incorrect sign [34], [35]. This reality should prompt one to investigate how effect size heterogeneity interacts with the trade-off between the resulted directional FDP and power. Another question of practical importance is to examine carefully how the impact of effect size heterogeneity depends on the noise level. As an aside, given that Proposition 3.1 remains true for forward stepwise regression and least angle regression, we conjecture that Proposition 3.2 and Theorem 4 also hold for the two model selection procedures. More broadly, it is of interest to investigate whether effect size heterogeneity retains its impact on other ℓ_1 regularized methods such as the two-stage Lasso [19] and the Dantzig selector.

APPENDIX

A. Technical Proofs for Section III

1) Proof of Proposition 3.2:

Proof of Proposition 3.2: We use the "primal-dual witness" argument in the Lasso literature (for example, see Theorem 2 in [24]). As a reminder, here we consider the standard form of Lasso as in (I.2).

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_1$$

with the model specified by (I.1),

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}.$$

We define a pair $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}}) \in \mathbb{R}^p \times \mathbb{R}^p$ to be *primal-dual optimal* if $\widehat{\boldsymbol{\beta}}$ is a minimizer of (I.2), and $\widehat{\boldsymbol{w}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$, satisfying the zero-subgradient condition

$$\boldsymbol{X}^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda\widehat{\boldsymbol{w}} = 0.$$

For the convenience of analysis, we denote $\lambda_n = \frac{\lambda}{n}$. Thus the condition above is equivalent to

$$\frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda_n\widehat{\boldsymbol{w}} = 0. \tag{A.1}$$

By the sufficiency of KKT condition, we know that if there exists some $\widehat{\boldsymbol{w}}$ such that the pair $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}}) \in \mathbb{R}^p \times \mathbb{R}^p$ satisfies (A.1), then $\widehat{\boldsymbol{\beta}}$ is the solution to the Lasso. So $\widehat{\boldsymbol{w}}$ can be seen as a "dual witness" showing $\widehat{\boldsymbol{\beta}}$ is indeed a solution. We are therefore going to construct a "dual witness" vector $\widehat{\boldsymbol{w}}$ to prove a certain $\widehat{\boldsymbol{\beta}}$ is the solution to the Lasso.

To concretely give our construction of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{w}}$, we fix an arbitrary small $\xi > 0$, and then let $s = [(1-\xi)\frac{n-1}{1-\xi+2\log p}]$.[4] Denote $S_0 \equiv \{1, 2, \ldots, s\}$, $S_1 \equiv \{s+1, s+2, .., k\}$, and $S = S_0 \cup S_1$, thus we have $S^C = \{k+1, \ldots, p\}$. Let $M(n) = n^a$ for some $a > \frac{1}{2}$, and let $\lambda_n = n^b$ for some b that satisfies $(k-s)a - 1 < b < (k-s+1)a - \frac{3}{2}$. We omit the dependence of M on n in the following proof. For clarity, for any subset of $T \subset \{1, 2, .., p\}$, we always use the notation $\boldsymbol{w}_T$ to denote the restricted vector $(w_i)_{i\in T}$ of a vector $\boldsymbol{w}$, and the notation $\boldsymbol{X}_T$ to denote the restricted column matrix $(x_{i,j})_{j\in T}$ of a matrix $\boldsymbol{X}$. We consider the following procedure to construct the pair $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}})$,

1) Let $\widehat{\boldsymbol{\beta}}_{S_0^C} = 0$;
2) Solve $(\widehat{\boldsymbol{\beta}}_{S_0}, \widehat{\boldsymbol{w}}_{S_0}) \in \mathbb{R}^s \times \mathbb{R}^s$ from the following oracle sub-problem

$$\widehat{\boldsymbol{\beta}}_{S_0} \in \underset{\boldsymbol{b}\in\mathbb{R}^s}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{S_0}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_1 \right\}, \tag{A.2}$$

   and choose $\widehat{\boldsymbol{w}}_{S_0} \in \partial\|\widehat{\boldsymbol{\beta}}_{S_0}\|_1$ such that

$$\frac{1}{n}{\boldsymbol{X}_{S_0}}^\top(\boldsymbol{X}_{S_0}\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{y}) + \lambda_n\widehat{\boldsymbol{w}}_{S_0} = 0;$$

3) Given $\widehat{\boldsymbol{\beta}}_{S_0}, \widehat{\boldsymbol{w}}_{S_0}$, and $\widehat{\boldsymbol{\beta}}_{S_0^C} = 0$, compute $\widehat{\boldsymbol{w}}_{S_0^C} \in \mathbb{R}^{p-s}$ by equation (A.1), and check whether the *strict dual feasibility* condition $\|\widehat{\boldsymbol{w}}_{S_0^C}\|_\infty < 1$ holds.

The primal-dual witness construction guarantees that if a pair $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}})$ satisfies all the three conditions above, then $\widehat{\boldsymbol{\beta}}$ is the unique solution of the Lasso [24]. Once we prove our construction satisfies the conditions above, the second claim of Proposition 3.2 is an easy corollary as we explicitly require $\widehat{\boldsymbol{\beta}}_j = 0$ for all $j \in S_0^C$, and this gives

$$\#\left\{ j : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0 \right\} = 0.$$

And from this construction, it is also not hard to prove the first claim of the Proposition 3.2. With this protocol in mind, we proceed to prove that we can construct such a pair of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}})$. Now, we solve $\widehat{\boldsymbol{\beta}}_{S_0}, \widehat{\boldsymbol{w}}_{S_0}$ from the subproblem in condition 2. Then, we set $\widehat{\boldsymbol{\beta}}_{S_0^C} = 0$ as in condition 1, and solve $\widehat{\boldsymbol{w}}_{S^C} \in \mathbb{R}^{p-s}$ from (A.1). To prove Proposition 3.2, we only need to prove that with this construction, the strict dual

[4] This is only for technical convenience. One can easily verify that this condition is equivalent to $s = (1 - o_{\mathbb{P}}(1))\frac{2n}{\log p}$.

feasibility condition holds with high probability as $n, p \to \infty$. To prove this, we first simplify condition (A.1) by substituting $\widehat{\boldsymbol{\beta}}_{S_0^C} = 0$, and write it in block matrix form as follows,

$$\frac{1}{n}\begin{bmatrix} \boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0} & \boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_1} & \boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S^C} \\ \boldsymbol{X}_{S_1}{}^\top \boldsymbol{X}_{S_0} & \boldsymbol{X}_{S_1}{}^\top \boldsymbol{X}_{S_1} & \boldsymbol{X}_{S_1}{}^\top \boldsymbol{X}_{S^C} \\ \boldsymbol{X}_{S^C}{}^\top \boldsymbol{X}_{S_0} & \boldsymbol{X}_{S^C}{}^\top \boldsymbol{X}_{S_1} & \boldsymbol{X}_{S^C}{}^\top \boldsymbol{X}_{S^C} \end{bmatrix} \cdot \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0} \\ -\boldsymbol{\beta}_{S_1} \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \boldsymbol{X}_{S_0}{}^\top \boldsymbol{z} \\ \boldsymbol{X}_{S_1}{}^\top \boldsymbol{z} \\ \boldsymbol{X}_{S^C}{}^\top \boldsymbol{z} \end{bmatrix} + \lambda_n \begin{bmatrix} \widehat{\boldsymbol{w}}_{S_0} \\ \widehat{\boldsymbol{w}}_{S_1} \\ \widehat{\boldsymbol{w}}_{S^C} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

or equivalently,

$$\frac{1}{n}\boldsymbol{X}_{S_0}^\top \boldsymbol{X}_{S_0}(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}) - \frac{1}{n}\boldsymbol{X}_{S_0}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} - \frac{1}{n}\boldsymbol{X}_{S_0}^\top \boldsymbol{z} + \lambda_n \widehat{\boldsymbol{w}}_{S_0} = 0, \tag{A.3}$$

$$\frac{1}{n}\boldsymbol{X}_{S_1}^\top \boldsymbol{X}_{S_0}(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}) - \frac{1}{n}\boldsymbol{X}_{S_1}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} - \frac{1}{n}\boldsymbol{X}_{S_1}^\top \boldsymbol{z} + \lambda_n \widehat{\boldsymbol{w}}_{S_1} = 0, \tag{A.4}$$

$$\frac{1}{n}\boldsymbol{X}_{S^C}^\top \boldsymbol{X}_{S_0}(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}) - \frac{1}{n}\boldsymbol{X}_{S^C}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} - \frac{1}{n}\boldsymbol{X}_{S^C}^\top \boldsymbol{z} + \lambda_n \widehat{\boldsymbol{w}}_{S^C} = 0. \tag{A.5}$$

By (A.3), we have

$$\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0} = \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1}\left[\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{X}_{S_0}{}^\top \boldsymbol{z}\right] - \lambda_n n \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0}. \tag{A.6}$$

By substituting (A.6) into (A.4) and (A.5), we can solve $\widehat{w}_j$ for any $j \in S_0^C$ as

$$\begin{aligned} &\widehat{w}_j \\ =&-\frac{1}{\lambda_n n} X_j{}^\top \boldsymbol{X}_{S_0}(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}) + \frac{1}{\lambda_n n}\left[X_j{}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + X_j{}^\top \boldsymbol{z}\right] \\ =&X_j{}^\top \boldsymbol{X}_{S_0}\left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1}\widehat{\boldsymbol{w}}_{S_0} \\ &-\frac{1}{\lambda_n n} X_j{}^\top \boldsymbol{X}_{S_0}\left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1}\left[\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{X}_{S_0}{}^\top \boldsymbol{z}\right] \\ &+\frac{1}{\lambda_n n}\left[X_j{}^\top \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + X_j{}^\top \boldsymbol{z}\right] \\ =&\underbrace{X_j{}^\top \boldsymbol{X}_{S_0}\left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1}\widehat{\boldsymbol{w}}_{S_0}}_{v_j} + \underbrace{X_j{}^\top \mathcal{P}_{S_0^\perp}\left[\frac{\boldsymbol{z}}{\lambda_n n} + \boldsymbol{X}_{S_1}\frac{\boldsymbol{\beta}_{S_1}}{\lambda_n n}\right]}_{u_j}, \end{aligned} \tag{A.7}$$

where $\mathcal{P}_{S_0^\perp} = \boldsymbol{I} - \boldsymbol{X}_{S_0}\left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1}\boldsymbol{X}_{S_0}{}^\top$. As mentioned previously, our goal is to show the strict dual feasibility condition $\max_{j \in S_0^C} |\widehat{w}_j| < 1$ holds with high probability. We will prove it by analyzing the two terms u_j and v_j separately. Specifically, we prove that $v_j < 1 - \frac{\xi}{16}$ and $u_j \to 0$ with high probability.

Denote $M_n = \frac{1}{n}\widehat{\boldsymbol{w}}_{S_0}^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1}\widehat{\boldsymbol{w}}_{S_0}$. Conditioning on the event $E = \{\widehat{\boldsymbol{w}}_{S_0} = \operatorname{sgn}(\boldsymbol{\beta}_{S_0})\}$ and its complement gives us

$$\mathbb{P}\left(\max_{j \in S_0^C}|v_j| \geq c\right) \leq \mathbb{P}\left(\max_{j \in S_0^C}|v_j| \geq c \,\middle|\, E\right) + \mathbb{P}(E^C).$$

It can be seen through Lemma 1.10 that the second term of the last display tends to zero; For the first term, we let $T(\vartheta)$ denote the event $\{|M_n - \mathbb{E}M_n| \geq \vartheta \mathbb{E}M_n\}$. Similar as before, conditioning on the event $T(\vartheta)$ and its complement gives for any $c \in (0,1)$,

$$\mathbb{P}\left(\max_{j \in S_0^C}|v_j| \geq c \middle| E\right) \leq \mathbb{P}\left(\max_{j \in S_0^C}|v_j| \geq c \,\middle|\, T(\vartheta)^C \cap E\right) + \mathbb{P}(T(\vartheta) \cap E).$$

By Lemma 1.2 and Lemma 1.4, the second term in the last display goes to 0 as $n \to \infty$ faster than $\frac{2}{\vartheta^2(n-s-3)}$. And for the first term, we tackle it by considering $\max_{j \in S_0^C} v_j$ and $\min_{j \in S_0^C} v_j$ separately. Denote the event $T = T(\vartheta)^C \cap E$ for convenience, we have

$$\mathbb{P}\left(\max_{j \in S_0^C} v_j \geq c \middle| T\right) \leq \mathbb{P}\left(\max_{j \in S_0^C} \widetilde{v}_j \geq c\right),$$

where $\widetilde{v}_j$ are i.i.d. from $\mathcal{N}(0, (1+\vartheta)\mathbb{E}[M_n|E]) = \mathcal{N}(0, (1+\vartheta)\frac{s}{n-s-1})$. This inequality follows from Lemma 1.9, which states that the probability of the event $\{\max_{i \in S_0^C} v_i \geq c\}$ increases as the mean and variance of v_i increase for Gaussian variables v_i. Given the event T, the maximum variance of v_j is simply $(1+\vartheta)\mathbb{E}[M_n|E]$, and thus we have the inequality above. Set $c = \varpi + \mathbb{E}\max_{j \in S_0^C} \widetilde{v}_j$. From Lemma 1.5, we have

$$\begin{aligned} &\mathbb{P}\Big(\max_{j \in S_0^C} v_j \geq c \Big| T\Big) \leq \mathbb{P}\Big(\max_{j \in S_0^C} \widetilde{v}_j > \varpi + \mathbb{E}\max_{j \in S_0^C} \widetilde{v}_j\Big) \\ \leq&(p-s)\exp\left(-\frac{\varpi^2}{2(1+\vartheta)\mathbb{E}[M_n|E]}\right). \end{aligned}$$

A similar argument for $\min_{j \in S_0^C} \widetilde{v}_j$ gives us

$$\begin{aligned} &\mathbb{P}\Big(\min_{j \in S_0^C} v_j < -c \Big| T\Big) = \mathbb{P}\Big(-\min_{j \in S_0^C} v_j > \varpi + \mathbb{E}\max_{j \in S_0^C} \widetilde{v}_j \Big| T\Big) \\ \leq&(p-s)\exp\left(-\frac{\varpi^2}{2(1+\vartheta)\mathbb{E}[M_n|E]}\right). \end{aligned}$$

Combining the two inequalities above yields

$$\begin{aligned} &\mathbb{P}\left(\max_{j \in S_0^C}|v_j| > \varpi + \mathbb{E}\max_{j \in S_0^C} \widetilde{v}_j \middle| T\right) \\ \leq&2(p-s)\exp\left(-\frac{\varpi^2}{2(1+\vartheta)\mathbb{E}[M_n|E]}\right) \to 0. \end{aligned} \tag{A.8}$$

By Lemma 1.4 and (A.8), we get

$$\mathbb{P}\left(\max_{j \in S_0^C}|v_j| \geq \varpi + \mathbb{E}\max_{j \in S_0^C} \widetilde{v}_j\right) \leq \frac{2}{\vartheta^2(n-s-3)} + 2(p-s)\exp\left(-\frac{\varpi^2(n-s-1)}{2s(1+\vartheta)}\right). \tag{A.9}$$

With the relation of $s = [(1-\xi)\frac{n-1}{1-\xi+2\log p}]$, we can set $\varpi = \frac{\xi}{8}, \vartheta = \frac{(1-\frac{\xi}{4})^2}{1-\xi} - 1 > 0$, and obtain

$$\begin{aligned}
&\varpi + \sqrt{(1+\vartheta)\frac{s}{n-s-1}2\log(p-s)} \\
\leq &\varpi + \sqrt{(1+\vartheta)\frac{(1-\xi)\frac{n-1}{1-\xi+2\log p}+1}{n-(1-\xi)\frac{n-1}{1-\xi+2\log p}-1}2\log p} \\
= &\varpi + \sqrt{(1+\vartheta)\frac{(1-\xi)(n-1)+(1-\xi+\log p)}{2(n-1)\log p}2\log p} \\
= &\varpi + \sqrt{(1+\vartheta)\left((1-\xi)+\frac{(1-\xi+\log p)}{(n-1)}\right)} \\
\leq &\frac{\xi}{8} + (1-\frac{\xi}{4}) + \frac{C}{n-1} \\
< &1 - \frac{\xi}{16} \quad \text{for some large } n. \qquad \text{(A.10)}
\end{aligned}$$

Combining this with the well-known fact of the expectation of the maximum of i.i.d. Gaussian variables that $\mathbb{E}\max_{j\in S_0^C}|\widetilde{v}_j| \leq \sqrt{(1+\vartheta)\frac{s}{n-s-1}2\log(p-s)}$ and (A.9), we know

$$\begin{aligned}
\mathbb{P}(\max_{j\in S_0^C}|v_j| \geq 1-\frac{\xi}{16}) \leq &\frac{2}{\vartheta^2(n-s-3)} \\
&+2(p-s)\exp\left(-\frac{\varpi^2(n-s-1)}{2s(1+\vartheta)}\right) + \mathbb{P}(E^C). \qquad \text{(A.11)}
\end{aligned}$$

This bound is good enough for our purpose. We now proceed to obtain a similar bound of u_j.

By the Cauchy–Schwarz inequality, we have

$$|u_j| \leq \left\|{X_j}^\top \mathcal{P}_{S_0^\perp}\right\| \cdot \left\|\frac{\boldsymbol{z}}{\lambda_n n} + \boldsymbol{X}_{S_1}\frac{\boldsymbol{\beta}_{S_1}}{\lambda_n n}\right\|. \qquad \text{(A.12)}$$

Therefore, we can bound $|u_j|$ if we can control the two norms in (A.12) separately. Because all eigenvalues of $\mathcal{P}_{S_0^\perp}$ are less than 1, we have

$$\left\|{X_j}^\top \mathcal{P}_{S_0^\perp}\right\|^2 \leq \|X_j\|^2 = \sum_{i=1}^n \mathbf{X}_{ij}^2 \overset{\mathcal{D}}{=} \frac{1}{n}\sum_{i=1}^n W_i^2,$$

where $W_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. The summation $\sum_{i=1}^n W_i^2$ is thus a χ^2-distribution with degree of freedom n. By Lemma 1.6, for any $t_1 > 0$, we have

$$\mathbb{P}\left(\left|\frac{\|X_j\|^2}{n} - 1\right| \geq t_1\right) \leq 2\exp(-nt_1^2/8).$$

Combining this with $\left\|{X_j}^\top \mathcal{P}_{S_0^\perp}\right\| \leq \|X_j\|$ gives us

$$\|{X_j}^\top \mathcal{P}_{S_0^\perp}\|_2 \leq \sqrt{1+t_1}, \quad \text{w.p. } \geq 1 - 2\exp(-nt_1^2/8). \qquad \text{(A.13)}$$

Finally, let $\widetilde{u}_j$ denote

$$\widetilde{u}_j = \left\|\frac{\boldsymbol{z}}{\lambda_n n} + \boldsymbol{X}_{S_1}\frac{\boldsymbol{\beta}_{S_1}}{\lambda_n n}\right\|.$$

It is easy to realize that

$$e_i \cdot (\boldsymbol{z} + \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1}) \sim \mathcal{N}(0, \sigma'^2), \quad \text{for all } 1 \leq i \leq s,$$

where $e_i \in \mathbb{R}^n$ is the i-th standard unit vector and $\sigma'^2 = \sigma^2 + \frac{M^{2(k-s+1)}-1}{n(M^2-1)}$. By easy calculation, we obtain

$$\mathbb{E}(\|\widetilde{u}_j\|^2) = n \cdot \frac{\sigma'^2}{\lambda_n^2 n^2}.$$

Applying Lemma 1.6 with $Z_i = e_i \cdot (\boldsymbol{z} + \boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1})$ again, we know for any $t_2 \geq 0$

$$\mathbb{P}\left(\left|\frac{\|\widetilde{u}_j\|^2}{\mathbb{E}(\|\widetilde{u}_j\|^2)} - 1\right| \geq t_2\right) \leq 2\exp(-nt_2^2/8),$$

which is equivalent to

$$\|\widetilde{u}_j\| \leq \sqrt{1+t_2}\cdot\frac{\sigma'}{\lambda_n\sqrt{n}}, \quad \text{w.p. } \geq 1 - 2\exp(-nt_2^2/8). \qquad \text{(A.14)}$$

Using (A.12) by combining the two bounds (A.13) and (A.14), we get

$$\begin{aligned}
&\mathbb{P}(\max_{j\in S_0^C}|u_j| \geq \sqrt{1+t_1}\sqrt{1+t_2}\cdot\frac{\sigma'}{\lambda_n\sqrt{n}}) \\
&\qquad \leq 2(p-s)(\exp(-nt_1^2/8) + 2\exp(-nt_1^2/8)). \qquad \text{(A.15)}
\end{aligned}$$

Now, we can set $M = n^a$ for some $a > \frac{1}{2}$, $\lambda_n = n^b$ for some b that satisfies $(k-s)a - 1 < b < (k-s+1)a - \frac{3}{2}$, and $t_1 = t_2 = 1$ to obtain

$$\begin{aligned}
\sqrt{1+t_1}\sqrt{1+t_2}\cdot\frac{\sigma'}{\lambda_n\sqrt{n}} = &2\frac{\sqrt{\sigma^2 + \frac{M^{2(k-s+1)}-1}{n(M^2-1)}}}{\lambda_n\sqrt{n}} \\
\lesssim &2\ n^{(k-s)a-1-b} \to 0. \qquad \text{(A.16)}
\end{aligned}$$

Particularly, when n is large enough, $\sqrt{1+t_1}\sqrt{1+t_2}\cdot\frac{\sigma'}{\lambda_n\sqrt{n}}$ is less than $\frac{\xi}{32}$, which in turn gives

$$\mathbb{P}(\max_{j\in S_0^C}|u_j| \geq \frac{\xi}{32}) \lesssim 2(p-s)(\exp(-nt_1^2/8) + 2\exp(-nt_1^2/8)). \qquad \text{(A.17)}$$

And thus by a union bound and then (A.7, A.11, A.17), we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{j\in S_0^C}|\widehat{w}_j| \geq 1 - \frac{\xi}{32}\right) \\
\leq &\mathbb{P}\left(\max_{j\in S_0^C}|v_j| \geq 1 - \frac{\xi}{16}\right) + \mathbb{P}\left(\max_{j\in S_0^C}|u_j| \geq \frac{\xi}{32}\right) + \mathbb{P}(E^C) \\
\lesssim &\frac{2}{\vartheta^2(n-s-3)} + 2(p-s)\left(\exp\left(-\frac{\varpi^2(n-s-1)}{2s(1+\vartheta)}\right)\right. \\
&\left. + \exp(-nt_2^2/8) + \exp(-nt_1^2/8)\right). \qquad \text{(A.18)}
\end{aligned}$$

We observe that as $n \to \infty$,

$$\begin{aligned}
&\frac{2}{\vartheta^2(n-s-3)} + 2(p-s)\left(\exp\left(-\frac{\varpi^2(n-s-1)}{2s(1+\vartheta)}\right)\right. \\
&\qquad \left. + \exp(-nt_2^2/8) + \exp(-nt_1^2/8)\right) \to 0, \qquad \text{(A.19)}
\end{aligned}$$

which simply implies that as $n \to \infty$,

$$\mathbb{P}(\max_{j \in S_0^C} |\widehat{w}_j| \geq 1) \to 0.$$

Thus, we have proven under our construction, strict dual feasibility holds. And as we pointed out at the beginning of the proof, the second part of the proposition 3.2 holds, since we set $\widehat{\boldsymbol{\beta}}_j = 0$ for any $j \in S_0^C$. Therefore, we obtain

$$\#\left\{j : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\right\} = 0.$$

Now, we proceed to prove the first part of the proposition, that is,

$$\#\left\{j : \widehat{\beta}_j(\lambda) \neq 0, \beta_j \neq 0\right\} = s = (1 - o_{\mathbb{P}}(1))\frac{n}{2\log p}.$$

Observe that the second equality is due to our assumption on s. So we only need to prove the first equality, that is, for all $j \in S_0$, $\widehat{\beta}_j$ are non-zero, and thus the total number of non-zero $\widehat{\boldsymbol{\beta}}$'s is exactly s. To show this, observe that if $\beta_j - \widehat{\beta}_j < \beta_j$, it is clear that $\widehat{\beta}_j > 0$. Therefore, it suffices to show the following inequality

$$\max_{j \in S_0}(\beta_j - \widehat{\beta}_j) < \min_{j \in S_0} \beta_j = M^{k-s+1} \equiv \rho$$

holds with probability tending to 1. We denote

$$Y_i = -e_i^\top \cdot \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \boldsymbol{X}_{S_0}{}^\top \left[\boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{z}\right] + e_i \cdot \lambda_n n \left(\frac{\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}}{n}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0}, \quad \text{(A.20)}$$

where $e_i \in \mathbb{R}^n$ is the i-th standard unit vector. By (A.6), we know

$$\max_{j \in S_0}(\beta_j - \widehat{\beta}_j) = \max_{1 \leq i \leq n} Y_i.$$

So it is equivalent to show that $\max_{1 \leq i \leq n} Y_i \geq \rho$ holds with probability tending to zero. By Lemma 1.7, we know for $E_i = \mathbb{E}(Y_i \big| \boldsymbol{X}_{S_0})$, and $V_i = \mathrm{Var}(Y_i \big| \boldsymbol{X}_{S_0})$, the event

$$A \equiv \bigcup_{i=1}^{s} \left\{|E_i| \geq (1 + \sqrt{n})|\,\mathbb{E}(E_i)|, \text{ or } |V_i| \geq 2\,\mathbb{E}(V_i)\right\}$$

has probability

$$\mathbb{P}(A) \leq \frac{sK}{n - s} \to 0, \text{ as } n \to \infty.$$

By conditioning on the event A and its complement, we have

$$\mathbb{P}(\max_{i \in S_0} Y_i \geq \rho) \leq \mathbb{P}(\max_{i \in S_0} Y_i \geq \rho \big| A^C) + \mathbb{P}(A) \leq \mathbb{P}(\max_{i \in S_0} \widetilde{Y}_i \geq \rho) + \frac{K}{\frac{n}{s} - 1},$$

where $\widetilde{Y}_i \overset{i.i.d.}{\sim} \mathcal{N}((1 + \sqrt{n})\,\mathbb{E}(E_i), 2\,\mathbb{E}(V_i))$ and the second inequality used the fact in Lemma 1.9 that the probability of the event $\{\max_{i \in S_0} Y_i \geq \rho\}$ increases as the mean and variance increase as long as the mean is less than ρ, which can be directly verified by

$$(1 + \sqrt{n})\,\mathbb{E}(E_i) = (1 + \sqrt{n})\frac{\lambda_n n^2}{n - s - 1} < \rho. \quad \text{(A.21)}$$

Markov's inequality then gives us

$$\begin{aligned}\mathbb{P}(\max_{i \in S_0} \widetilde{Y}_i \geq \rho) &\leq \frac{1}{\rho}\,\mathbb{E}\left(\max_{i \in S_0} |\widetilde{Y}_i|\right) \\ &\leq \frac{1}{\rho}\left(\mathbb{E}(\widetilde{Y}_i) + \mathbb{E}\left(\max_{i \in S_0} |\widetilde{Y}_i - \mathbb{E}(\widetilde{Y}_i)|\right)\right) \\ &\leq \frac{1}{\rho}\left(\frac{(1 + \sqrt{n})\lambda_n n^2}{n - s - 1} + 3\sqrt{\frac{2\sigma'^2 \log s}{n - s - 1}}\right), \quad \text{(A.22)}\end{aligned}$$

where the last inequality uses the bound on Gaussian maxima in Lemma 1.8. By the relation in (A.21), we can easily verify that the probability in (A.22) converges to zero under our conditions of $M = n^a$, $\rho = M^{k-s+1}$ and $\lambda_n = n^b$, where a and b satisfy $a > \frac{1}{2}$ and $(k - s)a - 1 < b < (k - s + 1)a - \frac{3}{2}$, as $n \to \infty$. □

2) Miscellaneous Lemmas for Section III: We first state a well-known result in the random matrix theory (see, for example, Theorem 5.2 in [36]) that we use in the proof of Theorem 4. Then we list all the necessary lemmas for proving Proposition 3.2.

Lemma 1.1: Under the working assumptions, for any deterministic $1 \leq m \leq p/2$, the matrix spectrum norm $\|\cdot\|_2$ satisfies

$$\max_{|S| \leq m} \left\|\boldsymbol{X}_S^\top \boldsymbol{X}_S - \boldsymbol{I}\right\|_2 \leq C\sqrt{\frac{m \log(p/m)}{n}}$$

with probability $1 - 1/p^2$, where C is a universal constant and T is any set of column indices.

Lemma 1.2: For v_j defined in (A.7), and any $i, j \in S_0^C$, we have the following facts,

1) $\mathbb{E}(v_j | \boldsymbol{X}_{S_0}) = 0$;
2) $\mathrm{Var}(v_j | \boldsymbol{X}_{S_0}) = \frac{1}{n}\widehat{\boldsymbol{w}}_{S_0}^\top (\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0})^{-1} \widehat{\boldsymbol{w}}_{S_0}$;
3) $\mathrm{Cov}(v_j, v_i | \boldsymbol{X}_{S_0}) = 0$, if $i \neq j$.

Proof of Lemma 1.2:

Because $j \in S_0^C$, $X_j \perp\!\!\!\perp \boldsymbol{X}_{S_0}$ and $v_j = X_j{}^\top \boldsymbol{X}_{S_0} t(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0} t)^{-1} \widehat{\boldsymbol{w}}_{S_0}$, fact 1 follows from X_j being a centered Gaussian variable.

For fact 2 and fact 3, we observe

$$\begin{aligned}&\mathrm{Cov}(V_j, V_i | \boldsymbol{X}_{S_0}) \\ =&\mathbb{E}\Bigg(\widehat{\boldsymbol{w}}_{S_0}^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \boldsymbol{X}_{S_0}{}^\top X_j \mathbf{X}_i^\top \boldsymbol{X}_{S_0} \\ &\left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0} \Big| \boldsymbol{X}_{S_0}\Bigg) \\ =&\widehat{\boldsymbol{w}}_{S_0}^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \boldsymbol{X}_{S_0}{}^\top \mathbb{E}\left(X_j \mathbf{X}_i^\top \big| \boldsymbol{X}_{S_0}\right) \\ &\boldsymbol{X}_{S_0} \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0} \\ =&\begin{cases} \frac{1}{n}\widehat{\boldsymbol{w}}_{S_0}^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}\end{aligned}$$

□

Lemma 1.3: Consider $\boldsymbol{X}_{S_0} \in \mathbb{R}^{n \times s}$, and suppose each of its column $\mathbf{X}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{s \times s}$ is positive definite. Then $\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}$ is a Wishart distribution of degree of freedom n, and $(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0})^{-1}$ is the inverse Wishart distribution, with expectation and variance

1) $\mathbb{E}({\boldsymbol{X}_{S_0}}^\top \boldsymbol{X}_{S_0})^{-1} = \frac{\boldsymbol{\Sigma}^{-1}}{n-s-1}$;
2) $\mathrm{Var}[(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0})^{-1}_{i,j}] = \frac{(n-s+1)(\Sigma^{-1}_{i,j})^2+(n-s-1)\Sigma^{-1}_{i,i}\Sigma^{-1}_{j,j}}{(n-s)(n-s-1)^2(n-s-3)}$.

Proof of Lemma 1.3: See for example Lemma 7.7.1 of [37] and the formula for the second moment of the inverse Wishart matrices in [38]. □

Lemma 1.4: Let $M_n = \frac{1}{n}\widehat{\boldsymbol{w}}_{S_0}^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0}$. Conditioned on the event E, that is, $\widehat{\boldsymbol{w}}_{S_0} = \mathrm{sgn}(\boldsymbol{\beta}_{S_0})$, we have the following facts:

1) $\mathbb{E}(M_n|E) = \frac{s}{n-s-1}$;
2) $\mathrm{Var}(M_n|E) = \frac{2s^2}{(n-s-1)^2(n-s-3)}$;
3) $\forall \vartheta > 0$, $\mathbb{P}[|M_n - \mathbb{E}(M_n)| \geq \vartheta \mathbb{E}(M_n)|E] \leq \frac{2}{\vartheta^2(n-s-3)}$.

Proof of Lemma 1.4: Observe that $\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}$ follows the Wishart distribution with variance $\frac{1}{n}\boldsymbol{I}_{S_0}$, and thus by Lemma 1.3, the matrix $(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0})^{-1}$ is the inverse Wishart distribution with mean

$$\mathbb{E}(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0})^{-1} = \frac{n}{n-s-1}\boldsymbol{I}_{S_0}. \tag{A.23}$$

Notice that $\widehat{w}_i = \pm 1$ for all $i \in S_0$, and when conditioned on E, it is equal to $\mathrm{sgn}(\boldsymbol{\beta}_{S_0})$, which is independent of X_{S_0}. Therefore, we have

$$\mathbb{E}(M_n|E) = \frac{1}{n}\frac{n}{n-s-1}\widehat{\boldsymbol{w}}_{S_0}^\top \boldsymbol{I}_{S_0}\widehat{\boldsymbol{w}}_{S_0} = \frac{s}{n-s-1}.$$

To calculate the second moment of the inverse Wishart matrices ([38]), we have that for $n-s-3>0$,

$$\begin{aligned}\mathbb{E}(M_n^2|E) =& \frac{1}{n^2}\frac{1}{(n-s)(n-s-3)}(n\cdot\widehat{\boldsymbol{w}}_{S_0}^\top \boldsymbol{I}_{S_0}\widehat{\boldsymbol{w}}_{S_0})^2\frac{n-s}{n-s-1}\\ =& \frac{s^2}{(n-s-1)(n-s-3)}.\end{aligned}$$

Therefore, combining the two equations above, we obtain

$$\begin{aligned}\mathrm{Var}(M_n|E) =& \frac{s^2}{(n-s-1)(n-s-3)} - \frac{s^2}{(n-s-1)^2}\\ =& \frac{2s^2}{(n-s-1)^2(n-s-3)}.\end{aligned}$$

For the third statement, Markov's inequality gives us

$$\begin{aligned}&\mathbb{P}(|M_n - \mathbb{E}(M_n)| \geq \vartheta \mathbb{E}(M_n|E) \leq \frac{\mathrm{Var}(M_n|E)}{\vartheta^2(\mathbb{E}(M_n|E))^2}\\ =& \frac{\frac{2s^2}{(n-s-1)^2(n-s-3)}}{\vartheta^2\frac{s^2}{(n-s-1)^2}} = \frac{2}{\vartheta^2(n-s-3)}.\end{aligned}$$

□

Lemma 1.5: Consider i.i.d. Gaussian random variables $z_j \sim \mathcal{N}(0,\sigma^2)$, where $j = 1,\ldots,l$ for some $l \geq 2$. We have for any $\varpi > 0$,

$$\mathbb{P}\left(\max_{1\leq j\leq l} z_j > \varpi + \mathbb{E}\left(\max_{1\leq j\leq l} z_j\right)\right) \leq \mathrm{e}^{-\frac{\varpi^2}{2\sigma^2}}.$$

Proof of Lemma 1.5: By the Gaussian tail bound

$$\mathbb{P}(z_j > \varpi) \leq \frac{\sigma}{\sqrt{2\pi}\varpi}\mathrm{e}^{-\frac{\varpi^2}{2\sigma^2}},$$

and the well-known fact for the expectation of maximum of i.i.d. Gaussian variables

$$\mathbb{E}\left(\max_{1\leq j\leq l} z_j\right) \leq \sigma\sqrt{2\log l},$$

we have the following union bound

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq j\leq l} z_j > \varpi + \mathbb{E}[\max_{1\leq j\leq l} z_j]\right)\\ \leq& l\frac{1}{\sqrt{2\pi}(\frac{\varpi}{\sigma}+\sqrt{2\log l})}\cdot\exp\left(\frac{(\varpi+\sqrt{2\log(l)}\sigma)^2}{2\sigma^2}\right)\\ =& \frac{1}{\sqrt{2\pi}(\frac{\varpi}{\sigma}+\sqrt{2\log l})}\exp\left(-\frac{\varpi^2}{2\sigma^2}\right)\cdot\exp\left(\sqrt{2\log l}\frac{\varpi}{\sigma}\right)\\ \leq& \exp\left(-\frac{\varpi^2}{2\sigma^2}\right)\end{aligned}$$

holds as long as $l \geq 2$. □

Lemma 1.6: Consider $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\theta^2)$, and denote $Z = \sum_{i=1}^n Z_i^2$. For $t>0$, we have the inequality

$$\mathbb{P}\left(\left|\frac{Z}{\mathbb{E}(Z)} - 1\right| \geq t\right) \leq 2\exp(-nt^2/8).$$

Proof of Lemma 1.6: Let $\widetilde{Z}_i$ be defined as

$$\widetilde{Z}_i = \frac{Z_i}{\theta} \overset{i.i.d.}{\sim} \mathcal{N}(0,1).$$

We have

$$\mathbb{E}(Z) = \sum_{i=1}^n \theta^2 = n\theta^2,$$

and therefore

$$\frac{Z}{\mathbb{E}(Z)} = \frac{\sum_{i=1}^n \widetilde{Z}_i}{n} \overset{\mathcal{D}}{=} \frac{\chi^2}{n}.$$

By easy calculation, we obtain

$$\begin{aligned}\mathbb{E}\left(\mathrm{e}^{\lambda(\widetilde{Z}_i^2-1)}\right) =& \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{\lambda\left(z^2-1\right)}\mathrm{e}^{-z^2/2}dz\\ =& \frac{\mathrm{e}^{-\lambda}}{\sqrt{1-2\lambda}} \leq \mathrm{e}^{2\lambda^2}.\end{aligned}$$

This means Z is sub-exponential with parameter $(2,4)$ (definition of a sub-exponential variable is standard, so we refer the reader to, for example, the Definition 2.2 and Example 2.11 in [4]), and thus Z is a sub-exponential variable with parameter $(2\sqrt{n},4)$. By the Bernstein inequality, we obtain

$$\mathbb{P}\left(\left|\frac{Z}{\mathbb{E}(Z)} - 1\right| \geq t\right) \leq 2\exp(-nt^2/8).$$

□

Lemma 1.7: For Y_i defined in (A.20), we have the following facts:

1) Denote $E_i = \mathbb{E}(Y_i|\boldsymbol{X}_{S_0})$, then we have $|\mathbb{E}(Y_i)| = |\mathbb{E}(E_i)| = \frac{\lambda_n n^2}{n-s-1}$;
2) Denote $V_i = \mathrm{Var}(Y_i|\boldsymbol{X}_{S_0})$, then we have $\mathbb{E}(V_i) = 1$;
3) For n sufficiently large, the inequality $\mathbb{P}\left(|E_i| \geq (1+\sqrt{n})|\mathbb{E}(E_i)|, \text{ or } |V_i| \geq 2\,\mathbb{E}(V_i)\right) \leq \frac{K}{n-s}$ holds for some constant K independent of n and s.

Proof of Lemma 1.7: The idea of the following proof is adapted from Lemma 6 of [24].

For part (a), since $\boldsymbol{X}_{S_1} \perp\!\!\!\perp \boldsymbol{X}_{S_0}$ and $\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{X}_{S_0}$, we get

$$E_i = \mathbb{E}(Y_i \big| \boldsymbol{X}_{S_0}) = -\lambda_n n e_i^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0}.$$

Thus, we have

$$\begin{aligned} |\,\mathbb{E}(Y_i)| &= \left|\mathbb{E}\left(-\lambda_n n e_i \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0}\right)\right| \\ &= \left|-\lambda_n n e_i^\top \frac{n}{n-s-1} \boldsymbol{I}_{S_0}^{-1} \widehat{\boldsymbol{w}}_{S_0}\right| \\ &= \frac{\lambda_n n^2}{n-s-1}, \end{aligned}$$

where the second equality is by (A.23) for the mean of the inverse Wishart distribution.

Next, we turn to prove part (b). We observe that each entry of vector $(\boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{z})$ is i.i.d. distributed as $\mathcal{N}(0, \sigma'^2)$, and is independent of $\boldsymbol{X}_{S_0}$, where we denote $\sigma'^2 = \sigma^2 + \frac{M^{2(k-s+1)}-1}{n(M^2-1)}$. So we have

$$\begin{aligned} &\mathrm{Var}(Y_i \big| \boldsymbol{X}_{S_0}) \\ &= \mathbb{E}\left[(e_i^\top \cdot \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \boldsymbol{X}_{S_0}{}^\top \left(\boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{z}\right))^2 \big| \boldsymbol{X}_{S_0}\right] \\ &= e_i^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \boldsymbol{X}_{S_0}{}^\top \mathbb{E}\Big[(\boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{z}) \\ &\quad \cdot (\boldsymbol{X}_{S_1}\boldsymbol{\beta}_{S_1} + \boldsymbol{z})^\top \big| \boldsymbol{X}_{S_0}\Big] \boldsymbol{X}_{S_0} \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} e_i \\ &= \sigma'^2 e_i^T \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} \boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0} \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} e_i \\ &= \sigma'^2 e_i^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} e_i. \end{aligned}$$

Thus by (A.23) again, we obtain

$$\begin{aligned} \mathbb{E}(V_i) &= \mathbb{E}(\sigma'^2 e_i^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} e_i) \\ &= \sigma'^2 e_i^\top \frac{n}{n-s-1} \boldsymbol{I}_{S_0}^{-1} e_i \\ &= \frac{n\sigma'^2}{n-s-1}. \end{aligned}$$

To prove part (c), we use the formula for the second moment of the inverse Wishart distribution in the part (2) of Lemma 1.3. With $E_i = \mathbb{E}(Y_i \big| \boldsymbol{X}_{S_0})$, we get

$$\begin{aligned} \mathbb{E}(E_i^2) &= \mathbb{E}(\mathbb{E}(Y_i \big| \boldsymbol{X}_{S_0}))^2 \\ &= \frac{\lambda_n^2 n^2}{(n-s)(n-s-3)} \left[\left(e_i^\top \left(\frac{1}{n}\boldsymbol{I}_{S_0}\right)^{-1} \widehat{\boldsymbol{w}}_{S_0}\right)^2 \right. \\ &\quad \left. + \frac{n^2}{n-s-1} \left(\widehat{\boldsymbol{w}}_{S_0}^\top \widehat{\boldsymbol{w}}_{S_0}\right)\left(e_i^\top e_i\right)\right] \\ &= \frac{\lambda_n^2 n^2}{(n-s)(n-s-3)} \left[n^2 + \frac{1}{n-s-1} \cdot ns \cdot n\right] \\ &= \frac{\lambda_n^2 n^4 (ns + n - s^2 - 2s + 1)}{(n-s)(n-s-3)(n-s-1)}. \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathrm{Var}(E_i) &= \frac{\lambda_n^2 n^4 (ns + n - s^2 - 2s + 1)}{(n-s)(n-s-3)(n-s-1)} - \frac{\lambda_n^2 n^4}{(n-s-1)^2} \\ &= \frac{\lambda_n^2 n^4 (n-1)}{(n-s)(n-s-3)(n-s-1)}. \end{aligned} \tag{A.24}$$

By Chebyshev's inequality, we can get

$$\begin{aligned} \mathbb{P}(|E_i| \geq (1+\sqrt{n})\,\mathbb{E}(E_i)) &\leq \mathbb{P}(|E_i - \mathbb{E}(E_i)| \geq \sqrt{n}\,\mathbb{E}(E_i)) \\ &\leq \frac{\mathrm{Var}(E_i)}{n(\mathbb{E}(E_i))^2} \\ &= \frac{ns + n - s^2 - 2s + 1}{4n(n-s)(n-s-3)} \\ &\leq \frac{K_1}{n-s}, \end{aligned} \tag{A.25}$$

for some constant K_1 when n is large enough.

Similarly, by Lemma 1.3 (2) again for $i = j$, and $\Sigma = \frac{1}{n}\boldsymbol{I}$, we have

$$\begin{aligned} \mathrm{Var}(V_i^2) &= \sigma'^4 \mathrm{Var}\left[(e_i^\top \left(\boldsymbol{X}_{S_0}{}^\top \boldsymbol{X}_{S_0}\right)^{-1} e_i)^2\right] \\ &= \sigma'^4 \frac{(n-s+1+n-s-1)n^2}{(n-s)(n-s-1)^2(n-s-3)} \\ &= \frac{2\sigma'^4 n^2}{(n-s)(n-s-1)(n-s-3)}, \end{aligned}$$

and thus

$$\begin{aligned} &\mathbb{P}(V_i \geq 2E(V_i)) = \mathbb{P}(V_i - \mathbb{E}(V_i) \geq E(V_i)) \\ &\leq \frac{\mathrm{Var}(V_i)}{(\mathbb{E}(V_i))^2} = \frac{\frac{2\sigma'^4 n^2}{(n-s)(n-s-1)(n-s-3)}}{\left(\frac{n\sigma'^2}{n-s-1}\right)^2} \\ &\leq \frac{K_2}{n-s}, \end{aligned} \tag{A.26}$$

for some constant K_2 for large n. Therefore combining (A.25,A.26) with union bound, the statement in part 3 holds with $K = K_1 + K_2$. □

Lemma 1.8: Let $(X_1, \ldots, X_n)$ be independent and normally distributed. We have

$$\mathbb{E}[\max_{1\leq i\leq n} |X_i|] \leq 3\sqrt{\log n} \max_{1\leq i\leq n} \sqrt{\mathbb{E}\,X_i^2}.$$

Proof of Lemma 1.8: This is a well-known result of Gaussian maxima. We omit its proof and refer the reader to, for example, [24] Lemma 9. □

Lemma 1.9: Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. Suppose $\mu \leq \mu_0$, $\sigma \leq \sigma_0$, and $\rho \geq \mu_0$, then the probability $\mathbb{P}(Y \geq \rho) \leq \mathbb{P}(Z \geq \rho)$, where $Z \sim \mathcal{N}(\mu_0.\sigma_0)$.

Proof of Lemma 1.9: By definition, we have

$$\mathbb{P}(Y \geq \rho) = 1 - \Phi\left(\frac{\rho - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - \rho}{\sigma}\right).$$

And similarly, we have

$$\mathbb{P}(Z \geq \rho) = \Phi\left(\frac{\mu_0 - \rho}{\sigma_0}\right).$$

By the assumption, we know that

$$\frac{\mu - \rho}{\sigma} \leq \frac{\mu_0 - \rho}{\sigma_0},$$

and thus the lemma follows from the fact that Φ is increasing. □

Lemma 1.10: Under the working assumptions, the event $E = \{\widehat{\boldsymbol{w}}_{S_0} = \mathrm{sgn}(\boldsymbol{\beta}_{S_0})\}$ satisfies

$$\mathbb{P}(E^C) = o_n(1).$$

Proof of Lemma 1.10: This is the classic result of Theorem 3 in [24]. For the Lasso problem considered in Equation (A.2), with the identity covariance matrix, all conditions (26a), (26b), (26c) therein are easily satisfied with $C_{\min} = C_{\max} = 1$. As long as we have the condition

$$\min_{j=1,..,s} \beta_j > g(\lambda) = c_3\lambda + 20\sqrt{\frac{\sigma^2 \log s}{n}}, \tag{A.27}$$

for some constant $c_3 > 0$, we can guarantee that $\mathrm{sgn}(\boldsymbol{\beta}_{S_0}) = \mathrm{sgn}(\hat{\boldsymbol{\beta}}_{S_0}) \equiv \widehat{\boldsymbol{w}}_{S_0}$ with probability $1 - c_1\exp(-c_2\min\{s, \log(k-s)\})$ for some constant $c_1, c_2 > 0$.

To verify condition (A.27), recall that $\lambda = n^b$, $s = O(n/\log p)$, and $\min_{j=1,..,s}\beta_i = M(n)^{k-s+1}$. Because $M(n) = n^a$ and $b < (k-s+1)a - \frac{3}{2}$, we have

$$\begin{aligned}\min_{j=1,..,s}\beta_j = M(n)^{k-s+1} &= n^{(k-s+1)a} \\ &\geq n^{b+\frac{3}{2}} \\ &= \lambda n^{\frac{3}{2}} \\ &\geq c_3\lambda + 20\sqrt{\frac{\sigma^2\log s}{n}} = g(\lambda).\end{aligned}$$

for sufficiently large n. This implies $P(E) \to 1$, or $P(E^C) = o_n(1)$. □

B. Technical Proofs for Section IV

1) A Property of FDP and TPP: Any Trade-off Curve Is Strictly Increasing: A natural belief on the pair of (TPP, FDP) is that FDP (type-I error) should increase with TPP (power), which may be strengthened by our simulation plots. However, along a single Lasso path, this is in general not necessarily true. It is well-known that Lasso is not monotone [18], so it is possible that with more and more true variables entering the Lasso path, fewer and fewer noise variables retain in the Lasso path. In such a case, FDP is no longer a monotone function of TPP. This possibility complicates our analysis, yet the following lemma asserts that this possibility is impossible. We prove that the asymptotic FDP is strictly increasing with the asymptotic TPP. Formally speaking, as λ varies, $\mathrm{fdp}_\lambda^\infty$ can be seen as a function of $\mathrm{tpp}_\lambda^\infty$, and $\mathrm{fdp}_\lambda^\infty$ is a strictly increasing function of $\mathrm{tpp}_\lambda^\infty$. To be rigorous, in the following lemma—indeed throughout the paper—we consider the regime below the Donoho–Tanner phase transition. We refer interested readers to [19] for results above this phase transition.

Lemma 1.11: Fix ϵ, δ, σ, and $\Pi \neq 0$. We have that $\mathrm{fdp}_\lambda^\infty(\Pi)$ is a strictly increasing function of $\mathrm{tpp}_\lambda^\infty(\Pi)$. That is, $\mathrm{fdp}_\lambda^\infty(\mathrm{tpp}_\lambda^\infty)$ is a well-defined function, and $\mathrm{fdp}_\lambda^{\infty\prime}(\cdot)\big|_{\mathrm{tpp}_\lambda^\infty} > 0$ for any valid value of $\mathrm{tpp}_\lambda^\infty$.[5]

To prove this lemma, we need the following characterizations among α, λ, $\mathrm{fdp}_\lambda^\infty$ and $\mathrm{tpp}_\lambda^\infty$.

[5]As we will see in Lemma 1.15, the valid range of $\mathrm{tpp}_\lambda^\infty$ is the range $(0, u^\star)$. In this paper, we only focus on the case where $u^\star = 1$.

Lemma 1.12: Fix ϵ, δ, σ, and $\Pi \neq 0$. Consider any α, τ, λ that solve equations (IV.1). We have the following facts

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda} > 0 \tag{A.28}$$

$$\frac{\mathrm{dtpp}^\infty}{\mathrm{d}\alpha} < 0 \tag{A.29}$$

$$\frac{\mathrm{dfdp}^\infty}{\mathrm{d}\alpha} < 0 \tag{A.30}$$

$$\frac{\mathrm{dtpp}^\infty}{\mathrm{dfdp}^\infty} > 0 \tag{A.31}$$

We note that the denotations of fdp^∞ and tpp^∞ stand for $\mathrm{fdp}_\alpha^\infty$ and $\mathrm{tpp}_\alpha^\infty$, where we treat α as the free parameter. We often suppress this dependence on α in the following proof when it is clear from the context, and use the denotations of fdp^∞ and tpp^∞ for simplicity.

Proof of Lemma 1.12: The (A.28) is a well-known result, and one can refer to, for example, Lemma 4.11 of [15] for a proof.

To prove (A.29), we note that $\mathrm{tpp}^\infty = \mathbb{P}(|\Pi^\star + \tau W| > \alpha\tau)$, where $\Pi^\star$ is the distribution of an entry of β given it's not zero. For any Π that is a proper distribution and satisfies (IV.1), proving $\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{P}(|\Pi+\tau W| > \alpha\tau) < 0$ will suffice. And this result follows from Lemma 4.10 of [15]. Now we left to prove (A.30) and (A.31). We note that, however, (A.31) follows directly from (A.30), and therefore we only need to prove (A.30). To see this fact, we note that tpp^∞ is a strictly increasing function of α, so α is also a function of tpp^∞. By the chain rule, we have

$$\frac{\mathrm{dfdp}^\infty}{\mathrm{dtpp}^\infty} = \frac{\mathrm{dfdp}^\infty}{\mathrm{d}\alpha}\cdot\frac{\mathrm{d}\alpha}{\mathrm{dtpp}^\infty}.$$

Note that we have already proven $\frac{\mathrm{dtpp}^\infty}{\mathrm{d}\alpha} < 0$, which implies $\frac{\mathrm{d}\alpha}{\mathrm{dtpp}^\infty} < 0$. Therefore, we only need to show (A.30) is true to prove (A.31).

Now, to prove (A.30), we observe that

$$\mathrm{fdp}^\infty(\alpha) = \frac{1}{1 + \frac{\epsilon\mathbb{P}\left(\left|\frac{\Pi^\star}{\tau}+W\right|>\alpha\right)}{2(1-\epsilon)\Phi(-\alpha)}},$$

and thus

$$\frac{\mathrm{dfdp}^\infty}{\mathrm{d}\alpha} = \frac{\epsilon}{2(1-\epsilon)}\cdot\frac{\frac{\mathrm{d}\left(\frac{\mathbb{P}\left(\left|\frac{\Pi^\star}{\tau}+W\right|>\alpha\right)}{\Phi(-\alpha)}\right)}{\mathrm{d}\alpha}}{\left(1+\frac{\epsilon\mathbb{P}\left(\left|\frac{\Pi^\star}{\tau}+W\right|>\alpha\right)}{2(1-\epsilon)\Phi(-\alpha)}\right)^2}.$$

Since the denominator is positive, we only need to show the numerator is negative. For simplicity, we will abuse the notation a little bit by using Π for $\Pi^\star$ in the rest of the proof. We need to show for all $\Pi \neq 0$,

$$\frac{\mathrm{d}\left(\frac{\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)}{\Phi(-\alpha)}\right)}{\mathrm{d}\alpha} < 0,$$

or equivalently,

$$\frac{\mathrm{d}\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)}{\mathrm{d}\alpha}\cdot\Phi(-\alpha)+\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)\cdot\phi(\alpha) > 0. \tag{A.32}$$

We observe that

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)\\ =\;&\mathbb{E}\left[\mathbb{P}_W\left(W>\alpha-\frac{\Pi}{\tau}\middle|\,\Pi\right)+\mathbb{P}_W\left(W<-\alpha-\frac{\Pi}{\tau}\middle|\,\Pi\right)\right]\\ =\;&\mathbb{E}\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi(-\frac{\Pi}{\tau}-\alpha)\right].\end{aligned}\tag{A.33}$$

Substituting the expressions of (A.33) into (A.32), we obtain, (A.34), as shown at the bottom of the next page.

Note the denominator is positive, so we only need to prove that the numerator is positive. Let $g(u)=(\Phi(u-\alpha)+\Phi(-u-\alpha))\phi(\alpha)-(\phi(u-\alpha)+\phi(u+\alpha))\Phi(-\alpha)$. By Lemma 1.14, $g(u)>0$ for $u\neq 0$, and therefore we have

$$\Omega_1+\Omega_3=\frac{\sigma^2\delta}{\tau^3}\mathbb{E}_\Pi\left[g\left(\frac{\Pi}{\tau}\right)\right]>0.\tag{A.35}$$

For Ω_5, we observe that if $\Pi>0$, then $-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)<0$; and if $\Pi<0$, then $-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)>0$. Therefore, we have

$$\mathbb{E}_\Pi\left[\frac{\Pi}{\tau}\left(-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right)\right]\leq 0$$

So, by the fact that $\Phi(-\alpha)\leq\frac{\phi(\alpha)}{\alpha}$, the definition of $\Phi(x)$, we have

$$\begin{aligned}&\Omega_5=\mathbb{E}_\Pi\left[\frac{\Pi}{\tau^2}\left(-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right)\right]\\ &\cdot\mathbb{E}_\Pi\left[\left(\int_{\alpha-\frac{\Pi}{\tau}}^{\infty}\phi(w)\mathrm{d}w+\int_{-\infty}^{-\alpha-\frac{\Pi}{\tau}}\phi(w)\mathrm{d}w\right)\right]\cdot\alpha\Phi(-\alpha)\\ &\geq\mathbb{E}_\Pi\left[\frac{\Pi}{\tau^2}\left(-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right)\right]\\ &\cdot\mathbb{E}_\Pi\left[\left(\int_{\alpha-\frac{\Pi}{\tau}}^{\infty}\phi(w)\mathrm{d}w+\int_{-\infty}^{-\alpha-\frac{\Pi}{\tau}}\phi(w)\mathrm{d}w\right)\right]\cdot\phi(\alpha)\\ &=\mathbb{E}_\Pi\left[\frac{\Pi}{\tau^2}\left(-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right)\right]\\ &\cdot\mathbb{E}_\Pi\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi\left(-\frac{\Pi}{\tau}-\alpha\right)\right]\cdot\phi(\alpha)\\ &=\mathbb{E}_\Pi\left[\frac{\Pi}{\tau^2}\int_{-\alpha-\frac{\Pi}{\tau}}^{\alpha-\frac{\Pi}{\tau}}w\phi(w)\mathrm{d}w\right]\\ &\cdot\mathbb{E}_\Pi\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi\left(-\frac{\Pi}{\tau}-\alpha\right)\right]\cdot\phi(\alpha).\end{aligned}\tag{A.36}$$

Similarly, by (A.34) and (A.36), and then by the definition of $f(u)$ in Lemma 1.13, we obtain

$$\begin{aligned}&\Omega_2+\Omega_5\geq\mathbb{E}_\Pi\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi\left(-\frac{\Pi}{\tau}-\alpha\right)\right]\\ &\cdot\mathbb{E}_\Pi\left[\frac{\Pi}{\tau^2}\int_{-\alpha-\frac{\Pi}{\tau}}^{\alpha-\frac{\Pi}{\tau}}w\phi(w)\mathrm{d}w+\frac{\Pi^2}{\tau^3}\int_{-\alpha-\frac{\Pi}{\tau}}^{\alpha-\frac{\Pi}{\tau}}\phi(w)\mathrm{d}w\right]\cdot\phi(\alpha)\\ &=\mathbb{E}_\Pi\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi\left(-\frac{\Pi}{\tau}-\alpha\right)\right]\mathbb{E}_\Pi\left[\frac{1}{\tau}f\left(\frac{\Pi}{\tau}\right)\right]\phi(\alpha).\end{aligned}\tag{A.37}$$

Similarly, we have

$$\Omega_4+\Omega_6=-\mathbb{E}_\Pi\left[\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right]\mathbb{E}_\Pi\left[\frac{1}{\tau}f\left(\frac{\Pi}{\tau}\right)\right].\tag{A.38}$$

Combining the last display, (A.37), (A.38), Lemma 1.13 and the well-known result that $\Phi(-\alpha)\leq\frac{\phi(\alpha)}{\alpha}$, we obtain

$$\Omega_2+\Omega_4+\Omega_5+\Omega_6\geq\mathbb{E}_\Pi\left[g\left(\frac{\Pi}{\tau}\right)\right]\cdot\mathbb{E}_\Pi\left[\frac{1}{\tau}f\left(\frac{\Pi}{\tau}\right)\right]>0,\tag{A.39}$$

Put together (A.35), (A.34) and (A.39), we have for all $\Pi\neq 0$

$$\frac{\mathrm{d}\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)}{\mathrm{d}\alpha}\cdot\Phi(-\alpha)+\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)\cdot\phi(\alpha)>0,$$

which, by (A.32), amounts to (A.30), or

$$\frac{\mathrm{dfdp}^\infty}{\mathrm{d}\alpha}<0.$$

Therefore, combining with (A.29), we obtain that

$$\frac{\mathrm{dtpp}^\infty}{\mathrm{dfdp}^\infty}>0.$$

□

Summarizing the result we have proven, it is very easy to prove Lemma 1.11.

Proof of Lemma 1.11: Observe $\mathrm{tpp}^\infty(\alpha)$ is a strictly increasing function of α, and thus tpp^∞ is a one-to-one function of α. The inverse function therefore exists, so α is a strictly increasing function of tpp^∞. Similarly, fdp^∞ is also a strictly increasing function of α. Therefore, we conclude that $\mathrm{fdp}^\infty=\mathrm{fdp}^\infty(\alpha)=\mathrm{fdp}^\infty(\alpha(\mathrm{tpp}^\infty))=\mathrm{fdp}^\infty(\mathrm{tpp}^\infty)$ is a strictly increasing function of tpp^∞, and that $\frac{\mathrm{dfdp}^\infty}{\mathrm{dtpp}^\infty}>0$ holds for any valid value of tpp^∞. □

Now, we prove the lemmas that we have used in the proof of Lemma 1.12.

Lemma 1.13: Let $f(u)=u\int_{-\alpha-u}^{\alpha-u}(w+u)\phi(w)\mathrm{d}w$. We have $f(u)>0$, for all $u\neq 0\in\mathbb{R}$.

Proof of Lemma 1.13: Observe that

$$\begin{aligned}f(u)&=u\int_{-\alpha-u}^{\alpha-u}(w+u)\phi(w)\mathrm{d}w\\ &\overset{w'=w+u}{=}u\int_{-\alpha}^{\alpha}w'\phi(w'-u)\mathrm{d}w'\\ &=u\int_0^\alpha w'[\phi(w'-u)-\phi(-w'-u)]\mathrm{d}w'.\end{aligned}$$

So, if $u>0$, then $\phi(w'-u)-\phi(-w'-u)>0$, for any $w'\in(0,\alpha]$, thus $f(u)>0$; and if $u<0$, then $\phi(w'-u)-\phi(-w'-u)<0$, for any $w'\in(0,\alpha]$, thus $f(u)>0$. □

Lemma 1.14: For any fixed $\alpha>0$, let $g(u)=(\Phi(u-\alpha)+\Phi(-u-\alpha))\phi(\alpha)-(\phi(u-\alpha)+\phi(u+\alpha))\Phi(-\alpha)$. Then we have $g(u)\geq 0$, and the equality $g(u)=0$ holds if and only if $u=0$.

Proof of Lemma 1.14: We observe that

$$\begin{aligned} g(u) =&\phi(\alpha)\Bigg(\int_{\alpha-u}^{\infty}\phi(w)\mathrm{d}w+\int_{\alpha+u}^{\infty}\phi(w)\mathrm{d}w \\ &-\int_{\alpha}^{\infty}\phi(w)\mathrm{d}w\left(\frac{\phi(u-\alpha)}{\phi(\alpha)}+\frac{\phi(u+\alpha)}{\phi(\alpha)}\right)\Bigg) \\ =&\phi(\alpha)\Bigg(\int_{\alpha}^{\infty}\phi(w)\mathrm{e}^{wu}\cdot\mathrm{e}^{\frac{-u^2}{2}}\mathrm{d}w \\ &+\int_{\alpha}^{\infty}\phi(w)\mathrm{e}^{-wu}\cdot\mathrm{e}^{\frac{-u^2}{2}}\mathrm{d}w \\ &-\int_{\alpha}^{\infty}\phi(w)(\mathrm{e}^{\alpha u}+\mathrm{e}^{-\alpha u})\cdot\mathrm{e}^{\frac{-u^2}{2}}\mathrm{d}w\Bigg) \\ =&\phi(\alpha)\mathrm{e}^{\frac{-u^2}{2}}\int_{\alpha}^{\infty}\phi(w)\big[(\mathrm{e}^{wu}+\mathrm{e}^{-wu})-(\mathrm{e}^{\alpha u}+\mathrm{e}^{-\alpha u})\big]\,\mathrm{d}w. \end{aligned}$$

Since for any $w>\alpha>0$, we have

$$\mathrm{e}^{wu}+\mathrm{e}^{-wu}>\mathrm{e}^{\alpha u}+\mathrm{e}^{-\alpha u},\quad \text{for any } u\in\mathbb{R}.$$

We obtain $g(u)\geq 0$, and it is clear the equality holds if and only if $u=0$. □

2) Miscellaneous Proofs for Section IV-A: In this section, we prove all the necessary lemmas for Theorem 2. To start with, we state the following lemma that specifies the range of all valid tpp$^{\infty}$'s.

Lemma 1.15: [Lemma C.1 and Lemma C.4 in [5]] Put

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

Then

$$\mathrm{tpp}^{\infty}<u^{\star}(\delta,\epsilon).$$

$$\begin{aligned} &\frac{\mathrm{d}\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)}{\mathrm{d}\alpha}\cdot\Phi(-\alpha)+\mathbb{P}\left(\left|\frac{\Pi}{\tau}+W\right|>\alpha\right)\cdot\phi(\alpha) \\ &=\frac{1}{\frac{\sigma^2\delta}{\tau^3}+\mathbb{E}_{\Pi,W}\left[\frac{\Pi^2}{\tau^3}1_{\{-\alpha<\frac{\Pi}{\tau}+W<\alpha\}}\right]} \\ &\quad\cdot\Bigg[\overbrace{\mathbb{E}_{\Pi}\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi(-\frac{\Pi}{\tau}-\alpha)\right]\cdot\frac{\sigma^2\delta}{\tau^3}\cdot\phi(\alpha)}^{\Omega_1} \\ &\quad+\overbrace{\mathbb{E}_{\Pi}\left[\Phi\left(\frac{\Pi}{\tau}-\alpha\right)+\Phi(-\frac{\Pi}{\tau}-\alpha)\right]\cdot\mathbb{E}_{\Pi,W}\left[\frac{\Pi^2}{\tau^3}1_{\{-\alpha<\frac{\Pi}{\tau}+W<\alpha\}}\right]\cdot\phi(\alpha)}^{\Omega_2} \\ &\quad\overbrace{-\mathbb{E}_{\Pi}\left[\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right]\cdot\frac{\sigma^2\delta}{\tau^3}\cdot\Phi(-\alpha)}^{\Omega_3} \\ &\quad\overbrace{-\mathbb{E}_{\Pi}\left[\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right]\cdot\mathbb{E}_{\Pi,W}\left[\frac{\Pi^2}{\tau^3}1_{\{-\alpha<\frac{\Pi}{\tau}+W<\alpha\}}\right]\cdot\Phi(-\alpha)}^{\Omega_4} \\ &\quad+\overbrace{\mathbb{E}_{\Pi}\left[\frac{\Pi}{\tau^2}\left(-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right)\right]}^{\Omega_5} \\ &\quad\overbrace{\cdot\mathbb{E}_{\Pi}\left[\alpha\left(\int_{\alpha-\frac{\Pi}{\tau}}^{\infty}\phi(w)\mathrm{d}w+\int_{-\infty}^{-\alpha-\frac{\Pi}{\tau}}\phi(w)\mathrm{d}w\right)\right]\cdot\Phi(-\alpha)}^{\Omega_{5,\text{cont.}}} \\ &\quad+\overbrace{\mathbb{E}_{\Pi}\left[\frac{\Pi}{\tau^2}\left(-\phi\left(\alpha-\frac{\Pi}{\tau}\right)+\phi\left(\alpha+\frac{\Pi}{\tau}\right)\right)\right]}^{\Omega_6} \\ &\quad\overbrace{\cdot\mathbb{E}_{\Pi}\left[\int_{\alpha-\frac{\Pi}{\tau}}^{\infty}-w\phi(w)\mathrm{d}w+\int_{-\infty}^{-\alpha-\frac{\Pi}{\tau}}w\phi(w)\mathrm{d}w\right]\cdot\Phi(-\alpha)}^{\Omega_{6,\text{cont.}}}. \end{aligned}$$

(A.34)

Moreover, any u between 0 and $u^\star$ can be uniquely realized as tpp^∞, by setting $\alpha = t^\star(u)$ which is the root to (II.4).

From this lemma, we know for $\delta < 1$ and large ϵ, it is possible that the range of tpp^∞ is no longer $(0,1)$. In such a case, we are "above the Donoho–Tanner phase transition (DTPT)"; and symmetrically, when tpp^∞ has the range $(0,1)$, we are "below the DTPT". The purpose of this lemma is mainly for the completeness of the theory. In the following, however, we will always assume the range of tpp^∞ is $(0,1)$ to avoid extra complicity. This assumption will simplify our argument, but the proofs of the theorems can be extended to the case when this assumption is not true.

Now, we prove that the upper curve can be achieved by any (ϵ, M)-homogeneous prior (II.7). This implies that the homogeneous effect sizes are the least desired.

Lemma 1.16: Given (ϵ, δ) and $\sigma = 0$. Any (ϵ, M)-homogeneous prior gives the same unique trade-off curve q^∇ on $(0, u^\star)$. Furthermore, this curve has the expression specified in (II.8).

Proof of Lemma 1.16: We start with the proof to show the curve q^∇ is *unique* in the sense that any two different (ϵ, M)-homogeneous priors give the same trade-off curve. Consider any two (ϵ, M)-homogeneous priors Π_1 and Π_2. Let their nonzero conditional priors be $\Pi_1^\star \equiv M_1$ and $\Pi_2^\star \equiv M_2$. Treating $\alpha > \alpha_0$ as the free parameter, we denote the solution to τ in equation (IV.1) with prior Π_1 by τ_1. We have

$$\delta = (1-\epsilon)\,\mathbb{E}(\eta_\alpha(W)^2) + \epsilon\,\mathbb{E}\left(\eta_\alpha\left(\frac{M_1}{\tau_1} + W\right) - \frac{M_1}{\tau_1}\right)^2.$$

It is clear from a simple calculation that $\tau_2 = \tau_1 \frac{M_1}{M_2}$, α and Π_2 also solve the first equation in (IV.1), that is,

$$\delta = (1-\epsilon)\,\mathbb{E}(\eta_\alpha(W)^2) + \epsilon\,\mathbb{E}\left(\eta_\alpha\left(\frac{M_2}{\tau_2} + W\right) - \frac{M_2}{\tau_2}\right)^2, \tag{A.40}$$

which implies τ_2 is the solution to (IV.1) given α and prior Π_2. Observe the relationships $\tau_2 = \tau_1 \frac{M_1}{M_2}$, $\Pi_1^\star \equiv M_1$ and $\Pi_2^\star \equiv M_2$. We have

$$\mathbb{P}\left(\left|\frac{\Pi_1^\star}{\tau_1} + W\right| > \alpha\right) = \mathbb{P}\left(\left|\frac{\Pi_2^\star}{\tau_2} + W\right| > \alpha\right).$$

Therefore, combining the equality above with (IV.2), we obtain

$$\begin{aligned}\text{tpp}_\alpha^\infty(\Pi_1) &= \mathbb{P}\left(\left|\frac{\Pi_1^\star}{\tau_1} + W\right| > \alpha\right) = \mathbb{P}\left(\left|\frac{\Pi_2^\star}{\tau_2} + W\right| > \alpha\right)\\ &= \text{tpp}_\alpha^\infty(\Pi_2),\end{aligned}$$

and

$$\begin{aligned}\text{fdp}_\alpha^\infty(\Pi_1) &= \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon\,\mathbb{P}\left(\left|\frac{\Pi_1^\star}{\tau_1} + W\right| > \alpha\right)}\\ &= \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon\,\mathbb{P}\left(\left|\frac{\Pi_2^\star}{\tau_2} + W\right| > \alpha\right)}\\ &= \text{fdp}_\alpha^\infty(\Pi_1).\end{aligned}$$

This means that any point on $q^{\Pi_1}(\cdot)$ is also on $q^{\Pi_2}(\cdot)$. Similarly, any point on $q^{\Pi_2}(\cdot)$ is also on $q^{\Pi_1}(\cdot)$. By Lemma 1.11, they are both strictly increasing function, and thus must equal everywhere on the entire domain $(0, u^\star)$.

Now, we proceed to prove that this unique trade-off curve has the expression given by (II.8). Fix some (ϵ, M)-homogeneous prior Π^∇. Let u be some point between 0 and $u^\star = 1$, and set α such that $\text{tpp}_\alpha^\infty(\Pi^\nabla)$ equals to u. We have

$$u = \text{tpp}_\alpha^\infty = \mathbb{P}(|\Pi^{\nabla\star} + \tau W| > \alpha\tau) = \Phi(-\alpha + \widetilde{M}) + \Phi(-\alpha - \widetilde{M}),$$

where $\widetilde{M} = \frac{M}{\tau}$. Let $\varsigma = -\alpha + \widetilde{M}$, then the equation above becomes

$$\Phi(\varsigma) + \Phi(-2\alpha - \varsigma) = u. \tag{A.41}$$

According to (IV.1), we have

$$\delta = (1-\epsilon)\mathbb{E}[\eta_\alpha(W)]^2 + \epsilon\mathbb{E}[\eta_\alpha(\widetilde{M} + W) - \widetilde{M}]^2. \tag{A.42}$$

By a simple algebraic fact

$$\mathbb{E}[\eta_\alpha(W)]^2 = 2[(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha)],$$

and the fact

$$\begin{aligned}&\mathbb{E}[\eta_\alpha(\widetilde{M} + W) - \widetilde{M}]^2\\ =& -(\alpha + \widetilde{M})\phi(\alpha - \widetilde{M}) - (\alpha - \widetilde{M})\phi(\alpha + \widetilde{M})\\ &+ (1+\alpha^2)[\Phi(-\alpha + \widetilde{M}) + \Phi(-\alpha - \widetilde{M})]\\ &+ \widetilde{M}^2[\Phi(\alpha - \widetilde{M}) - \Phi(-\alpha - \widetilde{M})],\end{aligned}$$

we can plug-in the definition of ς into (A.42) and obtain

$$\begin{aligned}\delta =\ & 2(1-\epsilon)[(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha)]\\ &+ \epsilon[-(2\alpha+\varsigma)\phi(\varsigma) + \varsigma\phi(2\alpha+\varsigma) +\\ &(1+\alpha^2)[\Phi(\varsigma) + \Phi(-2\alpha-\varsigma)]\\ &+ (\varsigma+\alpha)^2[\Phi(-\varsigma) + \Phi(-2\alpha-\varsigma)]].\end{aligned} \tag{A.43}$$

So $\varsigma = \varsigma(\alpha; \epsilon, \delta)$ is the solution of the equation above. And finally combining the last equation with (A.41), we get an equation in α

$$\Phi(\varsigma(\alpha)) + \Phi(-2\alpha - \varsigma(\alpha)) = u, \tag{A.44}$$

Denote the solution of α to the equation above by $t^\nabla = t^\nabla(u; \epsilon, \delta)$. We have

$$q^\nabla(u; \epsilon, \delta) = \text{fdp}_{t^\nabla}^\infty(\Pi^\nabla) = \frac{2(1-\epsilon)\Phi(-t^\nabla(u))}{2(1-\epsilon)\Phi(-t^\nabla(u)) + \epsilon u}. \tag{A.45}$$

Therefore, the expression for the upper boundary is just defined by (A.45), where t^∇ is solved from (A.43) and (A.44). □ We comment about the existence of α in the proof above. Note that both equations (A.43) and (A.44) are derived from the AMP equations, which for any $\alpha > \alpha_0$, have unique solution τ. Note that ς is a function of τ, and thus it is also unique. Therefore, the solution to (A.43) also uniquely exists by Lemma 1.15.

3) Miscellaneous Proofs for Section IV-B: In this section, we prove all necessary lemmas needed for Theorem 1, and then prove Theorem 1. We start by giving the proof to Lemma 4.4.

Proof of Lemma 4.4: We treat τ as the free parameter instead of λ. To explicitly express the limiting process of $\boldsymbol{M}$, we consider a sequence of priors $\{\Pi^\Delta(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})\}_t$, where

$M_1^{(t)} \to \infty$ and $M_{i+1}^{(t)}/M_i^{(t)} \to \infty$ as $t \to \infty$. From (IV.2), the asymptotic TPP of $\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})$ at τ can be written as

$$\begin{aligned}\mathrm{tpp}_{\tau}^{\infty}(\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})) = \ & \mathbb{P}\left(|\Pi^{\Delta\star}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma}) + \tau W| > \alpha\tau\right) \\ = & \left[\gamma_1 \mathbb{P}\left(\left|W + \frac{M_1^{(t)}}{\tau}\right| > \alpha\right) + \gamma_2 \mathbb{P}\left(\left|W + \frac{M_2^{(t)}}{\tau}\right| > \alpha\right)\right. \\ & \left. + \cdots + \gamma_m \mathbb{P}\left(\left|W + \frac{M_m^{(t)}}{\tau}\right| > \alpha\right)\right], \end{aligned} \tag{A.46}$$

where α is solved from (IV.1). We denote the last display by $\mathrm{tpp}_{\tau}^{\infty(t)}$ for convenience. Similarly, we denote

$$\begin{aligned}\mathrm{fdp}_{\tau}^{\infty(t)} &\equiv \mathrm{fdp}_{\tau}^{\infty}(\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})) \\ &= \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon \mathrm{tpp}^{\infty(t)}(\tau)}. \end{aligned} \tag{A.47}$$

In the following proof, we will choose an m-tuple of $(\tau_i^{(t)})_{i=1}^m$ for each fixed t, such that as $t \to \infty$, $(\mathrm{tpp}_{\tau_i^{(t)}}^{\infty(t)}, \mathrm{fdp}_{\tau_i^{(t)}}^{\infty(t)}) \to (u_i, q^{\Delta}(u_i))$ at m different u_i's. This implies exactly that the limit of trade-off curves $q^{\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})}$ agrees with q^{Δ} at (at least) different m points in $(0, 1]$. A natural way to pick such an m-tuple $(\tau_i^{(t)})_{i=1}^m$ is

$$\tau_i^{(t)} = \sqrt{M_i^{(t)} M_{i+1}^{(t)}}, 1 \le i \le m-1,$$

and $\tau_m^{(t)} = m \times M_m^{(t)}$ when $i = m$.[6] Under the regime of $M_1^{(t)} \to \infty$ and $M_{i+1}^{(t)}/M_i^{(t)} \to \infty$ for all i, we know $\tau_i^{(t)}$ satisfies

$$|M_i^{(t)}| = o(\tau_i^{(t)}), \text{ and } \tau_i^{(t)} = o(|M_{i+1}^{(t)}|), \quad 1 \le i \le m-1 \tag{A.48}$$

and

$$|M_m^{(t)}| = o(\tau_m^{(t)}), \tag{A.49}$$

as $t \to \infty$. Moreover, for any $\alpha > \alpha_0$ and any $1 \le i \le m$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|W + \frac{M_j^{(t)}}{\tau_i^{(t)}}\right| > \alpha\right) \\ &= \begin{cases} \mathbb{P}(|W| > \alpha) + o_t(1), & \text{for } j \le i, \\ 1 + o_t(1), & \text{for } j \ge i+1, \end{cases} \end{aligned} \tag{A.50}$$

and

$$\eta_\alpha\left(W + \frac{M_j^{(t)}}{\tau_i^{(t)}}\right) - \frac{M_j^{(t)}}{\tau_i^{(t)}} = \begin{cases} \eta_\alpha(W) + o_t(1), & \text{for } j \le i, \\ W - \alpha + o_{\mathbb{P},t}(1), & \text{for } j \ge i+1. \end{cases} \tag{A.51}$$

Let $\gamma^{(j)} = \sum_{i=j+1}^m \gamma_i$ and use $\alpha_i^{(t)}$ to denote the solution of α to (IV.1) given $\tau_i^{(t)}$. We have

$$\begin{aligned}&\left(1 - \frac{\sigma^2}{\tau_i^{(t)2}}\right)\delta \\ &= \epsilon \cdot \underbrace{\mathbb{E}\left(\eta_\alpha\left(\frac{\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})}{\tau_i^{(t)}} + W\right) - \frac{\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})}{\tau_i^{(t)}}\right)}_{(*)} \\ &\quad + (1-\epsilon)\,\mathbb{E}(\eta_\alpha(W))^2. \end{aligned} \tag{A.52}$$

By (A.51), the $(*)$ part of the last display is

$$\begin{aligned}(*) &= \sum_{j=1}^{i} \gamma_j\, \mathbb{E}\left(\eta_\alpha\left(\frac{M_j^{(t)}}{\tau_i^{(t)}} + W\right) - \frac{M_i^{(t)}}{\tau^{(t)}}\right)^2 \\ &\quad + \sum_{j=i+1}^{m} \gamma_j\, \mathbb{E}\left(\eta_\alpha\left(\frac{M_j^{(t)}}{\tau_i^{(t)}} + W\right) - \frac{M_j^{(t)}}{\tau_i^{(t)}}\right)^2 \\ &= (1-\gamma^{(i)})\,\mathbb{E}(\eta_\alpha(W)^2) + \gamma^{(i)}\,\mathbb{E}((W-\alpha)^2) + o_t(1). \end{aligned}$$

Observe the fact that σ is fixed and thus $\frac{\sigma^2}{\tau_i^{(t)2}} \to 0$. With some simple calculation, (A.52) can be written as

$$\epsilon\gamma^{(i)}(1+\alpha^2) + 2(1-\epsilon\gamma^{(i)})[(1+\alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha)] = \delta + o_t(1).$$

Therefore the solution $\alpha_i^{(t)}$ of the equation above has a limit[7]

$$\alpha_i^{(t)} \to \alpha^{(i)}, \quad \text{as } t \to \infty, \tag{A.53}$$

which solves the equation

$$\epsilon\gamma^{(i)}(1+\alpha^2) + 2(1-\epsilon\gamma^{(i)})[(1+\alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha)] = \delta. \tag{A.54}$$

Note $\alpha^{(i)}$ is independent of the choice of $\{M^{(t)}\}_{t=1}^{\infty}$ and $(\tau_i^{(t)})_{i=1}^m$. Direct calculation can verify that each solution $\alpha^{(i)}$ also satisfies the equation (II.4) with setting

$$u = u_i = 2\Phi(-\alpha^{(i)})(1-\gamma^{(i)}) + \gamma^{(i)}. \tag{A.55}$$

This implies $\alpha^{(i)}$ is also the unique solution of (II.4), so

$$\alpha^{(i)} = t^{\Delta}(u_i). \tag{A.56}$$

Combining (A.46), (A.47), (A.50) and (A.53), the limits of $\mathrm{tpp}_{\tau_i^{(t)}}^{\infty(t)}$ and $\mathrm{fdp}_{\tau_i^{(t)}}^{\infty(t)}$ are

$$\begin{cases} \mathrm{tpp}_{\tau_i^{(t)}}^{\infty(t)} & \to \ 2\Phi(-\alpha^{(j)})(1-\gamma^{(i)}) + \gamma^{(i)}, \\ \mathrm{fdp}_{\tau_i^{(t)}}^{\infty(t)} & \to \frac{2(1-\epsilon)\Phi(-\alpha^{(i)})}{2(1-\epsilon)\Phi(-\alpha^{(i)}) + \epsilon(2\Phi(-\alpha^{(i)})(1-\gamma^{(i)}) + \gamma^{(i)})}. \end{cases} \tag{A.57}$$

By $\alpha^{(i)} = t^{\Delta}(u_i)$ and (A.55), we obtain from (II.5) that

$$\begin{aligned}q^{\Delta}(u_i; \delta, \epsilon) &= \frac{2(1-\epsilon)\Phi(-t^{\Delta}(u_i))}{2(1-\epsilon)\Phi(-t^{\Delta}(u_i)) + \epsilon u_i} \\ &= \frac{2(1-\epsilon)\Phi(-\alpha^{(i)})}{2(1-\epsilon)\Phi(-\alpha^{(i)}) + \epsilon(2\Phi(-\alpha^{(i)})(1-\gamma^{(i)}) + \gamma^{(i)})}. \end{aligned} \tag{A.58}$$

Combining (IV.3), (IV.2), (A.55), (A.57) and (A.58), we finally obtain as $t \to \infty$,

$$\begin{cases} \mathrm{tpp}_t^{\infty}(\tau_i^{(t)}) \to u_i, \\ \mathrm{fdp}_t^{\infty}(\tau_i^{(t)}) \to q^{\Delta}(u_i; \delta, \epsilon). \end{cases}$$

Therefore, the limiting function of the trade-off curves of priors $\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})$ agrees with the lower boundary $q^{\Delta}(\cdot; \delta, \epsilon)$ at $(u_i, q^{\Delta}(u_i))$, for $i = 1, 2, \cdots, m$. Since the m different

[6] In fact, one can pick any $\tau_i^{(t)}$ such that (A.48) and (A.49) hold.

[7] By the existence asserted by AMP theory, the equation (A.54) has a unique solution, denote it by $\alpha^{(i)}$, we know the solution $\alpha_i^{(t)}$ of (A.53) must converge into it, since the left-hand side of (A.53) is continuous in α.

points are nonzero, there must be at least $m-1$ points in $(0,1)$. □

An important set of equations is (A.55). Recall that (A.56) asserts $\alpha^{(i)} = t^{\Delta}(u_i)$ for all i, and therefore the equations (A.55) are, for all i,

$$u_i = 2\Phi(-t^{\Delta}(u_i))(1-\gamma^{(i)}) + \gamma^{(i)}. \tag{A.59}$$

The last display allows us to quantify the exact points u_i the limit of $\Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma})$ agrees with q^{∇}. This fact allows us to set $\boldsymbol{\gamma}$ cleverly so that the distance between any two consecutive u_i's are small enough. This is formalized in the following lemma.

Lemma 1.17: For any $\xi > 0$, there is some $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_m\}$, with $\sum_i \gamma_i = 1$, such that the m points specified by (A.59), together with $u_0 = 0$ and $u_{m+1} = 1$ [8] satisfy the following

$$\max_{1\le j\le m+1} |u_j - u_{j-1}| < \frac{\xi}{2}. \tag{A.60}$$

Proof of Lemma 1.17: We first prove that the difference $u_{m+1} - u_m = 0$. Since $\gamma^{(m)} = 1$, and by (A.55), we obtain that $u_m = 1$, and thus $u_{m+1} - u_m = 1 - 1 = 0$.[9] With this in mind, we only need to prove the following two quantities can be arbitrarily small to ensure (A.60),

$$u_1 - 0 = \gamma_m + 2(1-\gamma_m)\Phi(-\alpha^{(m)}), \tag{A.61}$$

and for all $m \ge j \ge 2$,

$$\begin{aligned} u_j - u_{j-1} = \gamma_{m-j+1} &+ 2(1-\gamma^{(j)})\Phi(-\alpha^{(j)}) \\ &- 2(1-\gamma^{(j-1)})\Phi(-\alpha^{(j-1)}), \end{aligned} \tag{A.62}$$

where we remind the reader that by definition, $\gamma^{(m)} = \gamma_m$ and $\gamma^{(j)} - \gamma^{(j-1)} = \gamma_{m-j+1}$.

For the expression in (A.62), we observe that

$$\begin{aligned} u_j - u_{j-1} \le& \gamma_{m-j+1}|1 + 2(\Phi(-\alpha^{(j)}) - \Phi(-\alpha^{(j-1)}))| \\ &+ 2(1-\gamma^{(j)})|\Phi(-\alpha^{(j)}) - \Phi(-\alpha^{(j-1)})| \\ \le& 5\gamma_{m-j+1} + 2|\Phi(-\alpha^{(j)}) - \Phi(-\alpha^{(j)} - \gamma_{m-j+1}))|. \end{aligned} \tag{A.63}$$

We observe that in equation (A.55) or (A.59), the dependence of $\alpha^{(j)}$ on $\gamma^{(j)}$ is only through linear functional of $\gamma^{(j)}$. Therefore, it is not hard to realize that $\alpha^{(j)}$ is continuous in $\gamma^{(j)}$. When all $\{\gamma_s\}_{s>m-j+1}$ are fixed, the $\alpha^{(j)}$ is a continuous function in γ_{m-j+1}, and so is the expression in (A.63). So we can pick γ_{m-j+1} sufficiently small to ensure the (A.63) is less than $\frac{\xi}{2}$.

For the expression in (A.61), we pick some M sufficiently large such that $\Phi(-M) < \frac{\xi}{8}$. By Lemma 1.18, we can pick $\gamma_m < \frac{\xi}{4}$ such that the solution to (II.4) with u being (A.61) satisfies that $\alpha^{(m)} > M$, or $\Phi(-\alpha^{(m)}) < \Phi(-M) < \frac{\xi}{8}$. Therefore

$$\gamma_m + 2(1-\gamma^{(m)})\Phi(-\alpha(\gamma^{(m)})) < \frac{\xi}{4} + 2\cdot 1\cdot\frac{\xi}{8} = \frac{\xi}{2}.$$

□

Lemma 1.18: For any fixed $\delta,\ \epsilon, \xi > 0$ and $M > 0$. There exists $\gamma < \frac{\xi}{4}$, such that the solution α to (II.4) with $u = \gamma + 2(1-\gamma)\Phi(-\alpha)$ satisfies $\alpha > M$.

Proof of Lemma 1.18: We will use the following fact: there exists $\gamma < \frac{\xi}{4}$ and large $M'' > M' > M$, such that:

$$\begin{aligned} &(1-\epsilon\gamma)A(M') + \epsilon\gamma(1+M'^2) < \delta, \\ &(1-\epsilon\gamma)A(M'') + \epsilon\gamma(1+M''^2) > \delta. \end{aligned} \tag{A.64}$$

where $A(M) = 2[(1+M^2)\Phi(-M) - M\phi(M)]$.

Taken this as given for the moment, we set $u = \gamma + 2(1-\gamma)\Phi(-\alpha)$ with γ such that (A.64) holds. Then (II.4) becomes[10]

$$\begin{aligned} &\frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]} \\ &\quad = \frac{1-\gamma-2(1-\gamma)\Phi(-t)}{1-2\Phi(-t)}. \end{aligned}$$

Observe the right hand side of the last display is just $1-\gamma$, so it is equivalent to

$$\begin{aligned} &2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta \\ &\quad = (1-\gamma)\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right], \end{aligned}$$

or,

$$2(1-\epsilon\gamma)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon\gamma(1+t^2) - \delta = 0. \tag{A.65}$$

By relationship (A.64) and the fact that there exists unique solution $\alpha = t^{\nabla}(u)$ to (II.4) and thus to (A.65), we know the solution $\alpha \in (M', M'')$, and thus especially $\alpha > M$.

Now, to prove (A.64), we first note that it is direct to verify $A(t) = E[\eta_t(W)^2]$, and thus it is decreasing in t. And as $t \to \infty$, $A(t) \to 0$. Therefore for any $\delta > 0$, we can pick M' large enough such that $A(M') < \frac{\delta}{2}$, and now pick $\gamma < \frac{\xi}{4}$ small enough such that $\epsilon\gamma(1+M'^2) < \frac{\delta}{2}$. Therefore, the left-hand side of the first equation of (A.64) is bounded by δ. For the second equation in (A.64), pick M'' large enough so that the term $\epsilon\gamma(1+M''^2) > \delta$, and since $(1-\epsilon\gamma)A(M'') > 0$, the second line also holds. □

The agreeing points asserted by Lemma 4.4 are close to each other in their x-coordinate distances. Therefore, by the uniform continuity of the lower curve q^{∇} and Cantor's diagonalization argument, we can extend the result from Lemma 4.4 to uniform convergence.

Lemma 1.19: There exist a sequence of prior of $\Pi^{(t)} = \Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma}^{(t)})$, such that their trade-off curve $q^{\Pi^{(t)}}$ converge uniformly to q^{Δ} on any compact interval in $(0,1)$.

Proof of Lemma 1.19: Fix any compact interval $\mathcal{I} = [a,b]$ in $(0,1)$. As in Lemma 4.4, we first consider prior $\Pi^{(t)} = \Pi^{\Delta}(\boldsymbol{M}^{(t)}, \boldsymbol{\gamma}^{(t)})$ with $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}$ being some fixed m-tuple. By Lemma 1.11, we know that both $q^{\Pi^{(t)}}(u)$ and $q^{\Delta}(u)$ are continuous and strictly increasing. Consider any

[8]Technically speaking, it should be $u_{m+1} = u^{\star}$, yet as discussed earlier, we will focus on the case when we are below the Donoho–Tanner phase transition, so always $u^{\star} = 1$.

[9]One might want to verify the existence of $\alpha^{(m)}$. Since we always assume that we are below the DTPT, then for any $u < u^{\star} = 1$, the $\alpha = t^{\Delta}(u_i)$ exists as the solution to (II.4) by Lemma 1.15. By setting $\gamma^{(m)} = 1$, one can directly verify this corresponds to set $u = u_m = 1^-$, and by the continuity of t in equation (II.4), we know $\alpha^{(m)} = t^{\Delta}(u_m)$ exists and less than infinite. And since all other $\alpha^{(i)}$'s are less than $\alpha^{(m)}$, they also exist.

[10]Since the solution to t is α here, we can plug in $u = \gamma + 2(1-\gamma)\Phi(-t)$.

two adjacent agreeing points u_j, u_{j+1} specified in (A.55), such that $q^{\Delta}(u_j) = \lim_{t\to\infty} q^{\Pi^{(t)}}(u_j)$ and $q^{\Delta}(u_{j+1}) = \lim_{t\to\infty} q^{\Pi^{(t)}}(u_{j+1})$. Since in the interval (u_j, u_{j+1}) the difference is controlled by

$$q^{\Pi^{(t)}}(u) - q^{\Delta}(u) \le q^{\Delta}(u_{j+1}) - q^{\Delta}(u_j), \text{ for any } u \in (u_j, u_{j+1})$$

by the monotonicity of $q^{\Pi^{(t)}}$. This difference will be small as long as the gap $q^{\Delta}(u_{j+1}) - q^{\Delta}(u_j)$ is small, so we proceed to prove we can select $\Pi^{(t)}$ to ensure the gaps $q^{\Delta}(u_{j+1}) - q^{\Delta}(u_j)$ are small for all i.

Fix any $\theta > 0$. Since q^{Δ} is uniformly continuous on the compact set $\mathcal{I}$, there exists $\xi > 0$ such that for any $u, v \in \mathcal{I}$,

$$|q^{\Delta}(u) - q^{\Delta}(v)| < \frac{\theta}{2}, \quad \text{as long as } |u - v| < \xi \tag{A.66}$$

By the proof of Lemma 4.4, we can construct $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}_\theta$ to be specified later, and $\boldsymbol{M}^{(t)}_{\boldsymbol{\gamma}_\theta}$, such that the limit of $q^{\Pi^{(t)}}$ agrees with q^{Δ} at m points $u_1, \cdots, u_m$. This implies there exists some T_θ such that for all $t \ge T_\theta$,

$$\max_j \left| q^{\Pi^{(t)}}(u_j) - q^{\Delta}(u_j) \right| < \frac{\theta}{2}. \tag{A.67}$$

To specify the choice of $\boldsymbol{\gamma}_\theta$, note that by Lemma 1.17, we can choose $\boldsymbol{\gamma}_\theta$ such that $u_1, .., u_m$ satisfies $u_1 - a < u_1 - 0 < \frac{\xi}{2}$, $u_m - b < \frac{\xi}{2}$ and

$$\max_{2\le j\le m} |u_j - u_{j-1}| < \frac{\xi}{2}.$$

With this choice of $\boldsymbol{\gamma}_\theta$ together with (A.66) and (A.67), we obtain

$$\sup_{t\ge T_\theta, u\in\mathcal{I}} \left| q^{\Pi^{(t)}}(u) - q^{\Delta}(u) \right| < \theta.$$

Specifically, we have the equation above holds for $t = T_\theta$,

$$\sup_{u\in\mathcal{I}} \left| q^{\Pi^{(T_\theta)}}(u) - q^{\Delta}(u) \right| < \theta.$$

Since $\Pi^{(T_\theta)} = \Pi^{\Delta}\left(\boldsymbol{M}^{(T_\theta)}_{\boldsymbol{\gamma}_\theta}, \boldsymbol{\gamma}_\theta\right)$, the inequality above is simply

$$\sup_{u\in\mathcal{I}} \left| q^{\Pi^{\Delta}(\boldsymbol{M}^{(T_\theta)}_{\boldsymbol{\gamma}_\theta}, \boldsymbol{\gamma}_\theta)}(u) - q^{\Delta}(u) \right| < \theta. \tag{A.68}$$

Now we can apply Cantor's diagonalization trick since the last display is true for any $\theta > 0$. We set $\theta_\zeta = \frac{1}{\zeta} \to 0, \zeta \ge 1$, and choose the priors

$$\Pi^{(\zeta)} = \left\{ \Pi^{\Delta}\left(\boldsymbol{M}^{(T_{\theta_\zeta})}_{\boldsymbol{\gamma}_{\theta_\zeta}}, \boldsymbol{\gamma}_{\boldsymbol{\theta}_{\boldsymbol{\zeta}}}\right) \right\}_\zeta.$$

Then we know from (A.68) that $q^{\Pi^{(\zeta)}}$ converges to q^{Δ} uniformly on $\mathcal{I}$ as $\zeta \to \infty$. □ Now, we can proceed to prove Theorem 1, whose proof is very similar to that of Theorem 2 in Section IV-A.

Proof of Theorem 1 (a): Consider any non-constant prior Π, we first prove that there exists some Π^{Δ} and $\nu > 0$ such that for all $c < \lambda, \lambda' < C$,

$$\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta}) < \text{tpp}^{\infty}_{\lambda'}(\Pi) + \nu \text{ and } \text{fdp}^{\infty}_{\lambda}(\Pi^{\Delta}) > \text{fdp}^{\infty}_{\lambda'}(\Pi) - \nu \tag{A.69}$$

cannot hold simultaneously. To see this fact, first find $0 < u_1 < u_2 < 1$ such that the asymptotic powers $\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta}), \text{tpp}^{\infty}_{\lambda'}(\Pi)$ are always between u_1 and u_2 for $c < \lambda, \lambda' < C$. Next, we know from Lemma 1.20 that q^{Δ} is the strictly below any trade-off curve. So, for any prior Π, we have $q^{\Delta}(u) < q^{\Pi}(u)$ for any $u \in \mathcal{I} = [u_1, u_2]$. Note that both q^{Δ} and q^{Π} are uniformly continuous on $\mathcal{I}$ and thus one can set $\nu' > 0$ to be

$$\nu' := \inf_{u_1\le u\le u_2} \left(q^{\Pi}(u) - q^{\Delta}(u) \right) > 0. \tag{A.70}$$

Since q^{Δ} is a continuous function on the closed interval $[0, 1]$, we can make use of its uniform continuity, which ensures

$$\left| q^{\Delta}(u) - q^{\Delta}(u') \right| < \frac{\nu'}{4} \tag{A.71}$$

as long as $|u - u'| \le \nu''$ for some $\nu'' > 0$. By the assertion of Lemma 1.19, we can choose a prior Π^{Δ} such that it is $\frac{\nu'}{4}$-close to q^{Δ} on $\mathcal{I}$,

$$\sup_{u_1\le u\le u_2} (q^{\Pi^{\Delta}}(u) - q^{\Delta}(u)) < \frac{\nu'}{4}. \tag{A.72}$$

Now we can prove (IV.4) cannot hold simultaneously with our choice of Π^{Δ} and $\nu = \min\{\nu'/2, \nu''\}$. To see this, suppose we already have $\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta}) < \text{tpp}^{\infty}_{\lambda'}(\Pi) + \nu$. Now observe that

$$\begin{aligned}
\text{fdp}^{\infty}_{\lambda}(\Pi^{\Delta}) &= q^{\Pi^{\Delta}}(\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta})) \\
&\le q^{\Delta}\left(\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta})\right) + \frac{\nu'}{4} \\
&< q^{\Delta}\left(\text{tpp}^{\infty}_{\lambda'}(\Pi^{\Delta}) + \nu\right) + \frac{\nu'}{4} \\
&\le q^{\Delta}\left(\text{tpp}^{\infty}_{\lambda'}(\Pi)\right) + \frac{\nu'}{2} \\
&\le q^{\Pi}\left(\text{tpp}^{\infty}_{\lambda'}(\Pi)\right) - \nu' + \frac{\nu'}{2} \\
&= q^{\Pi}\left(\text{tpp}^{\infty}_{\lambda'}(\Pi)\right) - \frac{\nu'}{2} \\
&\le q^{\Pi}\left(\text{tpp}^{\infty}_{\lambda'}(\Pi)\right) - \nu \\
&= \text{fdp}^{\infty}_{\lambda'}(\Pi) - \nu,
\end{aligned}$$

where the first inequality follows from (A.72); the second inequality follows from the fact that $\text{fdp}^{\infty}(\text{tpp}^{\infty})$ is strictly increasing, and $\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta}) < \text{tpp}^{\infty}_{\lambda'}(\Pi) + \nu$; the third inequality is by (A.71); the fourth inequality is by (A.70); the last inequality is by the definition of ν. As such, the first inequality in (A.69) leads to the violation of the second inequality. Having shown (A.69), it is easy to prove Theorem 1. Lemma 4.1 ensures that the following four terms

$$\begin{aligned}
&\left|\text{TPP}_{\lambda}(\Pi^{\Delta}) - \text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta})\right|, \left|\text{FDP}_{\lambda}(\Pi^{\Delta}) - \text{fdp}^{\infty}_{\lambda}(\Pi^{\Delta})\right|, \\
&\left|\text{TPP}_{\lambda'}(\Pi) - \text{tpp}^{\infty}_{\lambda'}(\Pi)\right|, \left|\text{FDP}_{\lambda'}(\Pi) - \text{fdp}^{\infty}_{\lambda'}(\Pi)\right|
\end{aligned} \tag{A.73}$$

are all smaller than $\nu/2$ for all $c < \lambda, \lambda' < C$, with probability tending to one as $n, p \to \infty$. On this event, it is easy to check that $\text{TPP}_{\lambda}(\Pi^{\Delta}) \le \text{TPP}_{\lambda'}(\Pi)$ implies $\text{tpp}^{\infty}_{\lambda}(\Pi^{\Delta}) < \text{tpp}^{\infty}_{\lambda'}(\Pi) + \nu$, and $\text{FDP}_{\lambda}(\Pi^{\Delta}) \ge \text{FDP}_{\lambda'}(\Pi)$ implies $\text{fdp}^{\infty}_{\lambda}(\Pi^{\Delta}) > \text{fdp}^{\infty}_{\lambda'}(\Pi) - \nu$. As such, in the event (A.73), the impossibility of (A.69) uniformly for all $c < \lambda$, $\lambda' < C$ implies the impossibility of

$$\text{TPP}_{\lambda}(\Pi^{\Delta}) \le \text{TPP}_{\lambda'}(\Pi) \text{ and } \text{FDP}_{\lambda}(\Pi^{\Delta}) \ge \text{FDP}_{\lambda'}(\Pi)$$

for all $c < \lambda, \lambda' < C$. □

It is the same as the comment after the proof of Theorem 2, we can prove part (b) of Theorem 1 similarly, and we omit for simplicity.

In closing, we present the following lemma to be self-contained. It shows that the lower boundary is strictly below any trade-off curve, on which the proof of Theorem 1 relies.

Lemma 1.20 (Lemma C.3 in [5]): Consider any ϵ-sparse prior Π. The lower boundary q^{Δ} is strictly below the trade-off curve $q^{\Pi}(\cdot)$, that is, $q^{\Delta}(u) < q^{\Pi}(u)$ for any u.

Proof of Lemma 1.20: This is just a re-statement of Lemma C.3 in [5]. They have proved that for any $\text{tpp}^{\infty} = u$, $\text{fdp}^{\infty} > q^{\Delta}(u)$, which implies $q^{\Pi}(u) > q^{\Delta}(u)$. □

ACKNOWLEDGMENT

The authors would like to thank Małgorzata Bogdan, Emmanuel Candès, Edward George, Pragya Sur, and Nancy Zhang for stimulating discussions. They are grateful to the referees and an associate editor for their constructive comments that helped improve the presentation of this work.

REFERENCES

[1] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.
[2] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009.
[3] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Feb. 1994.
[4] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[5] W. Su, M. Bogdan, and E. J. Candès, "False discoveries occur early on the Lasso path," *Ann. Statist.*, vol. 45, no. 5, pp. 2133–2150, Oct. 2017.
[6] W. J. Su, "When is the first spurious variable selected by sequential regression procedures?" *Biometrika*, vol. 105, no. 3, pp. 517–527, Sep. 2018.
[7] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated Gaussian designs," *J. Mach. Learn. Res.*, vol. 11, pp. 2241–2259, Aug. 2010.
[8] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, 2009.
[9] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.
[10] M. Bayati and A. Montanari, "The LASSO risk for Gaussian matrices," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 1997–2017, Apr. 2012.
[11] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4290–4308, Jul. 2013.
[12] M. Bayati, M. A. Erdogdu, and A. Montanari, "Estimating LASSO risk and noise level," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 944–952.
[13] Z. Bu, J. Klusowski, C. Rush, and W. Su, "Algorithmic analysis and statistical estimation of SLOPE via approximate message passing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9361–9371.
[14] A. Weinstein, R. Barber, and E. J. Candès, "A power and prediction analysis for knockoffs with Lasso statistics," 2017, *arXiv:1712.06465*.
[15] A. Mousavi, A. Maleki, and R. G. Baraniuk, "Consistent parameter estimation for LASSO and approximate message passing," *Ann. Statist.*, vol. 46, no. 1, pp. 119–148, 2018.
[16] H. Weng, A. Maleki, and L. Zheng, "Overcoming the limitations of phase transition by higher order analysis of regularization techniques," *Ann. Statist.*, vol. 46, no. 6A, pp. 3099–3129, Dec. 2018.
[17] P. Sur, Y. Chen, and E. J. Candès, "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square," *Probab. Theory Rel. Fields*, vol. 175, nos. 1–2, pp. 487–558, Oct. 2019.
[18] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philos. Trans. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 367, no. 1906, pp. 4273–4293, 2009.
[19] S. Wang, H. Weng, and A. Maleki, "Which bridge estimator is the best for variable selection?" *Ann. Statist.*, vol. 48, no. 5, pp. 2791–2823, Oct. 2020.
[20] H. Wang, Y. Yang, Z. Bu, and W. Su, "The complete Lasso tradeoff diagram," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20051–20060.
[21] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 6920–6941, Oct. 2011.
[22] A. Maleki, "Approximate message passing algorithms for compressed sensing," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2010.
[23] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[24] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
[25] W. Su, M. Bogdan, and E. J. Candès, "Supplement to 'False discoveries occur early on the Lasso path,'" *Ann. Statist.*, 2017. Accessed: Feb. 12, 2021. [Online]. Available: http://www-stat.wharton.upenn.edu/~suw/paper/LassoFDR_supp.pdf
[26] M. Bogdan, E. van den Berg, W. Su, and E. J. Candès, "Supplementary material to 'Statistical estimation and testing via the sorted ℓ_1 norm,'" 2013. [Online]. Available: https://statweb.stanford.edu/~candes/publications/downloads/SortedL1_SM.pdf
[27] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851–862, Oct. 2007.
[28] M. F. F. Bogdan, F. Frommlet, P. Szulc, and H. Tang, "Model selection approach for genome wide association studies in admixed populations," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013.
[29] M. Bogdan, E. van den Berg, W. Su, and E. J. Candès, "Statistical estimation and testing via the sorted ℓ_1 norm," 2013, *arXiv:1310.1969*.
[30] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE—Adaptive variable selection via convex optimization," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1103–1140, 2015.
[31] W. Su and E. J. Candès, "SLOPE is adaptive to unknown sparsity and asymptotically minimax," *Ann. Statist.*, vol. 44, no. 3, pp. 1038–1068, Jun. 2016.
[32] Z. Bu, J. Klusowski, C. Rush, and W. J. Su, "Characterizing the SLOPE trade-off: A variational perspective and the Donoho–Tanner limit," 2021, *arXiv:2105.13302*.
[33] V. Ročková and E. I. George, "The spike-and-slab LASSO," *J. Amer. Stat. Assoc.*, vol. 113, no. 521, pp. 431–444, Jan. 2018.
[34] A. Gelman and F. Tuerlinckx, "Type S error rates for classical and Bayesian single and multiple comparison procedures," *Comput. Statist.*, vol. 15, no. 3, pp. 373–390, Sep. 2000.
[35] R. F. Barber and E. J. Candès, "A knockoff filter for high-dimensional selective inference," *Ann. Statist.*, vol. 47, no. 5, pp. 2504–2537, Oct. 2019.
[36] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Construct. Approx.*, vol. 28, no. 3, pp. 253–263, Dec. 2008.
[37] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 1962.
[38] V. Siskind, "Second moments of inverse Wishart-matrix elements," *Biometrika*, vol. 59, no. 3, pp. 690–691, 1972.

Hua Wang received the B.S. degree in mathematics and applied mathematics from Fudan University in 2018. He is currently pursuing the Ph.D. degree with the Wharton Statistics and Data Science Department, University of Pennsylvania. His research interests include span high-dimensional statistics, privacy-presevering machine learning, and deep learning theory.

Yachong (Elsa) Yang received the B.S. degree in statistics from the School of Gifted Young, University of Science and Technology of China, in 2019. She is currently pursuing the Ph.D. degree in statistics and data science with the Wharton School, University of Pennsylvania.

Weijie J. Su (Member, IEEE) received the bachelor's degree from Peking University in 2011 and the Ph.D. degree from Stanford University in 2016. He is currently an Assistant Professor with the Wharton Statistics and Data Science Department and the Department of Computer and Information Science, University of Pennsylvania. He is the Co-Director of Penn Research in Machine Learning (https://priml.upenn.edu/). His research interests span privacy-preserving data analysis, optimization, high-dimensional statistics, and deep learning theory. He was a recipient of the Stanford Theodore W. Anderson Dissertation Award in 2016, NSF CAREER Award in 2019, the Alfred P. Sloan Research Fellowship in 2020, and the Society for Industrial and Applied Mathematics (SIAM) Early Career Prize in Data Science in 2022.