On Sure Early Selection of the Best Subset

Ziwei Zhu\textsuperscript{C}, Member, IEEE, and Shihao Wu

Abstract— The early solution path, which tracks the first few variables that enter the model of a selection procedure, is of profound importance to scientific discoveries. In practice, it is often statistically hopeless to identify all the important features with no false discovery, let alone the intimidating expense of experiments to test their significance. Such realistic limitation calls for statistical guarantee for the early discoveries of a model selector. In this paper, we focus on the early solution path of best subset selection (BSS), where the sparsity constraint is set to be lower than the true sparsity. Under a sparse high-dimensional linear model, we establish the sufficient and (near) necessary condition for BSS to achieve sure early selection, or equivalently, zero false discovery throughout its early path. Essentially, this condition boils down to a lower bound of the minimum projected signal margin that characterizes the gap of the captured signal strength between sure selection models and those with spurious discoveries. Defined through projection operators, this margin is insensitive to the restricted eigenvalues of the design, suggesting the robustness of BSS against collinearity. Moreover, our model selection guarantee tolerates reasonable optimization error and thus applies to near best subsets. Finally, to overcome the computational hurdle of BSS under high dimension, we propose the “screen then select” (STS) strategy to reduce dimension for BSS. Our numerical experiments show that the resulting early path exhibits much lower false discovery rate (FDR) than LASSO, MCP and SCAD, especially in the presence of highly correlated design. We also investigate the early paths of the iterative hard thresholding algorithms, which are greedy computational surrogates for BSS, and which yield comparable FDR as our STS procedure.

Index Terms— Sure early selection, best subset selection, false discovery rate, solution path, sure screening.

I. INTRODUCTION

HIGH dimensional sparse linear models have been receiving intense theoretical investigation and widely applied in the big data era. Suppose we have \( n \) independent and identically distributed (i.i.d.) observations \( \{(x_i, y_i)\}_{i=1}^{n} \) of \((x, Y)\) that follows the linear model:

\[
Y = X^\top \beta^* + \epsilon.
\]

Here \( x \) is a design vector valued in \( \mathbb{R}^p \), \( \beta^* \in \mathbb{R}^p \) is an unknown sparse coefficient vector, and \( \epsilon \) is random noise independent of \( x \). Write \( X = (x_1, x_2, \ldots, x_n)^\top \), \( y = (y_1, \ldots, y_n)^\top \) and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top \). Then in matrix form, we have that

\[
y = X\beta^* + \epsilon.
\]

A central problem for high dimensional sparse linear models is the variable selection problem, i.e., to estimate the true support \( S^* \) of \( \beta^* \). Write \( s^* = |S^*| \). Given a target sparsity \( s \), which is not necessarily equal to \( s^* \), the best subset selection (BSS) solves

\[
\hat{\beta}^\text{best}_s := \arg\min_{\beta \in \mathbb{R}^p, \|\beta\|_0 \leq s} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2.
\]

In words, BSS seeks for the size-\( s \) subset of the available variables that achieves the minimum \( L_2 \) error of fitting \( y \). Another related type of subset regression approaches, which emerged in the 1970s, is the \( \ell_0 \)-regularized approach, exemplified by Mallow’s \( C_p \) [27], Akaike Information Criterion (AIC) [1], [2] and Bayesian Information Criterion (BIC) [33]. Instead of directly constraining \( \|
\beta\|_0 \), these approaches penalize the loss function by a regularization term that is proportional to \( \|
\beta\|_0 \), which can be viewed as an indicator of model complexity. Nevertheless, both the \( \ell_0 \)-constrained and \( \ell_0 \)-regularized methods are notorious for their NP-hardness [30] and are thus computationally infeasible under high dimensions.

For a long period of time, the computational barrier of BSS has been shifting attention away from its exact discrete formulation to its surrogate forms that are amenable to polynomial algorithms. The past three decades or so have witnessed a flurry of profound works on this respect, giving rise to a myriad of variable selection methods with both statistical accuracy and computational efficiency, particularly in high-dimensional regimes. A partial list of them include LASSO [10], [38], SCAD [13], [14], [17], [25], [26], elastic net [49], adaptive LASSO [48], MCP [45] and so forth. The shared spirit of these approaches is to substitute the \( \ell_0 \)-regularization with a surrogate penalty of model complexity as follows:

\[
\hat{\beta}^\text{pen} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \rho_\lambda(\beta),
\]

where \( \rho_\lambda(\beta) \), parameterized by \( \lambda \), is a regularizer that encourages parsimony. All these methods are backed up by solid guarantee of model consistency. For example, [46] established the well-known irrepresentable conditions for model consistency of LASSO under fixed designs. Reference [45] showed that MCP achieves model consistency when the design satisfies a sparse Riesz condition and the minimum signal strength
is not too weak. Reference [16] proposed an iterative local adaptive majorize-minimization (I-LAMM) algorithm for general empirical risk minimization with folded concave penalty (e.g., SCAD) and showed that only a local Riesz condition suffices to ensure model consistency. A recent work [21] considered a hybrid of $L_0$ and $L_2$ regularization for problem (4), which we refer to as LOL2, to pursue sparsity, robustness and computational efficiency. Specifically, they choose $\lambda_{\gamma}(\beta) = \lambda\|\beta\|_0 + \gamma\|\beta\|_2^2$ in (4). They developed fast algorithms for this class of problems based on coordinate descent and local combinatorial search, which exhibited outstanding numerical performance among the state-of-the-art sparse learning algorithms in terms of prediction, estimation and exact recovery probability.

Unlike the previous works focusing on exact support recovery, this paper studies the selection behavior of BSS when the true sparsity is underestimated. It is motivated by common practical situations where exact model recovery is hopeless. For instance, in genetic association studies [3], [37], the signal SNPs (single nucleotide polymorphisms) in a genetic construct are likely to be weak at contributing to the phenotypes of interest. As a result, pinpointing all the signal SNPs without false discoveries is almost impossible. Similarly, in astrophysics, detecting true but weak periodic signals is fraught with the risk of discovering false sources [11]. False discoveries are misleading and incur substantial waste of time and resources in scientific investigation; it is thus desirable to identify the few most important features without false discoveries, especially in the presence of weak signals. In Section I-A, we introduce the concept of the early solution path, through which we can assess the accuracy of the first few discoveries of a selection procedure. Our goal is to pursue sure early selection, meaning zero false early selection, which we believe as a realistic and desirable goal in modern practice. Compared with the widely recognized FDR (false discovery rate) control techniques [4], our focus is on the accuracy of a variable selection method at a low target sparsity level, without a pre-specified FDR threshold as guarantee. Moreover, in our analysis for BSS, given that the exact BSS is often intractable, we accommodate potential optimization error in our statistical analysis. Section I-B reviews and proposes several approximate algorithms for BSS, which turn out to exhibit superior FDR-TPR tradeoff over penalized methods on their early solution paths in the numerical experiments. Finally, we summarize the major contributions of the paper in Section I-C.

A. The Early Solution Path

For any two subsets $S_1, S_2 \subseteq [p]$, let $S_1 \setminus S_2$ denote $S_1 \cap S_2^c$. Then for any estimate $\hat{S}$ of the true model $S^*$, the false discovery proportion (FDP) and true positive proportion (TPP) are defined as

$$\text{FDP}(\hat{S}) := \frac{\|\hat{S} \setminus S^*\|}{\max(|\hat{S}|, 1)} \quad \text{and} \quad \text{TPP}(\hat{S}) := \frac{|\hat{S} \cap S^*|}{|S^*|}. \quad (5)$$

The false discovery rate (FDR) and true positive rate (TPR) are defined as the expectations of FDP and TPP respectively. In the context of multiple hypothesis testing, FDR and $1 - \text{TPR}$ are essentially type I and type II errors respectively.

A solution path provides a comprehensive view of a model selection procedure: it displays tradeoff between type I and type II errors of the selected models as the regularization parameter $\lambda$ in (4) or the model size $s$ in (3) varies. In contrast, model consistency requires oracular knowledge of sparsity, which is often unavailable in practice. Moreover, model consistency is often too ambitious a goal to achieve in real-world problems, where the true sparsity or signal strength rarely satisfies the theoretical requirement. Therefore, the tradeoff between type I and type II error is inevitable, rendering the solution path a more meaningful evaluation criteria for comparing variable selectors. In particular, the early solution path, which tracks the first few selected variables, is of interest to scientific research; after all, it is impossible in practice to assess the causality of too many variables by experiments. Therefore, it is imperative to provide statistical guarantee, say FDR, for the early solution path to guide the subsequent scientific effort.

Formally, define the early solution path of BSS as the set $\{\hat{\beta}_{\text{best}}^s(s)\}_{s \leq s^*}$, i.e., the BSS estimators whose sparsity is not greater than $s^*$. In this paper, we explicitly characterize when BSS achieves zero false discovery throughout the early solution path, which we refer to as sure early selection. Regarding related works on solution paths, [34] showed that under the regime of linear sparsity, i.e., $s^*/p$ tends to a constant, even when the features are independent, false discoveries occur early on the LASSO path with high probability, regardless of the signal strength. Reference [43] further provided a complete FDR-TPR tradeoff diagram of LASSO. Reference [35] investigated when the first false variable is selected by sequential regression procedures, which include forward stepwise, the LASSO, and least angle regression. Su’s setup shares similar flavor with this paper, while the theoretical results therein are based on i.i.d Gaussian design. We instead work with fixed designs with possible collinearity. More details are in Section II.

B. Algorithmic Development for BSS

Recent advancement in computing hardware and optimization algorithms has made possible the implementation of BSS for real-world problems. Reference [5] recast the BSS problem (3) as a mixed integer optimization (MIO) problem and showed that for $n, p$ in thousands, a MIO algorithm implemented on optimization softwares such as Gurobi can achieve certified optimality within minutes. Reference [6] devised a new cutting plane method that solves to provable optimality the Tikhonov-regularized [39] BSS problem with $n, p$ in the 100,000s. A more recent work [47] proposed an iterative splicing method called ABESS, short for adaptive best subset selection, to solve the BSS problem. They showed that ABESS enjoys both statistical accuracy and polynomial computational complexity when the design satisfies the sparse
Riesz condition and the minimum signal strength is of order $\Omega\{(s^* \log p \log \log n/n)^{1/2}\}$.

Regarding the statistical performance in numerical experiments, [5] and [6] demonstrated that BSS enjoys higher predictive power and lower false discovery rate (FDR) than LASSO. Reference [47] presented similar numerical results of ABEES and also showed that ABEES is able to estimate the model sparsity more accurately than LASSO, MCP and SCAD. Reference [20] conducted extensive numerical experiments on comparison between LASSO, relaxed LASSO and BSS. They covered a wider lower range of signal-to-noise ratios (SNR) than [5]. The main message therein is that regarding prediction accuracy, LASSO outperforms BSS in the low SNR regime, while the situation is reversed in the high SNR regime. Furthermore, relaxed LASSO [28] is the overall winner, performing similarly or outperforming both LASSO and BSS in nearly all the cases in [20].

In this paper, we introduce a “screen then select” (STS) strategy to approximately solve BSS under high-dimensional sparse models and achieve sure early selection. Specifically, we first apply a computationally cheap model selector, say LASSO, to filter out massive useless variables. Then we run a BSS algorithm, say MIO, on the remaining variables to select the final model. The STS strategy is designed to accommodate the computational hardness of the BSS algorithms by first reducing the dimension for them. One particular STS-type algorithm that we investigate is LBSS (LASSO plus BSS), which uses LASSO to screen variables and MIO to select the model (more details are in Section III-A). To give some flavor of the statistical performance of LBSS, Figure 1 compares the FDR-TPR curves of the solution paths of LBSS, MCP, SCAD, LASSO, PGD [7], [8], [23] and CoSaMP [31] under autoregressive design. There the $x$-axis represents FDR, and the $y$-axis represents TPR. Note that a perfect model selection procedure yields a “T”-shaped FDR-TPR path, meaning that it selects no false variable until enforced to select more than $s^*$ variables. One can see from Figure 1 that the FDR-TPR paths of LBSS and CoSaMP are the closest to the perfect “T”-shape, suggesting their superiority over the competing approaches in terms of FDR-TPR tradeoff. More comparisons of this type can be found in Section IV, and the overall winner is LBSS.

It is worth mention that [44] explored another suite of two-stage variable selection (TVS) techniques. In the initial stage, an estimator is derived by solving equation (4) with $L_q$ penalization, i.e., $\rho_\lambda(\beta) = \lambda \|\beta\|^q$, with $q \geq 1$. Then the solution path is generated by ranking the magnitude of the coefficients in a decreasing order and adding them one by one. An important component in the two-stage algorithms is the selection of the tuning parameter in the first stage. Reference [44] pointed out that the optimal $\lambda$ for selection accuracy also minimizes the asymptotic mean square error (AMSE) of the bridge estimator. Subsequent estimation and tuning process is then introduced. This generic two-stage methodology can potentially benefit variable selection methods other than the $L_q$ approach, e.g., PGD.

C. Major Contributions

Below we summarize the major contributions of our work:

(1) In Section II-A, we establish a sufficient condition for near best subsets to achieve sure early selection for any target sparsity $s \leq s^*$ based on a quantity called minimum projected signal margin. This margin is insensitive to high collinearity of the design matrix and even accommodates degenerate covariance structure of the design.

(2) In Section II-B, we show that throughout the early path, i.e., for any $s < s^*$, the sufficient condition above is near necessary (up to a universal constant) for BSS to achieve sure early selection.

(3) In Section II-C, we explicitly derive a non-asymptotic bound of the minimum projected signal margin under random design to allow for verification of the established sufficient and near necessary conditions.

(4) In Section III-A, we introduce the “screen then select” (STS) strategy to efficiently solve BSS. We show that BSS within a sure screening set can achieve sure early selection. Our numerical experiments show that the early solution path of the STS strategy enjoys superior FDR-TPR tradeoff over competing approaches.

The proof for all the theorems and the technical lemmas are collected in the appendix.

D. Notation

We use regular letters, bold regular letters and bold capital letters to denote scalars, vectors and matrices respectively. Given any two sequences $\{a_n\}, \{b_n\}$ valued in $\mathbb{R}$, we say $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a universal constant $C > 0$ such that $a_n \leq C b_n$. We say $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For any positive integer $a$, we use $[a]$ to denote the set $\{1, 2, \ldots, a\}$. For any column vector $a$ and matrix $A$, we use $a^\top$ and $A^\top$ to denote the transpose of $a$ and $A$ respectively. For any matrix $X \in \mathbb{R}^{n \times p}$, define the projection operator for its column space

$$P_X := X(X^\top X)^+ X^\top,$$

where the superscript symbol $^+$ denotes the Moore-Penrose inverse. We use $X_S$ to denote the submatrix of $X$ with
columns indexed by $S \subset [p]$. For any random variable $X$ valued in $\mathbb{R}$, define

$$||X||_{\psi_2} := \inf \{ t > 0 : \text{E} \exp(X^2/t^2) \leq 2 \}$$
and

$$||X||_{\psi_1} := \inf \{ t > 0 : \text{E} \exp(||X||/t) \leq 2 \}.$$  

For any random vector $x$ valued in $\mathbb{R}^p$, define

$$\|x\|_{\psi_2} := \sup_{\|v\|_2 = 1} \|v^T x\|_{\psi_2}$$
and

$$\|x\|_{\psi_1} := \sup_{\|v\|_2 = 1} \|v^T x\|_{\psi_1}.$$  

For any estimator $\hat{\beta}$ of $\beta^*$, based on (5), we abuse the notation to define

$$\text{FDP}(\hat{\beta}) := \frac{|\text{supp}(\hat{\beta}) \setminus S^*|}{\max (|\text{supp}(\hat{\beta})|, 1)}$$
and

$$\text{TPP}(\hat{\beta}) := \frac{|\text{supp}(\hat{\beta}) \cap S^*|}{|S^*|}.$$  

II. ON THE EARLY SOLUTION PATH OF BEST SUBSET SELECTION

Consider $s \leq s^*$ in (3). Our focus in this section is to establish sufficient and near necessary conditions for BSS to achieve sure early selection, i.e., $\text{FDP}(\hat{\beta}_{\text{best}}^*(s)) = 0$. To start with, we introduce a measure called the minimum projected signal margin to characterize the fundamental difficulty for BSS to achieve sure early selection. Define

$$\mathcal{A}(s) := \{ S : S \subset [p], |S| = s, S \setminus S^* = \emptyset \}$$
and

$$\mathcal{A}^*(s) := \{ S : S \subset S^*, |S| = s \}.$$  

In words, $\mathcal{A}(s)$ is the set of all size-$s$ models with at least one false variable, while $\mathcal{A}^*(s)$ is the set of those without any false variable. We further define $\mathcal{A}_t(s) := \{ S \in \mathcal{A}(s) : |S \setminus S^*| = t \}$ for $t \in [s]$, which collects all the models in $\mathcal{A}(s)$ with exactly $t$ false variables. For any $S_1, S_2 \subset [p]$, the marginal projection operator

$$P_{S_1 \setminus S_2} := P_{S_1} - P_{X_{S_1 \cap S_2}}.$$  

Note that since $S_1 \cap S_2 \subset S_1$, $P_{S_1 \setminus S_2}$ is a projection operator: it captures the directions in the column space of $X_{S_1}$ that are orthogonal to the column space of $X_{S_1 \cap S_2}$, representing the margin of fitting power of $X_{S_1}$ on top of $X_{S_1 \cap S_2}$. Writing $\mu^* = X\beta^*$, we define the optimal feature swap $\Phi : \mathcal{A}(s) \rightarrow \mathcal{A}^*(s)$ as

$$\Phi(S) := \arg\max_{S' \in \mathcal{A}^*(s), |S' \cap S| \leq |S|} \|P_{X_{S_1 \setminus S_2}} \mu^*\|_2^2.$$  

As the name indicates, $\Phi$ swaps the false variables in $S$ for the true ones that maximize the fitting power.

We are now in position to define the projected signal margin of $S \in \mathcal{A}(s)$ as

$$m(S) := \frac{\|P_{X(S)S} \mu^*\|_2 - \|P_{[X(S)S]S} \mu^*\|_2}{\|\Phi(S) \setminus S\|^{1/2}},$$  
and

its minimum over $\mathcal{A}(s)$ as

$$m_*(s) := \min_{S \in \mathcal{A}(s)} m(S).$$  

Intuitively, for any model $S$, $m(S)$ quantifies the enhancement in explanatory power by switching from $S$ to its optimal feature swap, normalized by the square root of the number of false variables in $S$. Consequently, $m_*(s)$ characterizes the minimum normalized gain of explanatory power from the optimal feature swap. The smaller $m_*(s)$, the easier for some $S \in \mathcal{A}(s)$ to outperform $\Phi(S)$ in terms of goodness of fit, giving rise to potential false discoveries. The example below illustrates the rate of the minimum projected signal margin under homogeneous signals and orthonormal design:

**Example 1:** Consider the case where $n \geq p > s^* > 1$, $\beta_j^* = b$, $\forall j \in [s^*]$ and $n^{-1}X^T X = I_p$. Then for any $s < s^*$, we have $m(S) = bn^{1/2}$ for all $S \in \mathcal{A}(s)$, implying that $m_*(s) = bn^{1/2}$. Detailed calculation for this example can be found in Appendix B-H.

Regarding general cases, we refer the readers to Section II-C, where we establish non-asymptotic bounds for projected signal margin under random design. In the sequel, we will see that $m_*(s)$ dictates the statistical behavior of the early path of BSS: the sufficient and necessary condition for BSS to achieve sure early selection essentially boils down to a lower bound of $m_*(s)$. Throughout this section, we assume that $s^* < n$ and that $\|x\|_{\psi_2} \leq \sigma$ in (1).

A. Sufficient Conditions

The goal of this subsection is to explicitly characterize a sufficient condition for $\hat{\beta}_{\text{best}}^*(s)$ to achieve zero false discovery. Given the optimization challenge of obtaining $\hat{\beta}_{\text{best}}^*(s)$ exactly, we extend the scope of our statistical analysis to embrace all near best $s$-subsets, i.e., the size-$s$ subsets that yield comparable goodness of fit as the best subset. Specifically, for any $S \subset [p]$, let $L_S = y^T (I - P_{X_S}) y$ and $L_* = \min_{S \subset [p], |S| = s} L_S$. Given any tolerance level $\eta \in [0, 1)$, consider the following collection of near best $s$-subsets:

$$S(s, \eta) := \{ S : |S| = s, L_S \leq L_* + \eta m_2^2(s) \}.$$  

Here $\eta$ is a tolerance threshold that determines the condition for $S$ to be a near best $s$-subset: the larger $\eta$, the higher fitting error we can tolerate for a near best subset. Below we introduce a sufficient condition for all sets in $S(s, \eta)$ to achieve sure selection.

**Theorem 1:** Suppose that $\log p \geq s^*$. Fix a small $\delta \in (0, 1)$. There exists a universal constant $C \geq 1$, such that for any $0 \leq \eta < 1$, whenever

$$m_*(s) \geq 8\sigma C \left( 1 - \frac{\log(\delta/Cs)}{2\log p} \right)^{1/2} \left( \frac{\log p}{1 - \eta} \right)^{1/2}$$
we have that

$$\text{FDP}(\hat{S}) = 0, \forall \hat{S} \in S(s, \eta) \geq 1 - \delta.$$  

Theorem 1 says that whenever the minimum projected signal margin $m_*(s)$ satisfies the lower bound (9), all the sets in $S(s, \eta)$ are sure selection sets with high probability. Therefore, if BSS does not waste any sparsity quota, i.e., $|\text{supp}(\hat{\beta}_{\text{best}}^*(s))| = s$, then we have $\text{TPP}(\hat{S}) = s^*/s^*$ for all $\hat{S} \in S(s, \eta)$ with high probability. In fact, when the true
features are non-degenerate, meaning that $X_{S^*}$ has full column rank, and when sure early selection happens, BSS always selects the sets with the maximum possible size $s$. This is due to the fact that $\| (I - P_{X_{S^*}}) y \|_2^2 > \| (I - P_{X_{S_{S^*}}}) y \|_2^2$ for any nested subsets $S_1 \subset S_2 \subset S^*$ and any full column rank $X_{S^*}$. However, one should not always expect lower sparsity from more stringent sparsity constraint or regularization: [29] present an example in which the active set of LASSO expands as $\lambda$ increases within certain range of $\lambda$. Consequently, as the target sparsity $s$ increases and gets closer to the true sparsity $s^*$, the TPP guarantee grows higher and finally reaches 1.

To understand the role of $\eta$, fix the problem setup so that $m_\star(s)$ is fixed. By definition (8), the larger $\eta$, the larger family of sure selection sets we have. Therefore, as $\eta$ increases, in order to guarantee zero false discovery for the entire family of near best subsets as in (10), it is natural to require higher $m_\star(s)$, which is consistent with (9). When $\eta$ grows to the extent that (9) breaks, the theory then fails to provide the universal guarantee for all sure selection sets. Regarding $\xi$, by assessing its role on the right hand side of (9) and (10) respectively, we can deduce that larger minimum projected signal margin leads to higher confidence in sure selection.

One important message of this theorem is that the variable selection guarantee for BSS is robust against collinearity. Through the marginal projection operators, Theorem 1 unveils that it is the gap in capacity of capturing the true signal between sure selection sets and sets with false discoveries that determines if any false variable enters the early path of BSS. Given that projection matrices of the features are invariant with respect to linear transformations of them, the minimum projected signal margin is robust against collinearity. For instance, given $a, b \in \mathbb{R}^n$, consider a simple yet illustrative case where $X_{S^*} = a_1^{\top}$ and $X_{S^*\setminus S} = b_1^{\top}$, i.e., all the columns of $X_{S^*}$ equal $a$, and all the columns of $X_{S^*\setminus S}$ equal $b$. This case violates both the irrepresentable condition and the sparse Riesz condition but can satisfy the minimum projected signal margin condition. By applying Theorem 1, one can deduce that as long as $a$ and $b$ are not too correlated and the signal is not too weak, BSS achieves sure early selection with high probability. Section IV presents a wide range of numerical results to demonstrate the robustness of BSS against design correlation.

Now we discuss the difference between our work and a recent related work [19] that studies the model consistency of BSS. The main theoretical innovation of our work given [19] lies in defining the projected signal margin through the optimal feature swap. Reference [19] proposed the following “identifiability margin” to characterize the fundamental difficulty for BSS to achieve model consistency:

$$\tau_\star(s^*) := \min_{S \subset [p], |S| = s^*, S \neq S^*} \frac{\beta^{\top}_s (\Sigma S^\top S S S^\top S S^\top S S^\top S S^\top S S^\top S) \beta}{|S^\star \setminus S|},$$

where $\Sigma := n^{-1} X^\top X$ and for any two sets $S_1, S_2 \subset [p]$, $\Sigma_{S_1, S_2}$ is the submatrix of $\Sigma$ containing rows indexed by $S_1$ and columns indexed by $S_2$. The most essential difference between the identifiability margin and our projected signal margin is that for any set $S$ with false discoveries, the former compares the projected signal strength of $S$ with that of $S^*$, while the latter compares $S$ with its optimal swap set $\Phi(S)$. Note that to ensure BSS to recover $S^*$ with known sparsity $s^*$, naturally it suffices to impose a large gap of signal strength between $S^*$ and the rest of the models of size $s^*$. However, on the early path of BSS, there are multiple ($\binom{n}{s}$, more precisely) sure selection sets of size $s$, and it is non-trivial to identify which one to refer to to characterize the fundamental difficulty of achieving sure selection. Instead of fixing one sure selection set as the reference, we propose the optimal feature swap to pick the most similar and powerful sure selection set $\Phi(S)$ for any set $S$ and compute their projected signal strength margin. This turns out to be the measure of fundamental difficulty for BSS to attain sure early selection (see Theorems 2 and 3).

Another difference between the two margins $\tau_\star$ and $m_\star$ is that the numerator of $m_\star$ is the difference between the $\ell_2$-norms of projected signals instead of the squared $\ell_2$-norms of them in $\tau_\star$. Because of this difference, when $s = s^*$, $m_\star$ is narrower than $\tau_\star$ after normalization:

$$m_\star^2(s^*) = \min_{S \subset [n], |S| = s^*} \frac{\|P_{S^\star \setminus S} \mu^*\|_2^2 - \|P_{S|S} \mu^*\|_2^2}{|S| S^\star},$$

$$\leq \min_{S \subset [n], |S| = s^*} \frac{\|P_{S^\star \setminus S} \mu^*\|_2^2 - \|P_{S|S} \mu^*\|_2^2}{|S| S^\star},$$

$$\leq \min_{S \subset [n], |S| = s^*} \frac{\|\mu^*\|_2^2 - \|P_{S|S} \mu^*\|_2^2}{|S| S^\star} = n \tau_\star(s^*),$$

where the inequality is due to the fact that $\|P_{S^\star \setminus S} \mu^*\|_2 \geq \|P_{S|S} \mu^*\|_2$. Therefore, $m_\star(s^*) \gtrsim \sigma(\log p)^{1/2}$ as imposed in (9) is a stronger condition than $\tau_\star(s^*) \gtrsim \sigma^2 \log p/n$, which was shown by [19] as a sufficient and near necessary condition for BSS to recover the true support. As we shall see in Section II-B, this stronger form of margin is not due to technical limitation but rather the greater difficulty of achieving sure early selection than model consistency. The early solution path underrates underestimates the true sparsity $s^*$ and thus requires wider margin of signal strength to compensate for model misspecification and identify true variables.

Now we discuss a series of literature on the solution path of variable selectors under high dimension: [34], [35], and [43]. Reference [34] studied the early solution path of LASSO. Define the LASSO estimator $\hat{\beta}_\text{lasso}(\lambda)$ as

$$\hat{\beta}_\text{lasso}(\lambda) := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$
tends to one, where \( q^*(0) = 0 \). We refer the readers to [34] for specific definition and visual illustration of \( q^* \). Simply speaking, this implies that false discoveries occur early on the solution path of LASSO however strong the signal is. While it is tempting to conclude theoretical superiority of BSS over LASSO on the early solution path, we emphasize that our Theorem 1 is actually not directly comparable with Theorem 2.1 of [34], given that we mainly focus on the ultra-high dimensional setup, where \( \log p \gtrsim s^* \), while [34] focused on linear sparsity regimes. Nevertheless, the simulation experiments in Section IV clearly demonstrate the numerical superiority of BSS over LASSO in terms of FDR and TPR on their early solution paths. A follow-up work [43] characterized the complete LASSO tradeoff terms of FDR and TPR on their early solution paths. A follow-up work [43] characterized the complete LASSO tradeoff diagram, which shows not only a lower bound on FDR but also an upper bound on TPR. This further explores the limitation of LASSO in terms of FDR-TPR tradeoff. Finally, [34] considered the \( L_0 \)-penalized estimator:

\[
\hat{\beta}_{L_0} = \arg \min_{\beta \in \mathbb{R}^p} |y - X\beta|_2^2 + \lambda \|eta\|_0
\]

and shows that under the linear sparsity regime with the signal strength going to infinity, \( \hat{\beta}_{L_0} \) can achieve (FDP, TPR) = (0,1) asymptotically with a proper choice of \( \lambda \). Note that this result revolves around exact model recovery and does not characterize the early solution path. It thus remains open how the early solution path of BSS behaves under the linear sparsity setup and how that is compared with LASSO.

Reference [35] considered the rank of the first false variable selected by sequential regression procedures including forward stepwise regression, LASSO, and the least angle regression. It assumes that \( X \in \mathbb{R}^{n \times p} \) has independent \( \mathcal{N}(0,1/n) \) entries, \( \epsilon \sim \mathcal{N}(0, \sigma^2 I) \) and nonzero components in \( \beta^* \) are equal to some \( M \neq 0 \). It focuses on the regime of near linear sparsity in the sense that \( c_1p/\log p \leq n \leq c_3p \) and that \( c_4n \leq s^* \leq \min\{0.99p, c_5n \log(99p)/\log p\} \) for arbitrary positive constants \( \{c_i\}_{i \in [5]} \). Let \( T \) denote the rank of the first false variable selected by any of the aforementioned three sequential regression methods. Reference [35] showed that as long as \( \sigma/\sqrt{M} \to 0 \), \( \log T = \{1 + o_P(1)\} \{2n(\log p)/s^*\}^{1/2} - n/(2s^*) + \log(n/(2p \log p)) \} \). When \( n \approx p \approx s^* \), it means that each of the sequential methods includes the first false variable after no more than \( O(\exp(\{\log s^*\}^{1/2}) \) steps asymptotically, thereby failing to achieve sure early selection. In our work, we show that BSS can achieve sure early solution with high probability whenever the projected signal margin exceeds the lower bound as in (9). Although we consider the ultra-high dimensional regime, which is far from the linear sparsity regime and makes our results not directly comparable with the results in [35], we emphasize that our theory does not require statistical independence between the features and accommodates nearly degenerate designs, while the results in [35] rely on independent Gaussian designs.

Beyond the guarantee for absolutely sure selection as in Theorem 1, we can extend our argument to provide any pre-specified level of FDP guarantee through adjusting the definition of the projected signal margin. Given any \( 0 < \eta < 1 \), consider the following variant of the minimum projected signal margin that corresponds to \( q \)-level FDP:

\[
m_q(s) := \min_{S \subseteq \cup_{i \leq [q]} A_i(s)} m(S),
\]

as well as the corresponding collection of near best \( s \)-subsets:

\[
S_q(s, \eta) := \{S : |S| = s, \mathcal{L}_S \leq \mathcal{L}_* + \eta m_q^2(s)\}.
\]

Now we are ready to present the following corollary on \( q \)-level FDP guarantee.

**Corollary 1:** Suppose that \( \log p \gtrsim s^* \). Fix a small \( \delta \in (0,1) \). There exists a universal constant \( C \geq 1 \), such that for any \( 0 \leq \eta < 1 \), whenever

\[
m_q(s) \geq 8\sigma C \left\{ 1 - \frac{\log(\delta/Cs)}{2 \log p} \right\}^{1/2} \frac{(\log p)^{1/2}}{1 - \eta},
\]

we have

\[
\mathbb{P}\{\text{FDP}(\hat{S}) \leq q, \forall \hat{S} \in S_q(s, \eta)\} \geq 1 - \delta.
\]

Corollary 1 is a variant of Theorem 1 by replacing \( A_i(s) \) in Theorem 1 with \( \cup_{i \leq [q]} A_i(s) \).

**B. Near Necessary Conditions**

In this section, we show that the order of the lower bound (9) for the minimum projected signal margin is near necessary (up to a universal constant) for BSS to achieve sure early selection in general setups (without the assumption that \( \log p \gtrsim s^* \)). The word “near” means to distinguish the condition from a strictly necessary condition: one might still achieve sure early selection if there are only few models in \( A(s) \) whose projected signal margins are below the required magnitude. In this section, we demonstrate that if there are sufficiently many models in \( A(s) \) whose projected signal margins fall below the order in (9), sure early selection is doomed to fail for BSS. We investigate the necessity in two scenarios separately: (i) \( s = 1 \); (ii) \( s \geq 2 \).

1) **Sure First Selection:** We start with the simplest case: \( s = 1 \), i.e., we aim to find only one true predictor. Let

\[
\hat{\gamma} := \arg \max_{j \in S^*} |\|X_j \mu^*\|_2^2|
\]

which denotes the best variable in \( S^* \) in terms of fitting the signal \( \mu^* \). Then for any \( j \in (S^*)^c \), \( \Phi(\{j\}) = \{\hat{\gamma}\} \) and

\[
m(\{j\}) = |\|X_j \mu^*\|_2^2 - |\|X_j \mu^*\|_2^2|.
\]

We say a set \( P \) is a \( \delta \)-packing set if and only if for any \( u, v \in P \), \( u - v \geq \delta \). Consider the following two assumptions on the problem setup:

**Assumption 1:** Let \( u_j = X_j/\|X_j\|_2 \) for \( j \in [p] \setminus S^* \). For some \( 0 < \delta_0 < 1 \), there exists an index set \( \mathcal{J}_{\delta_0} \subset (S^*)^c \) of delusive features satisfying:

(i) \( \{u_j\}_{j \in \mathcal{J}_{\delta_0}} \) is a \( \delta_0 \)-packing set;
(ii) \( |\mathcal{J}_{\delta_0}| \geq p^\alpha \) for some \( 0 < c_0 < 1 \);
(iii) Each feature in \( \mathcal{J}_{\delta_0} \) has small projected signal margin, i.e.,

\[
0 < m(\{j\}) < \frac{\delta_0 \sigma (c_0 \log p)^{1/2}}{20}, \forall j \in \mathcal{J}_{\delta_0}.
\]
Assumption 2: There exists a universal constant $\xi > \max\{8/(\delta(s_0)/2 \log p), s^{-1}\}$ such that
$$\|P_{X_j} \mu^\ast\|_2 \geq \xi\delta(s_0) \sigma(\log p)^{1/2}.$$ 

The main idea of Assumption 1 is that we can find sufficiently many delusive features with small projected signal margin as per (13). Furthermore, perceiving Euclidean distance between two vectors within $\{u_j\}_{j \in [p]}$ as the correlation proxy between the corresponding features, condition (ii) requires the delusive variables to be dissimilar to each other, ensuring that the impact of the dimension is nontrivial. Assumption 2 says that the best single variable has sufficient explanation power for the true signal $\mu^\ast$. This assumption is standard: one can deduce from the well-known $\beta$-min condition, i.e., $\min_{j \in S} |\beta_j^\ast| \geq \sigma(\log p/n)^{1/2}$, that $\sigma(\log p)^{1/2}$ is the minimum signal strength we need for $\|P_{X_j} \mu^\ast\|_2$ to identify $j^\ast$ as a true predictor.

To help our readers understand the assumptions, below we give a specific problem setup that satisfies the assumptions.

**Example 2:** Consider $n = [2p^n/n]$ for some $\epsilon > 0$ and $S^\ast = [S^\ast]$ such that $S^\ast \subseteq n/2$. Let $e_j \in \mathbb{R}^n$ be an orthogonal basis such that $e_j \in [n]$ and $e_j e'_{j} = 0$ for any $j \neq j'$. Let $X_j = \sqrt{n}e_j$ for $j \in S^\ast$ and $X_j = \sqrt{n}\epsilon_1$ for $j \in [n] \setminus S^\ast$ and some $\epsilon \in (0, 1)$. We do not make any assumption on $\{X_j\}_{j \in [p]}$. Under such design, the features in $S^\ast$ are orthogonal to each other and the features in [n] \ S^\ast are correlated with $X_1$, where the correlation depends on $\epsilon$. Fix $\gamma > \beta > 0$. Then $\beta_j^\ast = \gamma$ and $\beta_j^\ast = \beta$ for $j \in S^\ast \setminus \{1\}$. Then we have $\|P_{X_j} \mu^\ast\|_2 = \sqrt{n}\beta$, $\|P_{X_j} \mu^\ast\|_2 = \sqrt{n}\epsilon_1$ for $j \in S^\ast \setminus \{1\}$ and accordingly $j^\ast = 1$. $\{u_j\}_{j \in [n]}$ is a $\sqrt{2}(1 + q^2)^{-1/2}$-packing set and $\|n\ | S^\ast | n - S^\ast | \geq n/2 \geq \beta^\ast$. For each $j \in [n] \setminus S^\ast$, we have $m(j) = \sqrt{n}(1 - 1 + q^2)^{-1/2} \geq 0$. Then Assumptions 1 and 2 are satisfied as long as
$$\frac{\sqrt{2q}\sigma(\log p)^{1/2}}{1 + q^n}/2 \leq \gamma < \frac{\sqrt{2q}(\log p/n)^{1/2}}{20(1 + q^2)^{1/2} - 20}.$$

(14) can hold when $q$ is small enough, which indicates that $X_1$ and $X_j$ are similar for all $j \in [n] \setminus S^\ast$. By fixing $n, p$ and letting $q \to 0$, we have $0 < \gamma < \infty$ for (14). Notice that $\beta$ does not appear in (14), which indicates that we do not need further assumptions on the weak signals to satisfy Assumptions 1 and 2.

We refer our readers to Appendix B-H for calculation details for this example. The construction of this example is for interpretation purpose: one can easily extend it to a more general class of problem designs following the construction logic. Now we present our theorem on the near necessary condition for the very first selection of BSS to be true.

**Theorem 2:** Suppose that $\{e_j\}_{j \in [n]}$ are independent Gaussian noise with variance $\sigma^2$. Under Assumptions 1 and 2, we have that
$$\mathbb{P}(\min_{j \in J_{(\epsilon)}} \mathcal{L}(j) < \min_{j \in S^\ast} \mathcal{L}(j^\ast)) \geq 1 - 4s^2p^{-c_1}(c_0/2 - 1) - c_0p^{-c_1}c_0,$$
where $c_1$ is a universal constant.

Given Theorem 2, one can deduce by comparing (9) and (13) that the lower bound of the minimum projected signal margin in Theorem 1 is near necessary: whenever there are sufficiently many spurious features whose projected signal margins violate this lower bound up to a multiplicative constant, the first shot of BSS is false with high probability.

2) **Sure Early Multi-Selection:** Now we move on to the general case when $s \geq 2$. Define $S^\ast(I)$ to be the best size-$s$ subset of $S^\ast$ in terms of fitting the signal $\mu^\ast$, i.e., $S^\ast(I) := \arg\max_{S \subseteq \mathbb{A}(s)} \|P_{S} \mu^\ast\|_2^2$. For simplicity, we write $S^\ast(I)$ as $S^\ast$ in the sequel. Consider the following two assumptions on the problem setup:

**Assumption 3:** There exist $j_0 \in S^\ast$ and $0 < \delta_0 < 1$ such that if we let $\delta_0 = S^\ast \setminus \{j_0\}$, $u_j := (I - P_{S^\ast(I)} X_j)$ and $u_j = \sqrt{n}e_j$ for $j \in [p] \setminus S^\ast$, we can find an index set $J_{(\epsilon)} \in (S^\ast)^c$ of delusive features satisfying:
(i) $\{u_j\}_{j \in J_{(\epsilon)}}$ is a $\delta_0$-packing set;
(ii) $|J_{(\epsilon)}| \geq \rho_0$ for some $\epsilon < \epsilon_0 < 1$;
(iii) Consider all the models of size $s$ that are formed by replacing $j_0$ in $S^\ast$ with a spurious variable in $J_{(\epsilon)}$, namely, $A_{j_0} := \{S^\ast \cup \{j\} : j \in J_{(\epsilon)}\}$. Then each set in $A_{j_0}$ has small projected signal margin, i.e.,
$$0 < m(S) < \frac{\delta_0\sigma(\log p)^{1/2}}{20}, \forall S \in A_{j_0}.$$

**Assumption 4:** There exists a universal constant $\xi > \max\{8/(\sigma(\log p)/2 \log p), s^{-1}\}$ such that
$$\min_{S \in \mathbb{A}(s), S \neq S^\ast} \|P_{S^\ast} \mu^\ast\|_2 - \|P_{S} \mu^\ast\|_2 \geq \xi(\log p)^{1/2}.$$

Assumption 3 is similar to Assumption 1; the only difference is that in Assumption 3, each delusive model is not a singleton but a set constructed by replacing $j_0$ in $S^\ast$ with a spurious variable in $J_{(\epsilon)}$. Assumption 4 solely involves the true support $S^\ast$ and signal $\mu^\ast$. It guarantees that $S^\ast$ significantly outperforms all the other size-$s$ sure selection sets in terms of fitting $\mu^\ast$, thereby being the best $s$-subset within $S^\ast$ with high probability. In this vein, if we can find a size-$s$ model $S$ outside $S^\ast$ that fits $y$ better than $S^\ast$, BSS then favors $S$ and selects false variables. Following Example 2, we give a specific problem setup that satisfies Assumptions 3 and 4.

**Example 3:** Consider the same design matrix as in Example 2. Fix $\gamma > \beta > 0$. Let $\beta_j^\ast = \gamma$ for $j \in [s]$ and $\beta_j^\ast = \beta$ for $j \in S^\ast \setminus [s]$. Then $S^\ast(s) = [s]$, $\{u_j\}_{j \in [n]}$ is a $\sqrt{2q}(1 + q^2)^{-1/2}$-packing set and $\|n\ | S^\ast | n - S^\ast | = n - s^\ast \geq n/2 \geq \beta^\ast$. Let $j_0 = 1$. For each $S \in \mathcal{A}_{j_0} := \{S^\ast \cup \{j\} : j \in [n] \setminus S^\ast\}$, we have $m(S) = \sqrt{n}(1 - (1 + q^2)^{-1/2}) \gamma$. Also we have
$$\min_{S \in \mathbb{A}(s), S \neq S^\ast} \|P_{S^\ast} \mu^\ast\|_2 - \|P_{S} \mu^\ast\|_2 \geq \sqrt{n}(\gamma - \beta).$$

Accordingly, Assumptions 3 and 4 are satisfied if $p \geq 4$ and
$$\xi(\log p/n)^{1/2} \leq \gamma < \frac{\sqrt{2q}(\log p/n)^{1/2}}{20(1 + q^2)^{1/2} - 20},$$
which can also hold when $q \to 0$ and $p$ is large enough. Calculation details for this example and discussion for $q \to 0$ can be found in Appendix B-H.

The following theorem shows that with high probability, the best subset involves false selection once there are sufficiently
many models in $\mathcal{A}(s)$ with small projected signal margin as per (15). Therefore, together with Theorem 2, Theorem 3 shows the near necessity of the lower bound of the minimum projected signal margin (9) up to a constant uniformly over $s \in [s^*-1]$.

**Theorem 3:** Suppose that $\{e_i\}_{i \in [n]}$ are independent Gaussian noise with variance $\sigma^2$ and $s \geq 2$. Under Assumptions 3 and 4, we have that

$$\mathbb{P}\left(\min_{S \in A_{\delta}} \mathcal{L}_S < \min_{S \in A\beta} \mathcal{L}_S\right) \geq 1 - 4p^{-(c_1\delta^2_0-1)c_0+p^{-c_1\delta^2_0}c_0-6sp^{-r(2\xi^2-2)}},$$

where $c_1$ and $c_2$ are universal constants.

### C. Projected Signal Margin Under Random Design

The previous two subsections demonstrate the pivotal role of the minimum projected signal margin $m_\ast(s)$ in underpinning the sure selection of BSS. In this subsection, we explicitly derive $m_\ast(s)$ under random design, so that we can verify the sufficient condition (9) with ease. The following theorem gives an explicit lower bound of $m_\ast(s)$ when we have isotropic Gaussian design and $\beta^\ast$ has homogeneous entries.

**Theorem 4:** Consider $n$ independent observations $\{x_i\}_{i \in [n]}$ of $\mathcal{N}(0, I_p)$. Suppose that $\log p \gtrsim s^*$, and that $\beta^\ast_j = \beta, \forall j \in S^\ast$. Fix a small $\delta \in (0, 1)$. Then there exist universal constants $c > 0$ and $C > 1$ such that whenever $n \geq \left\{3 - \log(\delta/Cs^*)/\log(p)\right\}^2C^2s^*(\log(p))^2$, we have

$$\mathbb{P}\{m_\ast(s) \geq n^{1/2}c|\beta|\} \geq 1 - \delta.$$ 

Consequently, whenever

$$n \geq \left\{3 - \log(\delta/Cs^*)/\log(p)\right\}^2C^2s^*(\log(p))^2,$$

we have

$$\mathbb{P}\{m_\ast(s) \geq n^{1/2}c|\beta|, \forall s \in [s^* - 1]\} \geq 1 - \delta.$$ 

Theorem 4 provides an explicit characterization of the projected signal margin under independent Gaussian design. Combining Theorem 1 and Theorem 4, we see that $|\beta| \gtrsim \sigma(\log p/n)^{1/2}$ is sufficient for BSS to achieve sure early selection with high probability in this case. This lower bound for $|\beta|$ is aligned with the well-known $\beta$-min condition for support recovery. Specifically, Theorem 3.3 in [24] shows that when $\min_{j \in S^\ast} |\beta^\ast_j|/\sigma(\log p/n)^{1/2}) \rightarrow \infty$ as $n, p \rightarrow \infty$, one can achieve asymptotic power one through a false discovery rate control procedure based on debiased LASSO. Theorem 2 in [41] shows the necessity of the $\beta$-min condition in identifying true variables: if the $\beta$-min condition is violated, the failure probability for variable selection consistency can be lower bounded by $1/2$. Note that the novelty of Theorem 4 given the previous works is that it complements the selection accuracy guarantee for the entire early solution path when $s \in [s^* - 1]$ rather than for only the case of $s = s^\ast$.

Theorem 4 demonstrates that our lower bound (9) for the projected signal margin is achievable under independent Gaussian design. One can actually extend Theorem 4 to any covariance structure of the design; we choose to focus on the independent design for simplicity of presentation and comparison of the results. The independent Gaussian design has been frequently used to investigate the solution path of variable selectors, e.g., [34], [35], and [43]. For more general results, we refer our readers to Lemma 1 in Appendix B-E that discusses the distribution of the projected signal margin under arbitrary $\beta^\ast$ structure and covariance of $X$.

### III. Algorithmic Strategies

Sure early selection is an appealing property of BSS. Nevertheless, it can be difficult to achieve in practice because of the computational difficulty of exact BSS. In this section, we introduce three computational strategies for BSS to pursue sure early selection. Section III-A proposes a “screen then select” strategy that runs exact BSS within only the features that pass a preliminary screening. Section III-B introduces the iterative hard thresholding algorithms, exemplified by projected gradient descent [7], [8] and CoSaMP [31], as computational surrogates for BSS.

#### A. Screen Then Select

Recently, [5] proposed a MIO (Mixed Integer Optimization) formulation for best subset selection, which is amenable to a group of popular optimization solvers including CPLEX, GLPK, MOSEK and GUROBI. These solvers can handle BSS problems with $n, p$ in 1000s within minutes. Nevertheless, $p$ can go far beyond thousands in ultrahigh-dimensional problems, for which applying the MIO algorithm is again computationally burdensome or even infeasible.

To resolve the issue of high dimension, we consider the following “screen then select” (STS) strategy. In the screening stage, we generate a screening set $\widehat{S} \subset [p]$ of size $\tilde{s} > s$ through a preliminary feature screening procedure, which can be but is not limited to penalized least squares methods [13], [38], [45] and sure independence screening [15]. In the selecting stage, we solve the exact BSS problem with only the screened features, i.e., $X_{\widehat{S}}$, to finalize the model selection. Define the collection of near best $s$-subsets within $\widehat{S}$ as

$$\widehat{S}(s, \eta, \tilde{S}) := \{S \subset \widehat{S} : |S| = s, \mathcal{L}_S \leq \tilde{L}_\ast + \eta m^2_\ast(s)\},$$

where $\tilde{L}_\ast := \min_{S \subset \widehat{S}, |S| = s} \mathcal{L}_S$ is the optimal objective of the BSS problem on $\widehat{S}$, and where $\eta$ controls the tolerance level of optimization error. Define the sure screening event as $\mathcal{E} := \{S^\ast \subset \widehat{S}\}$. The following theorem shows that under event $\mathcal{E}$, near best $s$-subsets within $\widehat{S}$ can achieve sure early selection with high probability.

**Theorem 5 (Post-Screening Sure Early Selection):** Suppose that $\log p \gtrsim s^\ast$. Fix a small $\delta \in (0, 1)$. There exists a universal constant $C \geq 1$, such that for any $0 \leq \eta < 1$, whenever

$$m_\ast(s) \geq 8\sigma Cs\left\{1 - \frac{\log(\delta/Cs^*)}{2\log(p)}\right\}^{1/2} \frac{(\log(p))^{1/2}}{1 - \eta},$$

we have that

$$\mathbb{P}\{\text{FDP}(\widehat{S}) = 0, \forall \widehat{S} \in \widehat{S}(s, \eta, \tilde{S})\} \geq 1 - \delta - \mathbb{P}(\mathcal{E}^c).$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Theorem 7.21 in [42] says that LASSO enjoys sure screening when the design satisfies the sparse Riesz condition and the irrepresentable condition and when the signals are not too weak. This means that under these conditions, if we use LASSO as the feature screener in the STS strategy, LASSO can generate an \( \tilde{S} \) such that \( P(\tilde{E}^c) \to 0 \) as \( n, p \to \infty \). Combining this with the theorem above then implies that any near best \( s \)-subset of \( \tilde{S} \) achieves sure early selection. This motivates LBSS (LASSO plus BSS) that uses LASSO as the feature screener and then runs exact BSS for further selection. Section IV numerically shows that LBSS achieves as the feature screener and then runs exact BSS for further selection. This motivates LBSS (LASSO plus BSS) that uses LASSO to \([7]\) and \([8]\), which is basically a projected gradient descent \( \ell_2 \)-norm of \( \beta \). The \( \ell_2 \)-norm of \( \beta \) is a measure of the size of the model. Generally speaking, given a model selection procedure with tuning parameter \( \lambda \) regularized; the tuning parameter \( \lambda \) is the projection size \( \Pi \). PGD enjoys near-optimal error guarantee within few iterations for readers’ reference. Consider a set of projection sizes \( \{ \pi_1, \ldots, \pi_M \} \) with \( \pi_1 < \ldots < \pi_M \). For any \( i \in [M] \), let \( \hat{\beta}^\lambda(i) \) denote the solution of algorithm \( \lambda \) with \( \pi = \pi_i \). Instead of computing \( \hat{\beta}^\lambda(i) \) for each \( i \in [M] \) separately, we propose to use \( \hat{\beta}^\lambda(i) \) as a warm initializer to compute \( \hat{\beta}(i+1) \) (see line 3 of Algorithm 1), which turns out to substantially accelerate the convergence in our numerical experiments.

### Algorithm 1 \( \lambda \) - Path(\( \Xi_\lambda, \Pi \))

**Input:** \( \Xi_\lambda \), projection size set \( \Pi = \{ \pi_1, \ldots, \pi_M \} \) with \( \pi_1 < \ldots < \pi_M \).

1. \( \hat{\beta}^\lambda(0) = 0 \)
2. for \( i = 1:M \)
   1. \( \hat{\beta}^\lambda(i) \leftarrow \hat{\lambda}(\Xi_\lambda, \hat{\beta}^\lambda(i-1), \pi = \pi_i) \)
3. end for

**Output:** \( \{ \hat{\beta}^\lambda(1), \hat{\beta}^\lambda(2), \ldots, \hat{\beta}^\lambda(M) \} \)

In Section IV, we see that LBSS performs overall the best in terms of the FDR-TPR tradeoff on the early solution path among all the investigated methods including PGD, CoSaMP, LASSO, etc.

### IV. NUMERICAL EXPERIMENTS

This section numerically investigates the FDR and TPR of the early paths of LBSS, PGD, COSAMP as well as penalized methods including LASSO, SCAD, MCP and \( L_0L_2 \). Generally speaking, given a model selection procedure with tuning parameter \( \lambda \in \mathbb{R} \), the FDR-TPR path is formed by the set \( \{ \text{FDR}(S(\lambda)), \text{TPR}(S(\lambda)) \}_{\lambda \geq 0} \), where \( S(\lambda) \) is the output of the model selector with tuning parameter \( \lambda \). For LBSS, the tuning parameter \( \lambda \) is the target sparsity of best subset selection, while the screening size is specified as we discussed in Section III-A (around 300 variables in Section IV-A and 40 variables in Section IV-B). For both PGD and CoSaMP, \( \lambda \) is the projection size \( \pi \). Regarding the expansion size \( l \) in CoSaMP, we set it to be the size of the model selected by MCP that is tuned by 10-fold cross validation (CV) in terms of the mean squared error (MSE). For \( L_0L_2 \), \( \lambda \) is the tuning parameter for \( L_0 \) regularization; the tuning parameter for \( L_2 \) regularization, which we denote by \( \gamma \) after display (4), is chosen oracularly: we found that \( \gamma = 10^{-5} \) yields the best FDR-TPR curve. Note that a perfect FDR-TPR path is in the “T” shape, meaning that the method keeps recruiting (nearly) true predictors until having them all.
In the following, we evaluate the solution paths of all the approaches on both synthetic data and a skin cutaneous melanoma dataset. We use the R package bestsubset to solve exact BSS, the R package picasso [22] to implement L0L2. We simulate data as described in Section 4.1. We use the R package L0Learn [22] to implement LASSO, SCAD and MCP, and the R package L0L2 for solving exact BSS, the R package CoSaMP has no false discovery until TPR exceeds 80% but then recruits true variables in a relatively slow manner that prohibits it from selecting all the true variables even when the FDR is as high as 50%. LASSO instead has false discoveries relatively early on the solution path.

\section{A. Simulated Data}

\subsection{1) Design Correlation:} In this section, we investigate two different designs for numerical comparison: the autoregressive design and the equicorrelated design. Specifically, we set $p$ different designs for numerical comparison: the autoregressive design.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{FDR-TPR paths under autoregressive design. We set $\Sigma_{jk} = \rho^{j-k}$ for $\rho \in \{0, 0.5, 0.8\}$ and $\sigma \in \{0.5, 1\}$. The diamonds represent the solutions of different methods that are tuned by 10-fold cross validation.}
\end{figure}

\subsection{2) Equicorrelated Design.} We consider $\Sigma = \rho \mathbf{1}^\top + (1 - \rho) \mathbf{1}$ with $\rho \in \{0.5, 0.8\}$ and two noise levels $\sigma \in \{0.5, 1\}$. This is a much more challenging case than the previous one because of constant correlation. The FDR-TPR paths are presented in Figure 3. We have the following observations:

(i) In the independent settings ($\rho = 0$), LBSS, PGD and L0L2 perform similarly and outperform the other approaches on the early solution path. In particular, when $\rho = 0$ and $\sigma = 0.5$, both LBSS and PGD exhibit a “T”-shaped FDR-TPR paths. CoSaMP has no false discovery until TPR exceeds 80% but then recruits true variables in a relatively slow manner that prohibits it from selecting all the true variables even when the FDR is as high as 50%. LASSO instead has false discoveries relatively early on the solution path.

(ii) As the correlation between features increases, i.e., when $\rho \in \{0.5, 0.8\}$, PGD performs worse. CoSaMP and L0L2, in contrast, still have comparable early selection power as LBSS. Overall, LBSS performs the best in terms of the FDR-TPR tradeoff on the early solution path.

(iii) Regarding the TPR performance, when FDR is controlled at 20%, which is a widely used level, LBSS always enjoys the highest TPR.
CoSaMP and LBSS maintain to yield a nearly vertical FDR-TPR path at the early stage when \( \rho = 0.5 \) and \( \sigma = 0.5 \), meaning that most of their early discoveries are genuine. However, with higher design correlation or stronger noise, their performance deteriorates.

(iii) LASSO, SCAD and MCP select much more false variables than LBSS and CoSaMP on their early solution path. However, interestingly, at a latter stage, the non-convex penalty in SCAD and MCP corrects the models by replacing the false discoveries with true variables, leading to a north-west pivot of the path as \( \lambda \) decreases. LASSO instead does not exhibit such correction: its FDR-TPR path keep moving in the north-east direction as \( \lambda \) decreases, meaning that its FDR keeps increasing. PGD totally breaks down in such challenging cases with strongly dependent designs.

(iv) The performance of L0L2 is outstanding when \( \rho = 0.8 \) and \( \sigma = 0.5 \). In spite of the false discoveries on the early path, L0L2 gradually eliminates them and recruits true variables as \( \lambda \) decreases, so that in the latter part of the path, L0L2 can achieve more than 80% TPR given FDR is below 20%. We attribute this advantage to the \( L_2 \) regularization of L0L2 that guards against strong correlation among the features.

2) Homogeneous Signal Structure: In this section, we investigate the homogeneous signal structure, which is different from the signal structure considered in Section IV-A.1. We keep \( p = 5,000 \), \( s^* = 50 \) and \( n = [2s^* \log p] \). In each Monte Carlo experiment, we generate \( S^* \subset [p] \) by selecting \( s^* \) locations from \( [p] \) with uniform probability and generate \( \beta^* \) such that \( \beta^*_j = 0 \) for \( j \in (S^*)^c \) and that \( \beta^*_j = b \) for \( j \in S^* \), where \( b \in \{0.2, 0.3, 0.4\} \). Rows of the design matrix \( X \) are independently generated from \( \mathcal{N}(0, \Sigma) \), where \( \Sigma_{ij} = 0.5^{\lvert i-j \rvert} \) for \( i, j \in [p] \). We then generate \( y \) by \( y = X\beta^* + \epsilon \) with \( \epsilon \sim \mathcal{N}(0, I) \). In Figure 4, each point on the presented FDR-TPR curve represents the average of FDPs and TPPs of the estimators obtained in 100 independent Monte Carlo experiments, using the same tuning parameter. We have the following observations:

(i) For \( b = 0.4 \) and \( b = 0.3 \), all the methods perform reasonably well, achieving a TPR of 1 before the FDR surpasses 20%. LASSO and SCAD deliver similar results and their FDR-TPR trade-off is less favorable compared with other methods.
Fig. 4. FDR-TPR paths under the homogeneous signal structure. We set the noise level \( \sigma = 1 \) and choose \( \beta \in \{0.4, 0.3, 0.2\} \). The diamond points represent the 10-fold CV solutions.

(i) When \( b = 0.2 \), LBSS and L0L2 stand out with exceptional performance: a nearly vertical FDR-TPR path. PGD also delivers comparable results. CoSaMP, unlike other \( L_0 \)-type algorithms, presents a less desirable FDR-TPR trade-off. LASSO, SCAD and MCP paths pick false variables early.

B. Semi-Simulation With the Skin Cutaneous Melanoma Dataset

In this section, we take our design matrix \( X \) from the skin cutaneous melanoma dataset in the Cancer Genome Atlas (https://cancergenome.nih.gov/), which provides comprehensive profiling data on more than thirty cancer types [36]. The dataset contains \( p = 20,351 \) items of mRNA expression data of \( n = 469 \) patients. To choose reasonable locations of true variables, we adopt the top 20 genes that are found highly associated with cutaneous melanoma according to meta-analysis of 145 papers [9]. One can find a list of these genes at http://bioserver-3.bioacademy.gr/Bioserver/melGene/. We conduct a semi-simulation using this real design matrix. Fixing \( S^* \) to be the locations of the top 20 genes, we generate \( \beta^* \) by letting \( \beta_j^* = 0 \) for \( j \in (S^*)^c \) and \{\( \beta_j^*/\beta_{\min} \)\} \( j \in S^* \) \( \sim \chi_1^2 \), where \( \beta_{\min} \in \{1, 0.5, 0.25\} \). Then we generate the response vector by letting \( y = X\beta^* + \epsilon \), where \( X \) is the standardized \( X \), and where \( \epsilon \sim N(0, I) \).

Figure 5 presents the solution paths of all the methods investigated in Section IV-A. We have the following observations:

(i) Under the setups of relatively high signal-to-noise ratios \( (\beta_{\min} = 1, 0.5) \), the FDR-TPR paths of both LBSS and CoSaMP are nearly in the “Γ” shape, while the other methods have false discoveries on their early solution paths. In particular, when FDR is controlled at 20%, which is a widely used level, LBSS and CoSaMP always have the highest TPR. Similarly to Case 2 of Section [sec:simu]IV-A, SCAD, MCP and L0L2 correct their models at a latter stage by substituting false variables with true ones. The FDR of LASSO keeps increasing as \( \lambda \) decreases.

(ii) Regarding the 10-fold CV solutions, we see that all the methods with visible CV points have a non-negligible number of false variables in order to improve prediction power. The CV points of SCAD and MCP sometimes have lower FDR. LASSO tends to select false variables more than the other approaches.

V. CONCLUSION

In this paper, we establish the sufficient and (near) necessary conditions for BSS to achieve sure selection throughout its early path. We show that the underpinning quantity is the minimum projected signal margin that characterizes the
fundamental gap of fitting power between sure selection models and spurious ones. This margin is robust against collinearity of the design, justifying the low FDP of the early path of LBSS that we observed under highly correlated designs. The current results motivate the following two questions that we wish to answer in our future research:

1) Previous works [23] and [31] on IHT rely on restricted strong convexity and smoothness (or their variants) of the loss function to develop its optimization guarantee. Given that these conditions are dispensable for the statistical properties of BSS, and that IHT is shown to be closely related with BSS, it is natural to ask if these conditions are necessary for IHT to yield desirable statistical properties, say sure early selection. In particular, can we achieve any FDP guarantee for the early path of CoSaMP with only a lower bound of the minimum projected signal margin? Answering these questions should entail more delicate optimization analysis.

2) What are the sufficient and necessary conditions for BSS to achieve sure early selection under linear sparsity, i.e., $s^*/p$ tends to a constant?

3) Under more general contexts (say generalized linear models), what are the counterpart sufficient and necessary conditions for $\ell_0$-constrained methods to achieve sure early selection?

APPENDIX A

PSEUODECOD OF THE IHT ALGORITHMS

For any $v \in \mathbb{R}^p$ and $r \in \mathbb{N}$, define $T_{ab}(v, r) := \{ j : |v_j| \text{ is among the top } r \text{ largest values of } \{|v_k|\}_{k=1}^p \}$. For any $\beta \in \mathbb{R}^p$, let $L(\beta) = \sum_{i=1}^n (y_i - x_i^T \beta)^2$. We present the pseudocode of PGD and CoSaMP in Algorithms 2 and 3.

Algorithm 2 PGD($X, y, \hat{\beta}_0, \pi, \eta, \tau$)

| Input: Design matrix $X$, response $y$, initial value $\hat{\beta}_0$, projection size $\pi$, step size $\eta$, convergence threshold $\tau > 0$. |
|---|
| 1: $t \leftarrow 0$
| 2: repeat
| | 3: $\hat{\beta}_t \leftarrow \hat{\beta}_t - \eta \nabla L(\hat{\beta}_t)$
| | 4: $\hat{G}_t \leftarrow T_{ab}(\hat{\beta}_t, \pi)$
| | 5: $\hat{\beta}_t \leftarrow (X_{\hat{G}_t}^T X_{\hat{G}_t})^{-1} X_{\hat{G}_t}^T y$
| | 6: $t \leftarrow t + 1$
| 7: until $||\hat{\beta}_t - \hat{\beta}_{t-1}||_2 < \tau$
| 8: $\hat{\beta}_0 \leftarrow \hat{\beta}_t$
| Output: $\hat{\beta}_0$ |

Algorithm 3 CoSaMP($X, y, \hat{\beta}_0, \pi, l, \tau$)

| Input: Design matrix $X$, response $y$, initial value $\hat{\beta}_0$, projection size $\pi$, expansion size $l$, convergence threshold $\tau > 0$. |
|---|
| 1: $t \leftarrow 0$
| 2: repeat
| | 3: $\mathcal{G}_t \leftarrow T_{ab}(\nabla L(\hat{\beta}_t), l)$
| | 4: $\mathcal{S}_t \leftarrow \text{supp}(\hat{\beta}_t) \cup \mathcal{G}_t$
| | 5: $\hat{\beta}_t \leftarrow (X_{\mathcal{S}_t}^T X_{\mathcal{S}_t})^{-1} X_{\mathcal{S}_t}^T y$
| | 6: $\mathcal{S}_t \leftarrow T_{ab}(\hat{\beta}_t, \pi)$
| | 7: $\hat{\beta}_{t+1} \leftarrow (X_{\mathcal{S}_t} X_{\mathcal{S}_0})^{-1} X_{\mathcal{S}_t}^T y$
| | 8: $t \leftarrow t + 1$
| 9: until $||\hat{\beta}_t - \hat{\beta}_{t-1}||_2 < \tau$
| 10: $\hat{\beta}_0 \leftarrow \hat{\beta}_t$
| Output: $\hat{\beta}_0$ |

APPENDIX B

PROOFS OF TECHNICAL RESULTS

A. Proof of Theorem 2.1

For any $S \in [p]$, define $\gamma_S := n^{-1/2}(I - PX_S)\mu^*$. For any $S \in \mathcal{A}(S)$, we have

$$n^{-1} L_S = n^{-1}(y^T(I - PX_S)y)$$

Algorithm 3 CoSaMP($X, y, \hat{\beta}_0, \pi, l, \tau$)

| Input: Design matrix $X$, response $y$, initial value $\hat{\beta}_0$, projection size $\pi$, expansion size $l$, convergence threshold $\tau > 0$. |
|---|
| 1: $t \leftarrow 0$
| 2: repeat
| | 3: $\mathcal{G}_t \leftarrow T_{ab}(\nabla L(\hat{\beta}_t), l)$
| | 4: $\mathcal{S}_t \leftarrow \text{supp}(\hat{\beta}_t) \cup \mathcal{G}_t$
| | 5: $\hat{\beta}_t \leftarrow (X_{\mathcal{S}_t}^T X_{\mathcal{S}_t})^{-1} X_{\mathcal{S}_t}^T y$
| | 6: $\mathcal{S}_t \leftarrow T_{ab}(\hat{\beta}_t, \pi)$
| | 7: $\hat{\beta}_{t+1} \leftarrow (X_{\mathcal{S}_t} X_{\mathcal{S}_0})^{-1} X_{\mathcal{S}_t}^T y$
| | 8: $t \leftarrow t + 1$
| 9: until $||\hat{\beta}_t - \hat{\beta}_{t-1}||_2 < \tau$
| 10: $\hat{\beta}_0 \leftarrow \hat{\beta}_t$
| Output: $\hat{\beta}_0$ |

We wish to show that the following two bounds hold with high probability:

$$\inf_{S \in \mathcal{A}(S)} \left\{ 2^{-1} (1 - \eta) \left( ||\gamma_S||_2^2 - ||\gamma_{\Phi(S)}||_2^2 \right) + 2(\gamma_S - \gamma_{\Phi(S)})^T e \right\} > 0$$

(22)

and

$$\inf_{S \in \mathcal{A}(S)} \left\{ 2^{-1} (1 - \eta) \left( ||\gamma_S||_2^2 - ||\gamma_{\Phi(S)}||_2^2 \right) + \frac{1}{n} e^T (P X_S - P X_{\Phi(S)}) e \right\} > 0.$$
We first consider (22). Given that $\|e\|_{\psi_2} \leq \sigma$, applying Hoeffding’s inequality yields that for any $x > 0$,
\[
\mathbb{P}\left\{ (\gamma_S - \gamma_{\Phi(S)})^\top e \geq x\sigma \|\gamma_S - \gamma_{\Phi(S)}\|_2 \right\} \leq 2e^{-x^2/2}.
\]
(24)

Note that $A_t(s) := \{S \in \mathcal{A}(s) : |S\setminus\Phi(S)| = t\}$ for $t \in [s]$. Then $A(s) = \cup_{t \in \mathbb{N}} A_t(s)$. A union bound over $S \in A_t(s)$ yields that
\[
\mathbb{P}\left\{ \exists S \in A_t(s) \text{ s.t. } \frac{2(\gamma_S - \gamma_{\Phi(S)})^\top e}{n^{1/2}} \geq 2x\sigma \|\gamma_S - \gamma_{\Phi(S)}\|_2 \right\} \leq 2|A_t(s)|e^{-x^2/2}.
\]
(25)

Writing $\xi_0 = \frac{(1-\eta)m_\ast s}{\lambda(t\log p)^{1/2}}$ and substituting $x = 2\xi_0(t\log p)^{1/2}$ into (25), we have that
\[
\mathbb{P}\left\{ \exists S \in A_t(s) \text{ s.t. } \frac{2(\gamma_S - \gamma_{\Phi(S)})^\top e}{n^{1/2}} \geq 4\xi_0\sigma \|\gamma_S - \gamma_{\Phi(S)}\|_2 \left(\frac{\log p}{n}\right)^{1/2} \right\} \leq 2|A_t(s)|e^{-2\xi_0^2t\log p}.
\]
(26)

Note that
\[
\gamma_S - \gamma_{\Phi(S)} = n^{-1/2}(P_{\Phi(S)|S} - P_{\Phi(S)})\mu^* - n^{-1/2}(P_{\Phi(S)|S} - P_{S|\Phi(S)})\mu^*,
\]
which implies that
\[
\|\gamma_S - \gamma_{\Phi(S)}\|_2 \leq n^{-1/2}(\|P_{\Phi(S)|S}\mu^*\|_2 + \|P_{S|\Phi(S)}\mu^*\|_2).
\]
(27)

Consequently,
\[
\|\gamma_S\|_2^2 - \|\gamma_{\Phi(S)}\|_2^2 = n^{-1/2}\|P_{\Phi(S)|S} - P_{\Phi(S)}\mu^*\|_2 = n^{-1/2}(\|P_{\Phi(S)|S}\mu^*\|_2 + \|P_{S|\Phi(S)}\mu^*\|_2).
\]
\[
\geq \|\gamma_S - \gamma_{\Phi(S)}\|_2^2.
\]
(28)

Combining (28) with the definition of $\xi_0$ and $m_\ast(s)$ yields that
\[
\frac{1-\eta}{2}(\|\gamma_S\|_2^2 - \|\gamma_{\Phi(S)}\|_2^2) \geq 4\xi_0\sigma \|\gamma_S - \gamma_{\Phi(S)}\|_2 \left(\frac{\log p}{n}\right)^{1/2}.
\]

Therefore, we deduce from (26) that
\[
\mathbb{P}\left\{ \exists S \in A_t(s) \text{ s.t. } \frac{2(\gamma_S - \gamma_{\Phi(S)})^\top e}{n^{1/2}} \geq \frac{1-\eta}{2}(\|\gamma_S\|_2^2 - \|\gamma_{\Phi(S)}\|_2^2) \right\} \leq 2|A_t(s)|e^{-2\xi_0^2t\log p}.
\]

Note that
\[
|A_t(s)| = \binom{s}{t} \binom{p-t}{s-t} = \binom{p}{s} \binom{p-s}{t} = \binom{p}{s} \binom{p-s}{s-t}.
\]

By Stirling’s formula and the fact that $\log p \geq s^*$, we have
\[
\log \left( \binom{s}{s-t} \right) \leq \log \left( \binom{s^*}{s^*/2} \right) \leq s^* \log p.
\]

Hence, we have
\[
\mathbb{P}\left\{ \exists S \in A_t(s) \text{ s.t. } \frac{2(\gamma_S - \gamma_{\Phi(S)})^\top e}{n^{1/2}} \geq \frac{1-\eta}{2}(\|\gamma_S\|_2^2 - \|\gamma_{\Phi(S)}\|_2^2) \right\} \leq 2C_1p^{-2(\xi_0^2-1)}.
\]
(29)

Next we aim to show that (23) is a high-probability event. Fix any $t \in [s]$. For any $S \in \mathcal{A}_t(s)$, let $U, V$ be the orthogonal complement of $W := \text{col}(X_{S\setminus\Phi(S)})$ as a subspace of $\text{col}(X_{S})$ and $\text{col}(X_{\Phi(S)})$ respectively. Then $\dim(U) = \dim(V) = t$. We have
\[
\frac{1}{n} e^\top (P_{S\setminus\Phi(S)} - P_{X_{S\setminus\Phi(S)}}) e = \frac{1}{n} e^\top (P_U + P_{U^\perp}) e - \frac{1}{n} e^\top (P_{U^\perp} + P_V) e = \frac{1}{n} e^\top (P_U - P_V) e.
\]

By Theorem 1.1 in [32], there exists a universal constant $c > 0$ such that for any $x > 0$,
\[
\mathbb{P} \left( |e^\top P_U e - E(e^\top P_U e)| > \sigma^2 x \right) \leq 2e^{-c\min(x^2/\|P_U\|_2^2, x/\|P_U\|_2)} = 2e^{-c\min(x^2/t, x)}.
\]

Similarly,
\[
\mathbb{P} \left( |e^\top P_V e - E(e^\top P_V e)| > \sigma^2 x \right) \leq 2e^{-c\min(x^2/t, x)}.
\]

Note that
\[
E(e^\top P_V e) = \text{Etr}(P_V e e^\top) = \text{var}(\epsilon_1) \text{tr}(P_V) = t \text{ var}(\epsilon_1) = E(e^\top P_U e).
\]

Combining the above two inequalities yields
\[
\mathbb{P} (|e^\top P_U e - e^\top P_V e| > 2\sigma^2 x) \leq 4e^{-c\min(x^2/t, x)}.
\]
(30)

If we have $\xi_0 \geq (16\log 2)^{-1/2}$, then applying a union bound over $S \in A_t(s)$ and taking $x = 16\xi_0^2t\log p$ yields that
\[
\mathbb{P}\left\{ \exists S \in A_t(s) \text{ s.t. } \frac{1}{n} e^\top P_{X_{S\setminus\Phi(S)}} e - \frac{1}{n} e^\top P_{X_{S\setminus\Phi(S)}} e \right\} \leq 4e^{-c\min(x^2/t, x)}.
\]
Recall that \( \bar{\gamma} \) and (32) yields that
\[
 \| \bar{\gamma} \|_2^2 - \| \gamma(\phi(s)) \|_2^2 \\
\geq \frac{1 - \eta}{2n} \left( (\| \gamma_S \|_2^2 - \| \gamma(\phi(s)) \|_2^2) + (\| \gamma_S - \gamma(\phi(s)) \|_2^2) \right)
\]
\[
\geq \frac{1 - \eta}{2n} \left( (\| \gamma_S \|_2^2 - \| \gamma(\phi(s)) \|_2^2) \right)
\]
\[
\geq \frac{1 - \eta}{2n} \left( (\| \gamma_S \|_2^2 - \| \gamma(\phi(s)) \|_2^2) \right)
\]
\[
\geq 32 \xi_0^2 \sigma^2 t \left( \frac{\log p}{n} \right).
\]

A further union bound over \( t \in [s] \) yields
\[
\Pr \left\{ \exists S \in \mathcal{A}(s) \text{ s.t.} \left| \frac{m^T P_X S - e^T P_X \phi(s) e}{n} \right| \\
> \frac{1 - \eta}{2n} \left( (\| \gamma_S \|_2^2 - \| \gamma(\phi(s)) \|_2^2) \right) \right. \}
\]
\[
\leq 4C_1 \exp^{-2(8\xi_0^2)}.
\]

Finally, let \( C = \max \{1, (8c)^{-1/2}, 6C_1\} \). Combining (21), (29) and (32) yields that
\[
\Pr \left\{ \inf_{S \in \mathcal{A}(s)} \left( \frac{1}{n} \mathcal{L}_S - \mathcal{L}_{\phi(s)} - \eta(\| \gamma_S \|_2^2 - \| \gamma(\phi(s)) \|_2^2) \right) \geq 0 \right. \}
\]
\[
\geq 1 - Csp^{-2(2c_0^2)}.
\]

Note that
\[
\eta(\| \gamma_S \|_2^2 - \| \gamma(\phi(s)) \|_2^2) \geq \frac{\eta m^2(s)}{n} \quad \text{and}
\]
\[
\mathcal{L}_S - \mathcal{L}_C \geq \mathcal{L}_{S} - \mathcal{L}_{\phi(s)} \quad \text{for all} \quad S \in \mathcal{A}(s).
\]

for any \( \xi > C \), the conclusion then follows immediately if \( \xi_0 \geq \xi \) and setting \( Csp^{-2(2c_0^2)} = \delta \).

B. Proof of Theorem 2.2

We wish to show that if (13) in the article is satisfied, then there exists \( j \in \mathcal{J}_0 \) such that \( \mathcal{L}_{(j)} < \mathcal{L}_{(k)} := \min_{j^* \in S^*} \mathcal{L}_{(j^*)} \) with high probability. To see this, for any \( j^* \in S^* \) and any \( j \in \mathcal{J}_0 \), we have that
\[
n^{-1}(\mathcal{L}_{(j)} - \mathcal{L}_{(j^*)}) = \| \gamma(j) \|_2^2 - \| \gamma(j^*) \|_2^2 + \frac{2(\gamma(j) - \gamma(j^*)^T \epsilon}{n^{1/2}}
\]
\[
- \frac{1}{n} e^T (P_Xj - P_Xj^*) e
\]
\[
\leq \| \gamma(j) \|_2^2 - \| \gamma(j^*) \|_2^2 + \frac{2(\gamma(j) - \gamma(j^*)^T \epsilon}{n^{1/2}}
\]
\[
- \frac{1}{n} e^T (P_Xj - P_Xj^*) e.
\]

Recall that \( \bar{u}_j := X_j / \| X_j \|_2, \forall j \in [p] \). For convenience, write
\[
\Delta = \max_{j \in \mathcal{J}_0} \{ \| P_Xj \|_2^2 - \| P_X \mu^* \|_2^2 \}
\]
\[
= \max_{j \in \mathcal{J}_0} \{ \| \bar{u}_j \|_2^2 \}
\]

We then have that
\[
\gamma_{(j)} - \gamma_{(j^*)} = n^{-1/2}(P_Xj - P_Xj^*) \mu^*
\]
\[
= n^{-1/2}(\bar{u}_j \bar{u}_j^T - \bar{u}_j \bar{u}_j^T) \mu^*,
\]

and that
\[
\| \gamma_{(j)} \|_2^2 - \| \gamma_{(j^*)} \|_2^2
\]
\[
= \frac{1}{n} \mu^T (P_Xj - P_Xj^*) \mu^*
\]
\[
= \frac{1}{n} \mu^T (\bar{u}_j \bar{u}_j^T - \bar{u}_j \bar{u}_j^T) \mu^*
\]
\[
= \frac{1}{n} \left( \| \bar{u}_j \mu^* \|_2^2 - \| \bar{u}_j \mu^* \|_2^2 \right)
\]
\[
\leq \frac{2}{n} \| \bar{u}_j \mu^* \|_2^2.
\]

Note that Assumption 2.2 implies that \( |\bar{u}_j \mu^*| \geq 4\Delta \); we thus deduce from (35) that
\[
\| \bar{u}_j \mu^* \|_2^2 \geq \frac{4(\bar{u}_j \mu^* \mu^*)}{2}, \forall j \in \mathcal{J}.
\]

By Lemma 6.1 of [19], (36) and then (13 in the article), we have for any \( j, k \in \mathcal{J}_0 \) and \( j \neq k \) that
\[
\| \gamma_{(j)} - \gamma_{(k)} \|_2 \leq n^{-1/2} \| \bar{u}_j \bar{u}_j^T \mu^* - \bar{u}_k \bar{u}_k^T \mu^* \|_2
\]
\[
\leq \frac{1}{n} \min_{j, k \in \mathcal{J}_0} \| \bar{u}_j \|_2 \mu^*, \| \bar{u}_k \|_2 \mu^* \|_2
\]
\[
\leq \delta_0(2n^{-1/2} \| \bar{u}_j \mu^* \|_2^2 + \| \bar{u}_k \mu^* \|_2^2)
\]
\[
\leq \delta_0^2 2n^{-1/2} \| \bar{u}_j \mu^* \|_2^2.
\]

Therefore, \( \{ \gamma_{(j)} - \gamma_{(j^*)} \} \in \mathcal{J}_0 \) is a \( \delta_0 \)-packing set of itself. By Sudakov’s lower bound [42, Theorem 5.30], we deduce that
\[
\| \gamma_{(j)} - \gamma_{(j^*)} \|_2 \leq \| \gamma_{(j)} - \gamma_{(j^*)} \|_2
\]
\[
\leq \frac{1}{n} \left( \| \bar{u}_j \mu^* \|_2^2 + \| \bar{u}_j \mu^* \|_2^2 \right)
\]
\[
\leq \frac{2}{n} \| \bar{u}_j \mu^* \|_2^2.
\]

Furthermore, (34) and (13) in the article imply that
\[
\| \gamma_{(j)} - \gamma_{(j^*)} \|_2 \leq n^{-1/2}(\| \bar{u}_j \mu^* \|_2^2 + \| \bar{u}_j \mu^* \|_2^2)
\]
\[
\leq 2n^{-1/2} \| \bar{u}_j \mu^* \|_2^2.
\]

Therefore, by Lemma 4,
\[
\| \gamma_{(j)} - \gamma_{(j^*)} \|_2 \leq \frac{\sigma^2}{n^{1/2}} \max_{j \in \mathcal{J}_0} \| \gamma_{(j)} - \gamma_{(j^*)} \|_2^2
\]
\[
\leq \frac{\sigma^2}{n^{1/2}} |\bar{u}_j \mu^*|^2
\]
\[
\leq \frac{\sigma^2}{n^{1/2}} \| \bar{u}_j \mu^* \|_2^2.
\]
Choosing \( t = 2^{-3/2} C^{-1} \delta_0 c_0^{1/2} \log^{1/2} p \), we then reduce the bound above to

\[
\Pr \left\{ \min_{j \in \mathcal{J}_0} \frac{2(\gamma_j - \gamma_{j^*})^\top \epsilon}{\sigma_j \| \hat{\mu}_j^\top \mu^* \| \log^{1/2} p} \geq \frac{\delta_0 c_0^{1/2} \sigma_j \| \hat{\mu}_j^\top \mu^* \| \log^{1/2} p}{2^{3/2}} \right\} \leq p^{-\delta_0 c_0 \delta_0 \sigma / (8c_0^2)}. \tag{39}
\]

Besides, Assumption 2.2 yields that

\[ \| \hat{\mu}_j^\top \mu^* \| > 8 \sigma / \{ \delta_0 c_0 \log(p) \}^{1/2}. \]

Therefore, applying a union bound over \( j \in \mathcal{J}_0 \) to (30) with \( x = (8 \sigma)^{-1} \delta_0 c_0^{1/2} \| \hat{\mu}_j^\top \mu^* \| \log^{1/2} p \), we deduce by Assumption 2.2 that

\[
\Pr \left\{ \max_{j \in \mathcal{J}_0} \frac{\| P_{X_j} \epsilon - \epsilon \| P_{X_j} \epsilon}{\sigma_j \| \hat{\mu}_j^\top \mu^* \| \log^{1/2} p} \right\} \leq 4p^{-(\delta_0 c_0 \delta_0 \sigma / (8c_0^2))}. \tag{40}
\]

Finally, note that

\[
\min_{j \in \mathcal{J}_0} n^{-1}(\mathcal{L}_{(j)} - \mathcal{L}_{(j^*)}) \leq \max_{j \in \mathcal{J}_0} \left( \| \gamma_j \|_2^2 - \| \gamma_{j^*} \|_2^2 \right) + \min_{j \in \mathcal{J}_0} \frac{2(\gamma_j - \gamma_{j^*})^\top \epsilon}{\sigma_j} \| \hat{\mu}_j^\top \mu^* \| \log^{1/2} p \]  
+ \max_{j \in \mathcal{J}_0} \frac{\| P_{X_j} \epsilon - \hat{\mu}_j^\top \mu^* \|}{\sigma_j \| \hat{\mu}_j^\top \mu^* \| \log^{1/2} p} > 0 \tag{41}
\]

Combining (41), (39), (40) and (35) yields that when \( \Delta < \delta_0 \sigma (c_0 \log p)^{1/2} / 20 \),

\[
\Pr \left\{ \min_{j \in \mathcal{J}_0} \mathcal{L}_{(j)} < \mathcal{L}_{(j^*)} \right\} \geq 1 - 4p^{-(\delta_0 c_0 \delta_0 \sigma / (8c_0^2))}. \tag{42}
\]

The conclusion immediately follows once we apply a union bound over \( j^* \in S^\dagger \).

**C. Proof of Theorem 2.3**

An important observation is that \( \Phi(S) = S^\dagger \) and

\[ m(S) = \| P_{S^\dagger} \mu^* \|_2 - \| P_{S \setminus S^\dagger} \mu^* \|_2 \]

for any \( S \in \mathcal{A}_0 \). This motivates us to take the following two main steps to establish the theorem: (i) we show that \( \mathcal{L}_{S^\dagger} \) is the smallest among \( \{ \mathcal{L}_S \}_{S \in \mathcal{A}^*(s)} \) with high probability; (ii) we show that \( \min_{S \in \mathcal{A}_0} \mathcal{L}_S < \mathcal{L}_{S^\dagger} \), which implies that \( S^\dagger \) is not the best subset any more, and thus that the best subset must have false discoveries.

**Step (i).** We aim to show that

\[
\Pr \left\{ \min_{S \in \mathcal{A}^*(s)} \mathcal{L}_S - \mathcal{L}_{S^\dagger} \leq 0 \right\} \leq 4p^{-(\delta_0 c_0^{1/2} / 8c_0^2 - 2)} + 2p^{-(\delta_0 c_0^{1/2} / 32c_0^2 - 2)}. \tag{43}
\]

This step follows closely the proof strategy of Theorem 2.1. For any \( S \in \mathcal{A}^*(s) \),

\[
n^{-1}(\mathcal{L}_S - \mathcal{L}_{S^\dagger}) = \| \gamma_S \|_2^2 - \| \gamma_{S^\dagger} \|_2^2 + \frac{2(\gamma_S - \gamma_{S^\dagger})^\top \epsilon}{n^{1/2}} \]  
- \frac{1}{n} \epsilon^\top (P_{X_S} - P_{X_{S^\dagger}}) \epsilon \]  
+ \frac{\| \gamma_S \|_2^2 - \| \gamma_{S^\dagger} \|_2^2}{2} - \frac{1}{n} \epsilon^\top (P_{X_S} - P_{X_{S^\dagger}}) \epsilon. \tag{44}
\]

We wish to show that

\[
\inf_{S \in \mathcal{A}^*(s)} \left\{ \| \gamma_S \|_2^2 - \| \gamma_{S^\dagger} \|_2^2 \right\} > 0 \tag{45}
\]

and

\[
\inf_{S \in \mathcal{A}^*(s)} \left\{ \| \gamma_S \|_2^2 - \| \gamma_{S^\dagger} \|_2^2 \right\} > 0 \tag{46}
\]

with high probability in the sequel. Define \( \mathcal{A}^*_s \) : \( S \in \mathcal{A}^*(s) : |S \setminus S^\dagger| = t \). To prove (45), first fix some \( S \in \mathcal{A}^*_s \). By Hoeffding’s inequality, we have for any \( x > 0 \) that

\[
\Pr \left\{ (\gamma_S - \gamma_{S^\dagger})^\top \epsilon > x \sigma \| \gamma_S - \gamma_{S^\dagger} \|_2 \right\} \leq 2e^{-x^2 / 2}.
\]

A union bound over \( S \in \mathcal{A}^*_s \) yields that

\[
\Pr \left\{ \exists S \in \mathcal{A}^*_s \text{ s.t.} \frac{2(\gamma_S - \gamma_{S^\dagger})^\top \epsilon}{n^{1/2}} \geq \frac{2x \sigma \| \gamma_S - \gamma_{S^\dagger} \|_2}{n^{1/2}} \right\} \leq 2|\mathcal{A}^*_s|e^{-x^2 / 2} = 2 \left( \begin{array}{c} s \cr t \end{array} \right) \left( \frac{s - s}{t} \right) \leq 2p e^{-x^2 / 2 - 2t},
\]

Let \( x = \xi (t \log p)^{1/2} / 4 \). Then we have that

\[
\Pr \left\{ \exists S \in \mathcal{A}^*_s \text{ s.t.} \frac{2(\gamma_S - \gamma_{S^\dagger})^\top \epsilon}{n^{1/2}} \geq \frac{\xi \sigma \| \gamma_S - \gamma_{S^\dagger} \|_2}{2} \left( \frac{t \log p}{n} \right)^{1/2} \right\} \leq 2p^{-t (\xi^2 / 32 - 2)},
\]

Note that

\[
\| \gamma_S \|_2^2 - \| \gamma_{S^\dagger} \|_2^2 = n^{-1} \mu^\top (P_{X_{S^\dagger}} - P_{X_S}) \mu^*.
\]
\[
\begin{align*}
\gamma_s - \gamma_{S'} &\geq \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2 \\
&\geq \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2 \\
&\geq \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2.
\end{align*}
\] (47)

Further apply a union bound for \( t \in [s] \). We thus have that
\[
\Pr \left\{ \exists S \in \mathbb{A}^*(s) \text{ s.t. } \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2 \leq 2sp^{-\left(\epsilon^2/32 - 2\right)} \right\}.
\] (48)

Next we show that (46) holds with high probability. Similarly to (30), we can obtain that
\[
\Pr \left\{ \frac{2}{n} \|\gamma_s - \gamma_{S'}\|_2 > 2\sigma^2 x \right\} \leq 4e^{-c\min\left(s^2/t, x\right)}.
\] (49)

Note that \( \log p > 1 \) and \( \xi > 2 \). By taking \( x = t\xi^2 \log p/4 \), applying a union bound over \( S \in \mathbb{A}^*(s) \) yields that
\[
\Pr \left\{ \exists S \in \mathbb{A}^*(s) \text{ s.t. } \frac{2}{n} \|\gamma_s - \gamma_{S'}\|_2 > 2\sigma^2 x \right\} \leq 4p^{-t(\epsilon^2/4 - 2)}.
\] (50)

Note that
\[
\frac{2}{n} \|\gamma_s - \gamma_{S'}\|_2 = \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2
\]
\[
= \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2
\]
\[
= \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2
\]
\[
\geq \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2
\]
\[
\geq \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2
\]
\[
\geq \frac{2}{n} \left( \|\gamma_s - \gamma_{S'}\|_2 \right)^2.
\] (51)

Combining (48) and (51) yields (43).

**Step (ii).** We wish to show that if (14) in the article is satisfied, then \( \min_{S \in \mathbb{A}_{\delta_0}} \mathcal{L}_S < \mathcal{L}_S' \). Note that for any \( S \in \mathbb{A}_{\delta_0} \),
\[
\begin{align*}
\gamma_S - \gamma_{S'} &= n^{-1/2} (\mathbf{P}_{X_S} - \mathbf{P}_{X_{S'}}) \mathbf{e} \\
&= n^{-1/2} (\mathbf{u}_j^\top \mathbf{u}_j - \mathbf{u}_j^\top \mathbf{u}_j) \mathbf{e},
\end{align*}
\] (53)

and that
\[
\begin{align*}
\|\gamma_S\|_2^2 - \|\gamma_{S'}\|_2^2 \\
&= \frac{1}{n} \mathbf{e}^\top (\mathbf{P}_{X_S} - \mathbf{P}_{X_{S'}}) \mathbf{e}
\end{align*}
\] (52)

\[
\begin{align*}
\|\gamma_S\|_2^2 - \|\gamma_{S'}\|_2^2 \\
&= \frac{1}{n} \mathbf{e}^\top (\mathbf{P}_{X_S} - \mathbf{P}_{X_{S'}}) \mathbf{e}
\end{align*}
\] (55)

By Lemma 6.1 of [19] and then (55), we have for any \( j, k \in \mathcal{J}_{\delta_0} \) and \( j \neq k \) that
\[
\begin{align*}
\|\gamma_S\|_2^2 - \|\gamma_{S'}\|_2^2 \\
&= n^{-1/2} (\mathbf{u}_j^\top \mathbf{u}_j - \mathbf{u}_j^\top \mathbf{u}_j) \mathbf{e} \\
&\geq n^{-1/2} \min \{ |\mathbf{u}_j^\top \mathbf{u}_j|, |\mathbf{u}_j^\top \mathbf{u}_j| \} \delta_0 \\
&\geq \delta_0 (2n)^{-1/2} |\mathbf{u}_j^\top \mathbf{u}_j| = \delta_0.
\end{align*}
\] (56)

Therefore, \( \{ \gamma_S - \gamma_{S'} \}_{S \in \mathbb{A}_{\delta_0}} \) is a \( \delta_0 \)-packing set of itself. By Sudakov’s lower bound [42, Theorem 5.30], we deduce that
\[
\begin{align*}
\mathbb{E} \left\{ \min_{S \in \mathbb{A}_{\delta_0}} \frac{2(\gamma_S - \gamma_{S'})^\top \mathbf{e}}{n^{1/2}} \right\} \\
&\leq -\delta_0 \sigma \left( \frac{c_0 \log p}{n} \right)^{1/2}.
\end{align*}
\] (57)

Combining (56) and (57), we deduce that there exists a universal constant \( C > 0 \), such that for any \( t > 0 \),
\[
\begin{align*}
\mathbb{P} \left\{ \min_{S \in \mathbb{A}_{\delta_0}} \frac{2(\gamma_S - \gamma_{S'})^\top \mathbf{e}}{n^{1/2}} \\
&\geq -\delta_0 \sigma \left( \frac{c_0 \log p}{n} \right)^{1/2} + \left( C \sigma |\mathbf{u}_j^\top \mathbf{u}_j| \right) \right\} \\
&\leq \exp(-t^2).
\end{align*}
\] Choosing \( t = 2^{-3/2} C^{-1} \delta_0^{1/2} \log^{1/2} p \), we then reduce the bound above to
\[
\begin{align*}
\mathbb{P} \left\{ \min_{S \in \mathbb{A}_{\delta_0}} \frac{2(\gamma_S - \gamma_{S'})^\top \mathbf{e}}{n^{1/2}} \\
&\geq -\delta_0 \sigma \left( \frac{c_0 \log p}{n} \right)^{1/2} \right\}
\end{align*}
\] (58)
E. Lemma 1 and Its Proof

details for less redundancy.

we combine the bound above with (58), (59) and (54).

Besides, Assumption 2.4 yields that $|\hat{u}_j^\top \mu^*| > 8\sigma_0 \{\delta_0 \sigma_0^{1/2} (\log p)^{1/2}\}$. Therefore, applying a union bound over $S \in \mathbb{A}_0$ to (30) with $x = (8\sigma_0)^{-1} \delta_0 \sigma_0^{1/2} |\hat{u}_j^\top \mu^*| \log^{1/2} p$, we obtain that

$$
P \left\{ \max_{S \in \mathbb{A}_0} \frac{|e^\top P x_s - e^\top P x_{S^1}|}{n} > \frac{\delta_0 \sigma_0^{1/2} \sigma |\hat{u}_j^\top \mu^*| \log^{1/2} p}{4n^2} \right\} \leq 4p^{-(\xi\sigma_0^2/8-1)c_0}. 
$$

Finally, note that

$$
\min_{S \in \mathbb{A}_0} n^{-1}(L_S - L_{S^1}) \\
\leq \max_{S \in \mathbb{A}_0} \frac{|\gamma_{S^1}^2 - |\gamma_{S^1}||^2|}{n^{1/2}} \\
+ \min_{S \in \mathbb{A}_0} \frac{2(\gamma_{S^1} - |\gamma_{S^1}|)}{n^{1/2}} \\
+ \min_{S \in \mathbb{A}_0} \frac{1}{n} e^\top (P x_s - P x_{S^1}) e.
$$

When $\Delta < \delta_0 \sigma_0 (c_0 \log p)^{1/2}/20$, we reach the conclusion once we combine the bound above with (58), (59) and (54).

$$
P \left\{ \min_{S \in \mathbb{A}_0} L_S \leq L_{S^1} \right\} \\
\geq 1 - 4p^{-(\xi\sigma_0^2/8-1)c_0} - \delta_0^2 c_0/(8c_0^2).
$$

D. Proof of Corollary 1

The proof of Corollary 1 is analogous to that of Theorem 2.1 by simply replacing $\mathbb{A}_0(s)$ with $\mathbb{A}_q(s)$. We omit the details for less redundancy.

E. Lemma 1 and Its Proof

To start with, for any two size-$s$ sets $S \in \mathbb{A}_0(s)$ and $S^1 \in \mathbb{A}_0^*(s)$ such that $S \cap S^* \subset S^1$, we analyze two crucial components of the projection signal margin, $\mu^\top P_{S \setminus S^*} \mu^*$ and $\mu^\top P_{S \setminus S^1} \mu^*$, under Gaussian design. Suppose $\{x_i\}_{i \in [n]}$ are independent and identically distributed observations of $x \sim N(0, \Sigma)$. For notational simplicity, let $S_1 = S \cap S^1, S_2 = S^1 \setminus S_1$ and $S_3 = S \setminus S_1$. Figure 6 shows the relationship between the sets.

Next we introduce some quantities that are involved in the concentration bounds we establish. Define

$$
h := \text{cov}(x_{S_2}, \mu^*|x_{S_1}) \\
= (\Sigma_{S_2 \setminus S^*} - \Sigma_{S_2 \setminus S_1} \Sigma_{S_1 \setminus S^*}^{-1} \Sigma_{S_1 \setminus S^*}) \beta^* 
$$

and

$$
H := \text{cov}(x_{S_2}, x_{S_2}|x_{S_1}) \\
= \Sigma_{S_2 \setminus S_2} - \Sigma_{S_2 \setminus S_1} \Sigma_{S_1 \setminus S_2}^{-1} \Sigma_{S_1 \setminus S_2}.
$$

Then write

$$
\nu_1 := h^\top H^{-1} h \quad \text{and} \quad \nu_2 := \text{var}(\mu^*|x_{S_1}) - \nu_1 \\
= \beta^* (\Sigma_{S \setminus S^*} - \Sigma_{S \setminus S_1} \Sigma_{S_1 \setminus S}^{-1} \Sigma_{S_1 \setminus S^*}) \beta^* - \nu_1.
$$

Similarly, define

$$
h' := \text{cov}(x_{S_3}, \mu^*|x_{S_1}) \\
= (\Sigma_{S_3 \setminus S^*} - \Sigma_{S_3 \setminus S_1} \Sigma_{S_1 \setminus S_3}^{-1} \Sigma_{S_1 \setminus S^*}) \beta^* 
$$

and

$$
H' := \text{cov}(x_{S_3}, x_{S_3}|x_{S_1}) \\
= \Sigma_{S_3 \setminus S_3} - \Sigma_{S_3 \setminus S_1} \Sigma_{S_1 \setminus S_3}^{-1} \Sigma_{S_1 \setminus S_3}.
$$

Then write

$$
\nu_1' := h'^\top H'^{-1} h' \quad \text{and} \quad \nu_2' := \text{var}(\mu^*|x_{S_1}) - \nu_1 \\
= \beta^* (\Sigma_{S \setminus S^*} - \Sigma_{S \setminus S_1} \Sigma_{S_1 \setminus S}^{-1} \Sigma_{S_1 \setminus S^*}) \beta^* - \nu_1.'
$$

Now we are in position to present the concentration bounds for $\mu^\top P_{S \setminus S^*} \mu^*$ and $\mu^\top P_{S \setminus S^1} \mu^*$.

Lemma 1: For any two size-$s$ sets $S \in \mathbb{A}(s)$ and $S^1 \in \mathbb{A}_0^*(s)$ such that $S \cap S^* \subset S^1$ and any $\xi > 0$, we have

$$
P \left\{ \left| \mu^\top P_{S \setminus S^*} \mu^* - \left\{ (n-s+t) \nu_1 + t \nu_2 \right\} \right| > \xi \left\{ 3\nu_1 (n-s-t)^{1/2} \\
+ 6(\nu_1 \nu_2)^{1/2} (n-s+t)^{1/2} + 3 \nu_2 t^{1/2} \right\} \right\} \\
\leq 6e^{-c \min(\xi, \xi^2)} \\
\text{and} \\
P \left\{ \left| \mu^\top P_{S \setminus S^1} \mu^* - \left\{ (n-s+t) \nu_1' + t \nu_2' \right\} \right| > \xi \left\{ 3\nu_1' (n-s-t)^{1/2} + 6(\nu_1' \nu_2')^{1/2} (n-s-t)^{1/2} \\
+ 3 \nu_2' t^{1/2} \right\} \right\} \leq 6e^{-c \min(\xi, \xi^2)},
$$

where $t = |S \setminus S^1|$, and where $c$ is a universal constant.

Proof of Lemma 1:
Recall that $\mathbb{A}(s) := \{S \in \mathbb{A}(s) : |S \setminus S^*| = t\}$ for $t \in [s]$. For any $S \in \mathbb{A}_0(s)$, we start with analyzing $\|P_{S \setminus S^*} \mu^*\|_2^2$. Note that $|S_2| = t$ and $|S_1| = s - t$. For each $j \notin S$, regressing $X_j$ on $X_{S_1}$ yields that

$$
X_j = X_{S_1} \Sigma_{S_1 \setminus S_1}^{-1} X_{S_1} + (X_j - X_{S_1} \Sigma_{S_1 \setminus S_1}^{-1} X_{S_1}) =: X_{S_1} \theta_j + \Gamma_j.
$$
Consider the conditional distribution of $\|P_{S^1 \mid S} \mu^*\|^2_2$ given $X_{S_1}$. Let

$$I - P_{X_{S_1}} = VV^T = \sum_{j=1}^{n-(s-t)} v_j v_j^T$$

be an eigen-decomposition of $I - P_{X_{S_1}}$. Note that $V$ is independent of $\Gamma_{S_1}^*$, because $X_{S_1}$ is independent of $\Gamma_{S_1}$. Then we have

$$\mu^*^T P_{S^1 \mid S} \mu^* = \mu^*^T VV^T X_{S_1}(X_{S_1}^T VV^T X_{S_1})^{-1} X_{S_1}^T VV^T \mu^*$$

$$= \mu^*^T VV^T \Gamma_{S_2}^*(\Gamma_{S_2}^T VV^T \Gamma_{S_2})^{-1} \Gamma_{S_2}^T VV^T \mu^*$$

$$= \mu^*^T \Gamma_{S_2}^*(\Gamma_{S_2}^T \Gamma_{S_2})^{-1} \Gamma_{S_2}^T \mu^*,$$

where $\Gamma_{S_1}^* := V^T \Gamma_{S_1}^*$ and $\mu^* = V^T \mu^*$. Besides, conditional on $X_{S_1}$, $\Gamma_{S_1}^*$ and $\mu^* \neq 0$ have independent rows because of Gaussianity of $\Gamma_{S_1}^*$ and $\mu^*$, and orthogonality of $V$. Applying Lemma 2, we obtain that for any $\xi > 0$,

$$p\left[\mu^*^T P_{S^1 \mid S} \mu^* - \{n-s+t\nu_1 + \nu_2\}\right]$$

$$> \xi \left\{ 3\nu_1 (n-s)^{1/2} + 6(\nu_1 \nu_2)^{1/2} (n-s+t)^{1/2} + 3\nu_2 t^{1/2} \right\} \left| X_{S_1} \right| \leq 6e^{-c \min(\xi^2, \xi^4)}. \quad (62)$$

Taking expectation with respect to $X_{S_1}$ on both sides of $(62)$, we deduce that

$$P\left[\mu^*^T P_{S^1 \mid S} \mu^* - \{n-s+t\nu_1 + \nu_2\}\right]$$

$$> \xi \left\{ 3\nu_1 (n-s)^{1/2} + 6(\nu_1 \nu_2)^{1/2} (n-s+t)^{1/2} + 3\nu_2 t^{1/2} \right\} \leq 6e^{-c \min(\xi^2, \xi^4)}. \quad (63)$$

For $\|P_{S^1 \mid S} \mu^*\|^2_2$, we can reach the conclusion by simply replacing $S_2$ with $S_3$. \qed

**F. Proof of Theorem 2.4**

For any $S \in \mathcal{A}_t(s)$, we first consider any $S^1 \in \mathcal{A}^*_t(s)$ such that $S \cap S^* \subset S^1$. Note that $S = \mathcal{I}_p$, simple algebra yields that $\nu_1 = t \beta^2, \nu_2 = (s^* - s) \beta^2, \nu_1' = 0$ and $\nu_2' = (s^* - s + t) \beta^2$. Applying Lemma 1 yields that for any $\xi > 0$, we have

$$P\left[\mu^*^T P_{S^1 \mid S} \mu^* - \{n-s+t\beta^2 + t(s^* - s) \beta^2\}\right]$$

$$> \xi \left\{ 3\nu_1 (n-s)^{1/2} + 6(\nu_1 \nu_2)^{1/2} (n-s+t)^{1/2} + 3\nu_2 t^{1/2} \right\} \leq 6e^{-c \min(\xi^2, \xi^4)} \quad \text{and}$$

$$P\left[\mu^*^T P_{S^1 \mid S} \mu^* - t(s^* - s + t) \beta^2]\right]$$

$$> \xi \left\{ 3(s^* - s + t)^{1/2} \right\} \leq 6e^{-c \min(\xi^2, \xi^4)},$$

where $c$ is a universal constant.

For simplicity, let $\Delta := 3\nu_1 (n-s+t)^{1/2} + 6(\nu_1 \nu_2)^{1/2} (n-s+t)^{1/2} + 3\nu_2 t^{1/2}$ and $\Delta' := 3(s^* - s + t) \beta^2$. Whenever $n > s^*$, there exists a universal constant $C_0 > 0$ such that $\max\{\Delta, \Delta'\} \leq C_0 (ns^*t)^{1/2} \beta^2 =: M \beta^2$. Writing $A = \{t(n+s^* - 2s + t) - \xi M\}^{1/2} \beta/2$ and $B = \{t(s^* - s + t) + \xi M\}^{1/2} \beta$, we deduce from the previous two bounds that

$$P\left(\|P_{S^1 \mid S} \mu^*\|^2_2 - \|P_{S^1 \mid S} \mu^*\|^2_2 < A - B \right)$$

$$\leq 12e^{-c \min(\xi^2, \xi^4)}.$$

Now we derive a lower bound on $(A - B)/\tau^{1/2}$. We have

$$\frac{A - B}{\tau^{1/2}} \geq \frac{A^2 - B^2}{2\tau^{1/2} A} \frac{\{\{n-s\}t - 2|\xi M|\} |\beta|}{2\tau^{1/2} \{t(n+s^* - 2s + t) + \xi M\}^{1/2}}.$$

If we have $n \geq 2s$, choosing $\xi = \xi_0 := (n-s)/4M$ then yields that

$$\frac{A - B}{\tau^{1/2}} \geq n^{1/2}/24.$$

Therefore,

$$P\left(\|P_{S^1 \mid S} \mu^*\|^2_2 - \|P_{S^1 \mid S} \mu^*\|^2_2 < \frac{n^{1/2}/24}{\xi_0} \right) \leq 12e^{-c \min(\xi_0^2, \xi_0^4)}.$$

Define $F(S) := \{S^1 \in \mathcal{A}^*_t(s) : S \cap S^* \subset S^1\}$. Applying a union bound over $F(S)$ yields that

$$P\left(\sum_{S \in \mathcal{A}_t(s)} m(S) \leq \frac{n^{1/2}/24}{\xi_0} \right) \leq 12|F(S)| e^{-c \min(\xi_0^2, \xi_0^4)}.$$

According to Stirling’s formula and the ultra-high dimension assumption, we have

$$|F(S)| = \left(\frac{s^* - s + t}{t}\right)^t \leq \exp s^* \leq e^{p+1}.$$ and

$$|A_t(s)| = \left(\frac{p - s^*}{s - t}\right)^t \leq e^{p+1}.$$

A union bound over $A_t(s)$ yields that there exists a universal constant $C_1 > 0$ such that

$$P\left(3S \in A_t(s) \text{ s.t. } m(S) < \frac{n^{1/2}}{24} \right) \leq 12|A_t(s)| \cdot |F(S)| e^{-c \min(\xi_0^2, \xi_0^4)}$$

$$\leq 12C_1 e^{p+2} e^{-c \min(\xi_0^2, \xi_0^4)}$$

$$\leq 12C_1 e^{3t} e^{-c \min(\xi_0^2, \xi_0^4) - 3t \log p}.$$

Note that

$$\xi_0 = \frac{(n-s)t}{4M} \geq \frac{nt}{8M} = \frac{(nt)^{1/2}}{8c_0 (s^*)^{1/2}}.$$
Assume that \( n \geq k^* s \ast (\log p)^2 \) for some \( \kappa > 64C_0^2 \). Then we have \( \xi_0 \geq \kappa^{1/2} (8C_0)^{-1/2} \log p > 1 \) and
\[
P(\exists S \in H_0(s) \text{ s.t. } m(S) < \frac{n^{1/2}}{24} | \beta |) \]
\[
\leq 12C_1 p^{-\epsilon(c^{1/2}/(8C_0)^{1-3})}. \tag{64}
\]
Further applying a union bound over \( t \in [s] \) yields that
\[
P(\sum_a m_a(s) \geq \frac{n^{1/2}}{24} | \beta |) \geq 1 - 12C_1 sp^{-\epsilon(c^{1/2}/C_0^{1-3})}. \tag{64}
\]
Finally, let \( C = \max\{12C_1, 8C_0/3, 8C_0/c, 2^{1/2}/3\} \). The first statement follows by letting \( \delta \geq Csp^{-\epsilon(c^{1/2}/C_0^{1-3})} \) which implies
\[
\kappa \geq (3 - \log(\delta/C)) / (\log p)^2 C^2 > 9C_0^2.
\]
The second statement holds by applying a union bound over \( s \in [s^* - 1] \) for (64) and letting \( \delta \geq C \ast s^2 \ast p^{-\epsilon(c^{1/2}/C_0^{1-3})} \).

**G. Proof of Theorem 5**

By Theorem 1, we know that whenever
\[
m_a(s) \geq 8 \sigma C \left( 1 - \frac{\log(\delta/C)}{2 \log p} \right)^{1/2} \cdot \frac{(\log p)^{1/2}}{1 - \gamma}, \tag{65}
\]
it holds that
\[
P(\{\text{FDP} (S) = 0, \forall \tilde{S} \in S(\delta, \eta) \}) \geq 1 - \delta. \tag{66}
\]
Consider the event \( E_1 = \{\text{FDP} (S) = 0, \forall \tilde{S} \in S(\delta, \eta) \} \), the event indicates that \( \tilde{S}_n \) is obtained within \( S^* \). Consider the sure screening event \( \mathcal{E} = \{S^* \in \tilde{S}_n \} \). If \( E_1 \cap \mathcal{E} \) holds, we have \( \tilde{L}_n = \tilde{L}_n \) and thus
\[
\tilde{S}(\delta, \eta, \tilde{S}) = S(\delta, \eta) \text{ and } S \in S^*, \forall S \in \tilde{S}(\delta, \eta, \tilde{S}).
\]
So
\[
P(\{\text{FDP} (S) = 0, \forall \tilde{S} \in \tilde{S}(\delta, \eta, \tilde{S}) \}) \
\geq P(E_1 \cap \mathcal{E}) = 1 - P(E_1^c \cup \mathcal{E}^c) \
\geq 1 - P(E_1^c) - P(\mathcal{E}^c) = P(E_1) - P(\mathcal{E}^c) \
\geq 1 - \delta - P(\mathcal{E}^c).
\]

**H. Calculation for the Examples**

**Calculation for Example 1:** Let \( S \in \mathcal{A}_n(s) \), let \( S = \{S \setminus S^* \} \). Note that for any \( S^* \subset S^* \), \( X_1^T X_1 | S | \setminus S^* = 0 \) and accordingly we have \( \|X_1^T \ast \mu^* \|^2 = \|X_1^T \ast \mu^* \|^2 = n^2 \beta_1^2 \). Then \( \Phi(S) \) can be any subset of \( S^* \) that contains \( S \setminus S^* \). Note that \( \Phi(S) = n^{-1} \sum_{j \in \Phi(S)} X_j X_j^T - n^{-1} \sum_{j \in \Phi(S) \setminus S^*} X_j X_j^T = P_{\Phi(S)} - P_{\Phi(S) \setminus S^*} \). Also note that \( \|S \setminus S^* \| \leq \Delta \). Accordingly, one can get \( \|P_{\Phi(S)} \ast \mu^* \|^2 = \Delta n b^2 \) and \( \|P_{\Phi(S) \setminus S^*} \ast \mu^* \|^2 = 0 \). Then \( m(S) = \Delta n b^2 / \Delta n b^2 = n b^2 / \Delta n b^2 \) for all \( S \in \mathcal{A}_n(s) \). Thus \( m(S) \leq \min_{S \in \mathcal{A}_n(s)} m(S) = n b^2 / \Delta n b^2 \).

**Calculation for Example 2:** Note that \( \mu^* = \sum_j e_j \beta_j \) and
\[
\|P_{X_j} \ast \mu^* \|^2 = \|X_j \ast X_j^T \|^2 = \sqrt{n} \beta_j \quad \text{for } j \in S^*, \{1\} \text{ and similarly}
\]
\[
\|P_{X_j} \ast \mu^* \|^2 = \sqrt{n} \gamma \quad \text{for } j \notin S^*, \{1\}.
\]

Then \( m_j^\dagger = 1 \). For \( j = |n \setminus S^* \), \( \|u_j - u_j^\dagger \|_2 = \sqrt{2} (1 + q^2)^{-1}/2 \) for \( j \neq j^\dagger \). So \( \{u_j \}_{j \in [n] \setminus S^*} \) is a \( \{2q(1 + q^2)^{-1}/2\}-\text{packing} \), and \( \|u_j - u_j^\dagger \|_2 = \sqrt{2} (1 + q^2)^{-1}/2 \) for \( j \neq j^\dagger \). Then \( m_j^\dagger = \sqrt{3} (1 + q^2)^{-1}/2 \). For \( j \neq j^\dagger \), Taking the values into Assumptions 1 and 2 gives the conditions for \( \gamma \). To see how the right hand side of (14) behaves, note that by L’Hopital’s law,
\[
\lim_{q \to 0} \frac{q}{(1 + q^2)^{-1/2} - 1} = \lim_{q \to 0} \frac{1}{(1 + q^2)^{-1/2}} = \infty.
\]

**Calculation for Example 3:** Note that for any \( j \in [n] \setminus S^* \), \( X_j \ast \mu^* \) is orthogonal to \( X_j \), so \( u_j = X_j \) and \( u_j = (q e_j + e_1) / (1 + q^2)^{-1}/2 \). The \( \delta_0 \)-packing calculation follows the one in the calculation for Example 2. For \( j \in \{p/2 + 1, \ldots, p\} \), let \( S = S_0 \cup \{j\} \). We have
\[
P_{\Phi(S)} = n^{-1} \sum_{j \in S^*} X_j X_j^T - n^{-1} \sum_{j \in S \setminus S^*} X_j X_j^T = P_{\Phi(S)} - P_{\Phi(S) \setminus S^*} \quad \text{and}
\]
then \( \|P_{\Phi(S)} \ast \mu^* \|_2 = \|P_{\Phi(S) \setminus S^*} \ast \mu^* \|_2 / |S^* \setminus S^*|^{1/2} = \sqrt{n} (\gamma - \beta) \) for any \( S \in \mathcal{A}_n(S) \). Thus
\[
\min_{S \in \mathcal{A}_n(S) \setminus S^*} \|P_{\Phi(S) \setminus S^*} \ast \mu^* \|_2 = \sqrt{n} (\gamma - \beta).
\]

Then it suffices to get (16) by taking the values into Assumptions 3 and 4. Notice that Assumption 4 requires \( m \geq \max \left\{ (8/\sqrt{\delta e^2 q^2 / \log p}, 2 \right\} = \max \{4 \sqrt{2} (1 + q^2)^{-1}/(1 + q^2)^{-1}/2 \} \). We need to check that
\[
\text{(16): } \frac{4 \sqrt{2} (1 + q^2)^{-1/2} \sigma (\log p / n)^{1/2}}{\epsilon \log p} < \infty.
\]

happens as \( q \to 0 \). We have it by L’Hopital’s law,
\[
\lim_{q \to 0} \frac{\text{A2}(q)}{\text{A1}(q)} = \frac{\text{A2}(q)}{\text{A1}(q)} = \infty.
\]

which is bigger than 1 by having \( p \) go to infinity.
APPENDIX C
TECHNICAL LEMMAS

Lemma 2: Consider $n$ independent and identically distributed observations $(Y_i, x_i)_{i \in [n]}$ of $(Y, x) \sim \mathcal{N}(0, \Sigma)$, where $Y$ is valued in $\mathbb{R}$ and $x$ is valued in $\mathbb{R}^p$. Write $X = (x_1, \ldots, x_n)^T$ and $y = (y_1, \ldots, y_n)^T$. Define $\nu_1 := \Sigma_{Xx} \Sigma_{XX}^{-1} \Sigma_{XY}$ and $\nu_2 := \Sigma_{XY} - \nu_1$. Then write that

$$
\mathbb{P}\left[ y^T P_x y - \left( \nu_1 + p \nu_2 \right) \right] > \left( 3 \nu_1 n \right)^{1/2} \\
+ \left( 6 \nu_1 \nu_2 \right)^{1/2} \leq 6e^{-c \min(\xi^2, \xi)},
$$

where $c$ is a universal constant.

Proof of Lemma 2: Let $y^T P_x y = \theta^T X^T X \theta + z^T X \theta + z^T P_x z$. We bound the three terms on the right-hand side one by one. Note that

$$
\| \theta^T x \|_{\nu_1} \leq 3 \nu_1,
\| a^T z \|_{\nu_1} \leq 3 \nu_2, \quad \text{and} \quad \| Z(\theta^T x) \|_{\nu_1} \leq 6(\nu_1 \nu_2)^{1/2}.
$$

Given that $z$ is independent of $X$, applying Bernstein's inequality yields that for any $\xi > 0$,

$$
\mathbb{P}\left[ \left( \begin{array}{c} \| X \theta \|_{\nu_1} \\
\| a \cdot z \|_{\nu_1} \\
\| Z(\theta^T x) \|_{\nu_1}
\end{array} \right) \geq \xi \right] \\
\leq 2 \exp\left[ -cn \min\left( \left( 3 \nu_1 \right)^{-2} \xi^2, \left( 3 \nu_1 \right)^{-1} \xi \right) \right],
$$

and

$$
\mathbb{P}\left[ \left( \begin{array}{c} \| z^T P_x z \|_p \\
\| \nu_2 \|_p
\end{array} \right) \geq \xi \right] \\
\leq 2 \exp\left[ -cp \min\left( \left( 3 \nu_2 \right)^{-2} \xi^2, \left( 3 \nu_2 \right)^{-1} \xi \right) \right],
$$

where $c$ is a universal constant. Combining the three bounds above yields the conclusion.

Lemma 3: Given two random variables $X_1$ and $X_2$ valued in $\mathbb{R}$, $\mathbb{E}\{ \max(X_1, X_2) \} \leq \mathbb{E}\{ X_1 \} + \mathbb{E}\{ X_2 \}$.

Proof: $\mathbb{E}\{ \max(X_1, X_2) \} = \mathbb{E}\{ (X_1 + X_2)/2 + |X_1 - X_2|/2 \} \leq \frac{1}{2} \mathbb{E}\{ X_1 + X_2 \} + \frac{1}{2} \mathbb{E}\{ X_1 - X_2 \} = \mathbb{E}\{ X_1 \} + \mathbb{E}\{ X_2 \}$.

Lemma 4 (40), Lemma 6.12): Let $\{ X_t \}_{t \in T}$ be a separable Gaussian process. Then $\sup_{t \in T} \mathbb{E}\{ X_t \}$ is subgaussian.

ACKNOWLEDGMENT

The authors appreciate the suggestions and guidance from Prof. Dimitris Bertsimas at MIT that substantially improved the quality of this article. They thank the Editor, the Associate Editor, and three anonymous reviewers, for constructive comments that have led to a significantly improved manuscript.

REFERENCES

[1] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in Selected papers of Hirotugu Akaike. Cham, Switzerland: Springer, 1998, pp. 199–213.
[2] H. Akaike, “A new look at the statistical model identification,” IEEE Trans. Autom. Control, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
[3] I. Barnett, R. Mukherjee, and X. Lin, “The generalized higher criticism for testing SNP-set effects in genetic association studies,” J. Amer. Stat. Assoc., vol. 112, no. 517, pp. 64–76, Jan. 2017.
[4] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” J. Roy. Stat. Soc., Ser. B, vol. 57, no. 1, pp. 289–300, Jan. 1995.
[5] D. Bertsimas, A. King, and R. Mazumder, “Best subset selection via a modern optimization lens,” Ann. Statist., vol. 44, no. 2, pp. 813–852, Apr. 2016.
[6] D. Bertsimas and B. Van Parys, “Sparse high-dimensional regression: Exact scalable algorithms and phase transitions,” Ann. Statist., vol. 48, no. 1, pp. 300–323, Feb. 2020.
[7] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” J. Fourier Anal. Appl., vol. 14, nos. 5–6, pp. 629–654, Dec. 2008.
[8] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressed sensing,” Appl. Comput. Harmon. Anal., vol. 27, no. 3, pp. 265–274, 2009.
[9] F. Chatziadriou et al., “Comprehensive field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma,” J. Natl. Cancer Inst., vol. 103, no. 16, pp. 1277–1235, Aug. 2011.
[10] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” SIAM J. Comput., vol. 20, no. 1, pp. 33–61, Jan. 1998.
[11] O. De Jager, B. Rabukenheimer, and J. Swanepoel, “A powerful test for weak periodic signals with unknown light curve shape in sparse data,” Astron. Astrophys., vol. 221, no. 1, pp. 180–190, Aug. 1989.
[12] J. Fan, Y. Guo, and Z. Zhu, “When is best subset selection the ‘best,’” 2020, arXiv:2007.01478.
[13] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” J. Amer. Stat. Assoc., vol. 96, no. 456, pp. 1348–1360, 2001.
[14] J. Fan, H. Liu, Q. Sun, and T. Zhang, “I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error,” J. Amer. Stat. Assoc., vol. 106, no. 504, pp. 814–841, Apr. 2011.
[15] J. Fan and J. Lv, “Sure independence screening for ultrahigh dimensional feature space,” J. Roy. Statist. Soc. B, Stat. Methodol., vol. 70, no. 5, pp. 849–911, 2008.
[16] J. Fan and J. Lv, “Nonconcave penalized likelihood with NP-dimensionality,” IEEE Trans. Inf. Theory, vol. 57, no. 8, pp. 5467–5484, Aug. 2011.
[17] J. Fan and H. Peng, “Nonconcave penalized likelihood with a diverging number of parameters,” Ann. Statist., vol. 32, no. 3, pp. 928–961, Jun. 2004.
[18] J. Ge et al., “Picasso: A sparse learning library for high dimensional data analysis in R and Python,” J. Mach. Learn. Res., vol. 20, no. 1, pp. 1692–1696, 2019.
[19] Y. Guo, Z. Zhu, and J. Fan, “Best subset selection is robust against design dependence,” 2020, arXiv:2007.01478.
[20] T. Hastie, R. Tibshirani, and R. Tibshirani, “Sure independence screening for ultrahigh dimensional feature space,” 2011, J. Mach. Learn. Res., vol. 10, no. 1, pp. 1227–1235, Aug. 2011.
[21] H. Hazimeh and R. Mazumder, “Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms,” Oper. Res., vol. 68, no. 5, pp. 1517–1537, Sep. 2020.
[22] H. Hazimeh, R. Mazumder, and T. Nonet, “L0learn: A scalable package for sparse learning using L0 regularization,” J. Mach. Learn. Res., vol. 24, no. 205, pp. 1–8, 2023.
[23] P. Jain, A. Tewari, and P. Kar, “On iterative hard thresholding methods for high-dimensional m-estimation,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 685–693.

[24] A. Javanmard and H. Javadi, “False discovery rate control via debiased Lasso,” Electron. J. Statist., vol. 13, no. 1, pp. 1212–1253, Jan. 2019.

[25] P.-L. Loh and M. J. Wainwright, “Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima,” J. Mach. Learn. Res., vol. 16, pp. 559–616, Mar. 2015.

[26] P.-L. Loh and M. J. Wainwright, “Support recovery without incoherence: A case for nonconvex regularization,” Ann. Statist., vol. 45, no. 6, pp. 2455–2482, Dec. 2017.

[27] C. L. Mallows, “Some comments on $c_p$,” Technometrics, vol. 15, no. 4, pp. 661–675, 1973.

[28] N. Meinshausen, “Relaxed lasso,” Comput. Statist. Data Anal., vol. 52, no. 1, pp. 374–393, Sep. 2007.

[29] A. Mousavi, A. Maleki, and R. G. Baraniuk, “False discoveries occur early for high-dimensional m-estimation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 20051–20060.

[30] B. K. Natarajan, “Sparse approximate solutions to linear systems,” SIAM J. Comput., vol. 24, no. 2, pp. 227–234, Apr. 1995.

[31] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” Appl. Comput. Harmon. Anal., vol. 26, no. 3, pp. 301–321, May 2009.

[32] M. Rudelson and R. Vershynin, “Hanson-wright inequality and sub-Gaussian concentration,” Electron. Commun. Probab., vol. 18, pp. 1–9, Jan. 2013.

[33] G. Schwarz, “Estimating the dimension of a model,” Ann. Statist., vol. 6, no. 2, pp. 461–464, Mar. 1978.

[34] W. Su, M. Bogdan, and E. Candès, “False discoveries occur early on the lasso path,” Ann. Statist., vol. 45, no. 5, pp. 2133–2150, Oct. 2017.

[35] W. J. Su, “When is the first spurious variable selected by sequential regression procedures?” Biometrika, vol. 105, no. 3, pp. 517–527, Sep. 2018.

[36] Q. Sun, R. Zhu, T. Wang, and D. Zeng, “Counting process-based dimension reduction methods for censored outcomes,” Biometrika, vol. 106, no. 1, pp. 181–196, Mar. 2019.

[37] R. Sun and X. Lin, “Genetic variant set-based tests using the generalized Berk–Jones statistic with application to a genome-wide association study of breast cancer,” J. Amer. Stat. Assoc., vol. 115, no. 531, pp. 1079–1091, Jul. 2020.

[38] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. Roy. Stat. Soc., Ser. B, Methodol., vol. 58, no. 1, pp. 267–288, Jan. 1996.

Ziwei Zhu (Member, IEEE) received the Ph.D. degree from the Department of Operations Research and Financial Engineering, Princeton University, in 2018, supervised by Prof. Jianqing Fan. He was an Assistant Professor of statistics with the University of Michigan, Ann Arbor, MI, USA, from 2019 to 2022. He is currently a Quantitative Researcher with Radix Trading.

Shihao Wu received the B.S. degree in data science from Fudan University, Shanghai, China, in 2021. He is currently pursuing the Ph.D. degree with the Department of Statistics, University of Michigan, Ann Arbor, MI, USA.