Journal of Machine Learning Research 25 (2024) 1-66 Submitted 5/23; Revised 6/24; Published 12/24

A minimax optimal approach to high-dimensional double sparse linear regression

Yanhang Zhang ZHANGYH98@RUC.EDU.CN
School of Statistics, Renmin University of China
100872 Beijing, China

Zhifan Li * ZHIFANLI@BIMSA.CN
Beijing Institute of Mathematical Sciences and Applications
101408 Beijing, China

Shixiang Liu LIUSHIXIANG_STAT@RUC.EDU.CN
School of Statistics, Renmin University of China
100872 Beijing, China

Jianxin Yin † JYIN@RUC.EDU.CN
Center for Applied Statistics and School of Statistics, Renmin University of China
100872 Beijing, China

Editor: Daniel Hsu

Abstract

In this paper, we focus our attention on the high-dimensional double sparse linear regression, that is, a combination of element-wise and group-wise sparsity. To address this problem, we propose an IHT-style (iterative hard thresholding) procedure that dynamically updates the threshold at each step. We establish the matching upper and lower bounds for parameter estimation, showing the optimality of our proposal in the minimax sense. More importantly, we introduce a fully adaptive optimal procedure designed to address unknown sparsity and noise levels. Our adaptive procedure demonstrates optimal statistical accuracy with fast convergence. Additionally, we elucidate the significance of the element-wise sparsity level s_0 as the trade-off between IHT and group IHT, underscoring the superior performance of our method over both. Leveraging the beta-min condition, we establish that our IHT-style procedure can attain the oracle estimation rate and achieve almost full recovery of the true support set at both the element level and group level. Finally, we demonstrate the superiority of our method by comparing it with several state-of-the-art algorithms on both synthetic and real-world datasets.

Keywords: double sparsity, iterative hard thresholding, minimax optimality, fully adaptive procedure, oracle estimation rate.

1. Introduction

Over the last decade, the rapid growth of high-dimensional data has drawn broad attention to sparse learning across many scientific communities, with plenty of remarkable achievements in algorithms, theory, and applications. One of the well-studied problems

∗. Yanhang Zhang and Zhifan Li contributed equally to this work.
†. Corresponding author

is the sparsity-constrained linear regression, also known as the best subset selection. We consider a linear model

$$y = X\beta^* + \xi,$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the design matrix, $\beta^* \in \mathbb{R}^p$ is the underlying regression coefficient and $\xi \in \mathbb{R}^n$ is the sub-Gaussian random error with scale parameter σ^2. In the high-dimensional framework, we focus on the case where $p \gg n$ and the coefficient β^* is sparse in the sense that only a few covariates are important to the model. Traditionally, element-wise ℓ_0 sparse problem considers the parameter space

$$\beta^* \in \{\beta \in \mathbb{R}^p : \sum_{i=1}^{p} \mathrm{I}(\beta_i \neq 0) \leq s'\},$$

where β_i is the ith entry of β and s' is some positive integer, which controls the sparsity level of the sparsity-constrained linear regression problem. Best subset selection is a famous NP-hard problem (Natarajan, 1995), and it has been widely studied in the fields of statistics and machine learning (Bertsimas et al., 2016; Yuan et al., 2018; Huang et al., 2018; Zhu et al., 2020).

Recently, an increasing number of studies on high-dimensional variable selection have focused on the concept of structured sparsity. These studies assume that important variables form specific structures or patterns, with group-wise sparsity being one of the most prominent examples. The group-wise ℓ_0 sparsity considers the parameter space

$$\beta^* \in \{\beta \in \mathbb{R}^p : \sum_{j=1}^{m} \mathrm{I}(\beta_{G_j} \neq 0) \leq s\},$$

where $\{G_j\}_{j=1}^m$ are the indices of m non-overlapping groups such that $\cup_{j=1}^m G_j = \{1, \ldots, p\}$. Here positive integer s controls the number of nonzero groups in the model. The group sparsity means that within a group, the coefficients are either all zeros or at least one nonzero. In particular, when $|G_1| = \ldots = |G_m| = 1$, the group selection problem boils down to the standard best subset selection. To date, a variety of practical algorithms have been explored and investigated to conduct group ℓ_0 selection (Eldar et al., 2010; Huang et al., 2011; Hazimeh et al., 2023; Zhang et al., 2023).

When considering each group that has been selected, it is generally accepted that only a few of the variables that make up the group are actually significant. We refer to this idea as double sparsity and define it as follows:

Definition 1 (Double sparsity) *The regression coefficient $\beta^* \in \mathbb{R}^p$ is called (s, s_0)-sparse if*

$$\|\beta^*\|_{0,2} \coloneqq \sum_{j=1}^{m} \mathrm{I}(\beta^*_{G_j} \neq 0) \leq s \quad \textit{and} \quad \|\beta^*\|_0 \coloneqq \sum_{i=1}^{p} \mathrm{I}(\beta^*_i \neq 0) \leq ss_0. \tag{1}$$

Double sparsity promotes sparsity both within and between groups. Specifically, it restricts the number of nonzero groups included in the model to s, and within these s groups, the number of nonzero elements must be no more than ss_0. Intuitively, s_0 can be thought of as the average sparsity within the s selected groups, providing insight into the sparsity levels within the nonzero groups.

1.1 Related Work

Recently, sparse group selection has emerged as a prominent area of high-dimensional structured sparsity learning. To tackle this problem, a combination of two penalized methods is often considered. In order to perform sparse group selection, Friedman et al. (2010) and Simon et al. (2013) proposed sparse group Lasso (SGLasso), a combination of the Lasso penalty (Tibshirani, 1996) and the group Lasso penalty (Yuan and Lin, 2006) joined together. Numerous efforts have been dedicated to accelerating the convergence of SGLasso (Ida et al., 2019; Zhang et al., 2020).

The theoretical research on double sparsity began with Cai et al. (2022), which established the minimax lower bounds for the estimation error of the double sparse linear regression, and the near-optimal upper bounds for the estimation error of SGLasso are obtained under the irrepresentable condition. Moreover, they provided the theoretical guarantees for both the sample complexity and estimation error of SGLasso. Li et al. (2024) concentrated on the Gaussian location model with a double sparse structure. They established the minimax rates for the estimation error over $\ell_u(\ell_q)$ mixed-norm for $u, q \in [0, 1]$. Despite these advancements, there still remains a dearth of methods with optimal theoretical guarantees.

Traditional convex relaxation-based methods, such as SGLasso, inherently introduce estimation bias for the coefficients, especially when large coefficients undergo significant shrinkage. Moreover, Bellec (2018) demonstrated that convex estimators, such as the Lasso-type estimator, cannot attain the oracle estimation rate $O(\sigma\sqrt{\frac{ss_0}{n}})$, even when the beta-min condition is satisfied. This phenomenon motivates us to develop computationally feasible non-convex algorithms, with iterative hard thresholding (IHT, Blumensath and Davies (2009)) being a representative example. IHT and its variants have garnered increasing attention for their efficacy in addressing a variety of high-dimensional statistical inference problems (Blumensath and Davies, 2010; Jain et al., 2014; Yuan et al., 2020; Hao et al., 2021). Given sparsity level s', IHT performs a gradient descent step on the parameter β, followed by the selection of the s' largest absolute values at each subsequent step. Under restricted convexity/smoothness conditions, Jain et al. (2014) showed that IHT can obtain a minimax optimal estimator for high-dimensional M-estimation given a sufficient large sparsity level. Yuan et al. (2018) investigated the parameter estimation and support recovery of IHT for both $s = s^*$ and $s \gg s^*$ under RIP-type conditions. Giraud (2021) employed the IHT procedure in the context of linear regression with group sparsity and established the optimal upper bound for parameter estimation. However, most of the related works consider the known sparsity level s' as prior information, making it challenging to analyze theoretical guarantees in the non-asymptotic sense without the knowledge of s'. To tackle this problem, Ndaoud (2020) proposed a fully adaptive IHT-style procedure, which can achieve the optimal rates for parameter estimation with unknown s'.

1.2 Main Results and Contributions

In this paper, our goal is to construct feasible methods for double sparse linear regression that are not only efficient but also with optimal statistical properties guaranteed. To the best of our knowledge, our paper is the first to develop a fully adaptive optimal procedure for high-dimensional double sparse linear regression with unknown s, s_0, and σ.

Addressing the signal under the double sparse assumption was an unresolved challenge until Cai et al. (2022); Li et al. (2024). The approach employed in Cai et al. (2022) relies on sub-gradient and dual certificate constructions, applicable only in the context of ℓ_1-type penalties. An earlier work by Li et al. (2024) introduced an IHT-style algorithm for detecting signals with a double sparse structure. They demonstrated the minimax optimality of the proposed algorithm for parameter estimation. However, this algorithm is impractical because it depends on the unknown parameters s, s_0, and σ. Notably, achieving adaptivity for double sparsity is much more challenging than for element-wise or group-wise sparsity. A natural approach is using a grid search technique for tuning the unknown parameters s and s_0 such as Cai et al. (2022). However, the grid search approach is computationally infeasible, and difficult to establish optimal guarantees from a theoretical perspective. Motivated by the adaptive framework for element-wise sparsity (Verzelen, 2012; Ndaoud, 2020), we develop a two-step adaptive procedure for parameter estimation and variable selection in the context of double sparse linear regression.

Importantly, our procedure is not a simple combination of classical IHT (Ndaoud, 2020) and group IHT (Giraud, 2021). The sequence of our two-step IHT operators is critical and the order cannot be interchanged. Specifically, reversing the order of these two steps could compromise the logical framework of the proof by contradiction.

The advantages of double sparse IHT over convex counterparts, such as sparse group Lasso, are evident. Our theory is entirely based on the RIP-type condition, while the theory of sparse group Lasso (cf. Cai et al. (2022)) relies on a stronger irrepresentable condition. We further establish that under the beta-min conditions, our algorithm can achieve the oracle estimation rate $O(\sigma\sqrt{\frac{ss_0}{n}})$, showcasing the superiority of our algorithm over sparse group Lasso. Moreover, as far as we know, support recovery results in sparse group Lasso have not been established under mild assumptions, while we obtain the almost full recovery (Butucea et al., 2018) at both the element-wise and group-wise levels. This is further supported by synthetic and real-world data analyses.

In conclusion, the main contribution of this paper is summarized as follows:

- We introduce a novel double sparse IHT operator that ensures both element-wise and group-wise sparsity. This operator consists of two steps that control model complexity efficiently. Building upon the double sparse IHT operator, we introduce a novel IHT-style procedure that dynamically updates the threshold at each iteration. We analyze upper bounds on the estimation error of our method and establish matching minimax lower bounds for the estimation error $O\left(\sqrt{\frac{\sigma^2}{n}(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s})}\right)$, conclusively demonstrating the optimality of our proposed approach.

- We propose a fully adaptive optimal procedure that handles unknown sparsity levels s, s_0 and noise level σ. Through our research, we demonstrate that the estimator obtained by our adaptive procedure attains optimal performance in the minimax sense. As far as we know, it is the first minimax adaptive procedure for the double sparse linear regression. Furthermore, we discover the pivotal role of the element sparsity level s_0 as the trade-off between IHT and group IHT, underscoring the superior performance of our method over both. We have implemented our proposals in an open-source R package named `ADSIHT`.

- Under the element-wise and group-wise beta-min conditions, we establish that our algorithm attains the oracle estimation rate $O(\sigma\sqrt{\frac{ss_0}{n}})$. This result indicates that our procedure performs comparably to the ordinary least-squares estimator when given the true support set. It highlights the superiority of our DSIHT procedure over convex counterparts such as sparse group Lasso in theory. Additionally, we demonstrate that our procedure achieves almost full recovery of the true support set at both the element and group levels.

- We apply our proposed methods to both synthetic and real-world datasets, and comprehensive empirical comparisons with several state-of-the-art methods show the superiority of our method across a variety of metrics. Additionally, computational results for a real-world dataset demonstrate that our approach produces more accurate predictive power with fewer variables and groups.

1.3 Organization

The remainder of the paper is structured as follows. We introduce the notation used throughout the paper towards the end of this section. In Section 2, we introduce an IHT-style procedure with fast convergence and establish matching upper and lower bounds for estimation error. In Section 3, we firstly propose a novel information criterion to determine the optimal stopping time and develop an adaptive procedure for conducting sparse group selection with unknown s and σ. Then, we elucidate the connection between our work, IHT, and group IHT. We also present a minimax adaptive procedure to select the optimal value of s_0, which makes our method a fully adaptive optimal procedure. In Section 4, we establish that our DSIHT algorithm achieves the oracle estimation rate and accomplishes almost full recovery under the beta-min conditions. In Section 5, we present numerical experiments comparing our methods with several state-of-the-art approaches using both synthetic and real-world datasets. Finally, in Section 6, we provide a summary of our study and offer detailed proofs of our main results in the Appendix.

1.4 Notations

For the given sequences a_n and b_n, we say that $a_n = O(b_n)$ or $a_n \lesssim b_n$ (resp. $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$) when $a_n \leq cb_n$ (resp. $a_n \geq cb_n$) for some positive constant c. We write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. Let $d = \max_{1\leq j\leq m} |G_j|$ be the maximum group size. Denote $[m]$ as the set $\{1, 2, \ldots, m\}$, and $\mathrm{I}(\cdot)$ as the indicator function. Let $x \vee y$ be the maximum of x and y, while $x \wedge y$ is the minimum of x and y. Denote $S^* = \{i : \beta_i^* \neq 0\} \subseteq [p]$ as the support set of β^*. Similarly, let $G^* = \{j : \beta_{G_j}^* \neq 0, G_j \subseteq [p], \text{ and } G_j \cap G_{j'} = \emptyset, \forall j \neq j'\} \subseteq [m]$ be the group-wise support set of β^*. Let $S_{G^*} = \cup_{j\in G^*} G_j$ be all the elements contained in groups G^*. Obviously, $S^* \subseteq S_{G^*}$. For any set S with cardinality $|S|$, let $\beta_S^* = (\beta_j, j \in S) \in \mathbb{R}^{|S|}$ and $X_S = (X_j, j \in S) \in \mathbb{R}^{n\times|S|}$, and let $(X^\top X)_{SS} \in \mathbb{R}^{|S|\times|S|}$ be the submatrix of $X^\top X$ whose rows and columns are both listed in S. For a vector β, denote $\|\beta\|_2$ as its Euclidean norm. For a matrix A, denote $\|A\|_2$ as its spectral norm and $\|A\|_F$ as its Frobenius norm. Denote $\mathbb{I}_p$ as the $p \times p$ identity matrix. Let $C, C_0, C_1, \ldots$ denote positive constants whose actual values vary from time to time. Denote the parameter space of double sparsity as $\Theta^{m,d}(s, s_0)$. Denote $\mathcal{S}^{m,d}(s, s_0)$ as the space consisting of all the support sets

of (s, s_0)-sparse vector. Notably, according to the definition of double sparsity, we have $\mathcal{S}^{m,d}(a_1 s, b_1 s_0) \subseteq \mathcal{S}^{m,d}(a_2 s, b_2 s_0)$ for any positive constants $a_1 b_1 = a_2 b_2$ and $a_1 \leq a_2$. For example, $\mathcal{S}^{m,d}(2s, 2s_0)$ is a subspace of $\mathcal{S}^{m,d}(4s, s_0)$. To facilitate computation, we assume $\|X_j\|_2 = \sqrt{n}$, $\forall j \in [p]$.

2. Analysis of minimax optimality

In Section 2.1, we introduce the double sparse iterative hard thresholding (DSIHT) operator. In particular, we provide a clear explanation of its construction and develop a DSIHT algorithm with known sparsity and noise levels. Following this, in Section 2.2, we analyze the sources of estimation error. Then, we establish the upper bounds for parameter estimation of the DSIHT algorithm in Section 2.3. In Section 2.4, we derive the minimax lower bound for double sparse linear regression, which yields that the upper bound in Section 2.3 is minimax optimal.

2.1 Double sparse iterative hard thresholding operator

Given $\lambda, s_0 > 0$, we define the double sparse iterative hard thresholding operator $\mathcal{T}_{\lambda,s_0} : \mathbb{R}^p \to \mathbb{R}^p$ as the following two steps:

Step 1 (Element-wise Condition Checking): define an element-wise hard thresholding operator $\mathcal{T}_\lambda^{(1)} : \mathbb{R}^p \to \mathbb{R}^p$ on $\beta \in \mathbb{R}^p$ as

$$\{\mathcal{T}_\lambda^{(1)}(\beta)\}_j = \beta_j \mathrm{I}(|\beta_j| \geq \lambda), \quad \forall j \in [p].$$

The operator $\mathcal{T}_\lambda^{(1)}$ preserves the signal whose absolute magnitude is greater than or equal to λ, thus it can be seen as a preliminary screening process for identifying important variables.

Step 2 (Group-wise Condition Checking): denote

$$\mathcal{J}_{s_0} \coloneqq \{j \in [m] : \|\beta_{G_j}\|_2^2 \geq s_0 \lambda^2\}.$$

The definition of operator $\mathcal{T}_{\lambda,s_0}^{(2)} : \mathbb{R}^p \to \mathbb{R}^p$ is

$$\{\mathcal{T}_{\lambda,s_0}^{(2)}(\beta)\}_{G_j} = \begin{cases} \beta_{G_j}, & \text{if } j \in \mathcal{J}_{s_0}. \\ 0, & \text{if } j \in [m] \backslash \mathcal{J}_{s_0}. \end{cases}$$

The operator $\mathcal{T}_{\lambda,s_0}^{(2)}$ selects groups with large magnitudes, utilizing group information to further filter the important variables. The operator $\mathcal{T}_{\lambda,s_0} = \mathcal{T}_{\lambda,s_0}^{(2)} \circ \mathcal{T}_\lambda^{(1)}$ is a composition of these two steps. Unlike the classical IHT procedure, our procedure updates the threshold λ in $\mathcal{T}_{\lambda,s_0}$ at each step in order to achieve both optimal statistical accuracy and fast convergence. Given $\lambda_0 > \lambda_\infty > 0$ and $0 < \kappa < 1$, we provide the form of the sequence $\{\lambda_t\}_{t=1}^\infty$ as follows

$$\lambda_t = \kappa^t \lambda_0 \vee \lambda_\infty, \ t = 0, 1, 2, \ldots \tag{2}$$

For a given s_0 and sequence of threshold $\{\lambda_t\}_{t=1}^\infty$, we denote the estimators $\{\beta^t\}_{t=1}^\infty$ as

$$\beta^t = \mathcal{T}_{\lambda_t,s_0}\left(\beta^{t-1} + \frac{1}{n} X^\top (y - X\beta^{t-1})\right), \ t = 1, 2, \ldots. \tag{3}$$

Moreover, we denote the corresponding support set of $\{\beta^t\}_{t=1}^{\infty}$ as $\{S^t\}_{t=1}^{\infty}$. In the studies of variable selection, the misidentification of true support set S^*, i.e., $S^t \cap (S^*)^c$ is called type-I error, and the omission of S^*, i.e., $(S^t)^c \cap S^*$ is called type-II error. We summarize our procedure as the following algorithm:

Algorithm 1 Double **S**parse **IHT** (DSIHT) algorithm with known s, s_0 and σ.

Require: $X,\ y,\ \{G_j\}_{j=1}^m,\ \kappa,\ \lambda_0,\ s_0,\ s,\ \sigma$.

1: Initialize $t = 0$, $\beta^t = 0$ and $\lambda_\infty = 4\sqrt{\frac{\sigma^2}{n}(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})}$.
2: **while** $\lambda_t \geq \lambda_\infty$, **do**
3: $\quad \beta^{t+1} = \mathcal{T}_{\lambda_t, s_0}\left(\beta^t + \frac{1}{n}X^\top(y - X\beta^t)\right)$.
4: $\quad \lambda_{t+1} = \kappa\lambda_t$.
5: $\quad t = t + 1$.
6: **end while**

Ensure: $\hat{\beta} = \beta^t$.

Here we offer an intuitive explanation for the choice of λ_t. A large λ_t promotes sparsity in the estimator β^t, which significantly reduces the type-I error by preventing spurious variables from being incorporated into the model. However, excessive sparsity can result in a high type-II error by omitting too many true variables. As Section 2.2 shows, it leads to a high estimation error because the magnitude of β^* is drastically shrunk to zero. Conversely, a small λ_t can reduce the type-II error by increasing the complexity of the model. Nevertheless, this allows too many spurious variables into the model, resulting in a high type-I error. This intuition motivates us to choose the specific form of the sequence $\{\lambda_t\}_{t=1}^{\infty}$ by balancing these two types of errors.

In our procedure, we employ a decreasing sequence (2) instead of directly setting the threshold as this order. The reason is that such a small threshold can potentially result in the selection of too many unimportant variables at the initial step. This lack of sparsity makes our procedure hard to benefit from the contraction property of the DSRIP condition, and the estimation error cannot be well-controlled in iterations. In comparison, a sufficiently large λ_0 identifies a small set of variables, effectively controlling the false discoveries of the initial solution. With the decrease of the threshold, we optimize the solution in an appropriate direction iteratively without losing sparsity. A novelty of our procedure lies in the fact that it implicitly controls the type-I error at a low level at each step, and reduces the type-II error through iterations. In Theorem 4, we choose $\lambda_\infty \asymp \sqrt{\frac{\sigma^2}{n}(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})}$ and show its optimality in the minimax sense.

2.2 Analysis of estimation error

To conduct the theoretical analysis, we decompose the iterative term into three parts:

$$\begin{aligned} H^{t+1} :=& \beta^t + \frac{1}{n}X^\top(y - X\beta^t) \\ =& \beta^* + \left(\frac{1}{n}X^\top X - \mathbb{I}_p\right)(\beta^* - \beta^t) + \frac{1}{n}X^\top\xi \\ =& \beta^* + \Phi(\beta^* - \beta^t) + \Xi, \end{aligned} \tag{4}$$

where $\Phi \coloneqq \frac{1}{n}X^\top X - \mathbb{I}_p$ and $\Xi \coloneqq \frac{1}{n}X^\top \xi$. Equation (4) shows that the estimation error comes from three sources:

- The true parameters β^* shrunk by mistake.
- The optimization error that β^t approximates β^*.
- The randomness caused by the errors ξ.

Among these three sources, the optimization error corresponds to the iterative procedure, and the randomness of our proposed procedure mainly comes from the third term Ξ. In what follows, we detail how to upper bound the latter two sources of errors accurately. Firstly, we introduce an essential condition for the design matrix X in order to get a contraction of the optimization error.

Assumption 1 (DSRIP condition) *We say that $X \in \mathbb{R}^{n\times p}$ satisfies the Double Sparse Restricted Isometry Property $DSRIP(s, s_0, \delta)$ with constant $0 < \delta < 1$, if $\forall S \in \mathcal{S}^{m,d}(s, s_0)$ and $\forall u \neq 0, u \in \mathbb{R}^{|S|}$, it holds that*

$$1 - \delta \leq \frac{\|X_S u\|_2^2}{n\|u\|_2^2} \leq 1 + \delta.$$

Remark 2 *The Double Sparse Restricted Isometry Property (DSRIP) serves as a natural extension of the ordinary RIP condition (Candes and Tao, 2005) under the double sparse linear regression. For sub-Gaussian design, considering a p-dimensional ss_0-sparse structure, we require a sample size of $n = \Omega(ss_0 \log \frac{ep}{ss_0})$ to ensure that the RIP condition holds with high probability. However, for the satisfaction of the DSRIP condition, we only need $n = \Omega(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s})$. It is worth noting that, given $p = m \times d$, the DSRIP condition can be satisfied with a smaller sample size compared to RIP. Further details can be found in Appendix C.*

DSRIP serves as an essential component for analyzing the high-dimensional double sparse linear regression (Li et al., 2024). It imposes a less stringent condition than the ordinary RIP. Assuming the same element-wise sparsity, DSRIP only requires subsets of ss_0-sparse vectors with no more than s groups to be satisfied, whereas RIP requires all ss_0-sparse vectors to hold. If design matrix X satisfies DSRIP(s, s_0, δ), we have $\|\Phi\|_2 \leq \delta < 1$, demonstrating that Φ serves as the contraction factor for all (s, s_0)-sparse vectors. As a result, by leveraging both DSRIP and the sparse structure of the signal, the contraction factor Φ enables iterative reduction of the optimization error.

Next, we turn to the analysis of the random error term Ξ. To upper bound this source of error, we need to capture the complexity of the noise term.

Lemma 3 *Assume that X satisfies $DSRIP(s, s_0, \delta)$. Then, there exists a constant $C > 0$, the event*

$$\mathcal{E} \coloneqq \left\{ \forall S \in \mathcal{S}^{m,d}(s, s_0) : \sum_{i\in S} \Xi_i^2 \leq \frac{4\sigma^2}{n}\left(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s} \right) \right\}$$

holds with probability at least $1 - \exp\left\{-C(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s})\right\}$.

Lemma 3 provides the uniform upper bounds of the random error term with high probability. We now analyze the random term Ξ in detail and decompose the source of random errors into two parts:

- The random errors Ξ attached to the true support set S^*.
- The random errors Ξ caused by type-I error, the mis-identification of true parameters β^*. More concretely, some random errors escape from operator $\mathcal{T}_{\lambda,s_0}$, which we call these errors as pure errors below.

The errors caused by random errors Ξ can be attributed to two sources: the random errors corresponding to S^* and $(S^*)^c$, respectively. Since $S^* \in \mathcal{S}^{m,d}(s, s_0)$ is with a sparse prior, the random errors attached to S^* can be well-bounded by event $\mathcal{E}$ with high probability. However, it is difficult to find an upper bound for the pure errors since the amount of the pure errors is undetermined. Therefore, the central problem that operator $\mathcal{T}_{\lambda,s_0}$ addresses is to bound the support set of the pure errors. Intuitively, we want to collect the pure errors in some subsets belonging to $\mathcal{S}^{m,d}(s, s_0)$. Then, the magnitude of pure errors can be upper bounded by event $\mathcal{E}$.

We consider applying $\mathcal{T}_{\lambda,s_0}$ to the pure errors directly and show that if the pure errors overflow $\mathcal{S}^{m,d}(s, s_0)$, it will contradict with $\mathcal{E}$ with high probability. According to the structure of $\mathcal{S}^{m,d}(s, s_0)$, we decompose the discussion into two cases:

Case 1: Assume that the set selected by $\mathcal{T}_{\lambda,s_0}$ lies in no more than s groups but the amount exceeds ss_0. Element-wise condition checking ensures that all the selected entries are larger than λ. Then, for any (s, s_0)-shaped subset of this set with cardinality ss_0, the total magnitude of these subsets exceeds $ss_0\lambda^2$. With the choice of $\lambda \geq 2\sqrt{\frac{\sigma^2}{n}(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})}$, we have $ss_0\lambda^2 \geq 4\frac{\sigma^2}{n}(ss_0\log\frac{ed}{s_0} + s\log\frac{em}{s})$, which contradicts event $\mathcal{E}$ with high probability. We provide an illustrative example in Figure 1.

Case 2: Assume that the set selected by $\mathcal{T}_{\lambda,s_0}$ lies in more than s groups, yet within any s selected groups, the number of the selected entries does not exceed ss_0. Group-wise condition checking implies that the magnitude of each selected group is larger than $s_0\lambda^2$. Consequently, the (s, s_0)-shaped subset consisting of any s selected groups satisfies that the total magnitude exceeds $ss_0\lambda^2$. For $\lambda \geq 2\sqrt{\frac{\sigma^2}{n}(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})}$, it contradicts with event $\mathcal{E}$ with high probability. We provide an illustrative example in Figure 2. Notably, if there exist s selected groups with the number of selected entries exceeding ss_0, we analyze this case similarly to **Case 1**.

Overall, by applying operator $\mathcal{T}_{\lambda,s_0}$ directly, Ξ can be shrunk into a (s, s_0)-shaped subset with high probability.

2.3 Upper bound for estimation error

In Section 2.2, we have introduced the idea to control the estimation error caused by optimization error and randomness. Formally speaking, the three sources of estimation error can be bounded in sequence. In what follows, we analyze the error bounds of our proposed procedure. The main result of our theoretical analysis is given by Theorem 4.

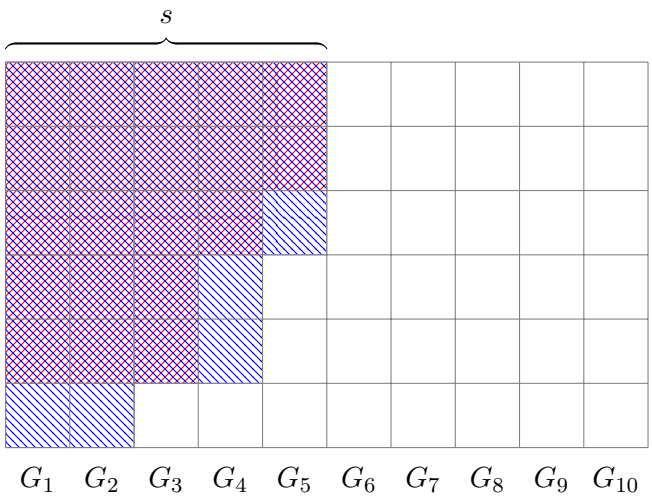

Figure 1: Illustrative example of **case 1**. There are 10 groups with equal group size $d = 6$, and we reshape the group structure as a 6×10 matrix with each column representing a group. Here $s = 5$ and $s_0 = 4$. The blue region represents the selected set, and the red region represents a (s, s_0)-shaped subset satisfying that total magnitude exceeds $ss_0\lambda^2$. Here the cardinality of the red-colored set is $s \times s_0 = 20$. Note that the whole vector of support is reshaped into a matrix with a particular group structure.

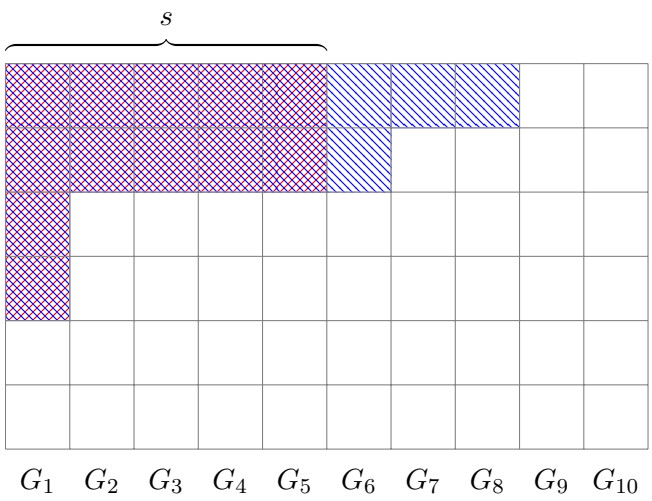

Figure 2: Illustrative example of **case 2**. The elements in Figure 2 are the same as in Figure 1. The entries of the red region cover $s = 5$ groups and its cardinality is less than $s \times s_0 = 20$.

Theorem 4 *Assume that β^* is (s, s_0)-sparse and X satisfies DSRIP$(3s, \frac{5}{3}s_0, \delta)$. Assume that $\delta < 0.11 \wedge \kappa^{10}$, $\|\beta^*\|_2 \leq \sqrt{ss_0}\lambda_0$ and $\lambda_\infty \geq 4\sqrt{\frac{\sigma^2}{n}(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})}$. We run Algorithm 1 and obtain the corresponding solution sequence $\{\beta^t\}, t = 1, 2, \cdots$. Then, with probability at least $1 - \exp\left\{-C(ss_0\log\frac{ed}{s_0} + s\log\frac{em}{s})\right\}$, we have*

(i) Inside groups G^*, the type-I error can be controlled by a (s, s_0)-shaped subset, that is,

$$S_{G^*} \cap S^t \cap (S^*)^c \in \mathcal{S}^{m,d}(s, s_0). \tag{5}$$

(ii) Outside groups G^*, the type-I error can be controlled by a (s, s_0)-shaped subset, that is,

$$S^c_{G^*} \cap S^t \in \mathcal{S}^{m,d}(s, s_0). \tag{6}$$

(iii) The upper bounds for estimation error are

$$\|\beta^* - \beta^t\|_2 \leq \frac{3}{2}(1 + \sqrt{2})\sqrt{ss_0}\lambda_t. \tag{7}$$

Part (i) of Theorem 4 shows that the type-I error of $\{\beta^t\}$ within the true groups G^* can be controlled in a (s, s_0)-shaped set. Part (ii) of Theorem 4 asserts that our procedure selects fewer than s incorrect groups into the model, and at most ss_0 variables outside groups G^*. Together, they show that the solution sequence $\{\beta^t\}$ generated by our procedure is $(2s, \frac{3}{2}s_0)$-sparse at each step, affirming that our procedure effectively controls false discoveries at both the element and group levels. The non-convexity of the IHT-style method may cause the parameter estimation error to not decrease at each step. To address this issue, a common approach to get around this issue is constructing a surrogate function of the upper bound that decreases exponentially (Yuan et al., 2018; Zhu et al., 2020; Zhang et al., 2023). With the choice of $\{\lambda_t\}$, (7) gives a decreasing upper bound for the parameter estimation error. Notably, with the choice of $\lambda_\infty \asymp \sqrt{\frac{\sigma^2}{n}(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})}$, the upper bound decays geometrically to the minimax lower bounds in (8), which demonstrates the optimality of our procedure in the minimax sense.

Remark 5 *In the above discussion, we have discussed the idea of the construction of $\mathcal{T}_{\lambda,s_0}$ by applying it to Ξ directly. In our practical procedure, we apply $\mathcal{T}_{\lambda,s_0}$ to H^t rather than Ξ. Referring to the two cases above, we can show that*

(i) Inside the true groups G^*, if $S^t \cap (S^*)^c \notin \mathcal{S}^{m,d}(s, s_0)$, there exists a (s, s_0)-shaped subset $\tilde{S}_{1,t} \subseteq S^t \cap S_{G^*} \cap (S^*)^c$ such that $ss_0\lambda^2_{t+1} \leq \sum_{i\in\tilde{S}_{1,t}}\{\mathcal{T}_{\lambda_{t+1},s_0}(H^{t+1})\}^2_i$.

(ii) Outside the true groups G^*, if $S^t \cap (S^*)^c \notin \mathcal{S}^{m,d}(s, s_0)$, there exists a (s, s_0)-shaped subset $\tilde{S}_{2,t} \subseteq S^t \cap S^c_{G^*}$ such that $ss_0\lambda^2_{t+1} \leq \sum_{i\in\tilde{S}_{2,t}}\{\mathcal{T}_{\lambda_{t+1},s_0}(H^{t+1})\}^2_i$.

Notably, our proof mainly relies on the method of mathematical induction. Assuming the results (5),(6),(7) in Theorem 4 hold for step t, we first prove that (5) and (6) hold for step $t + 1$ by induction hypothesis. We then combine the induction hypothesis with (5) and (6) for step $t + 1$ to establish (7), completing the inductive steps.

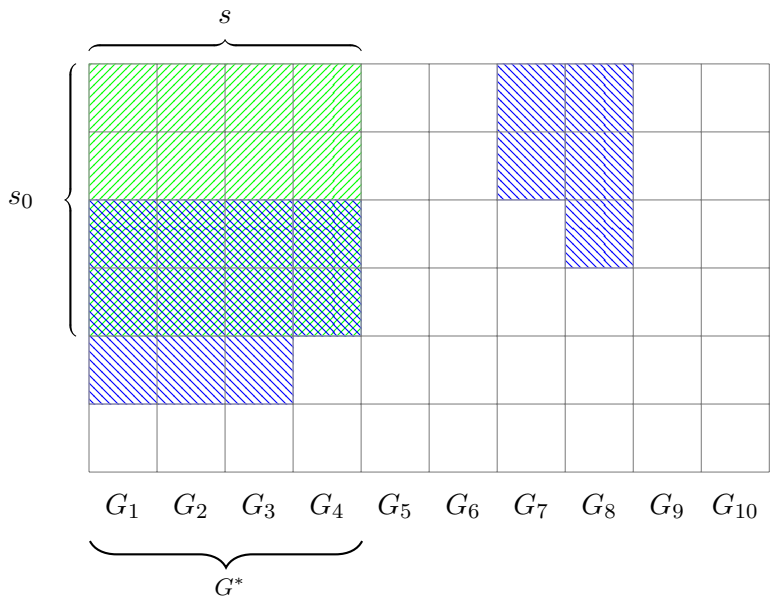

Figure 3: Illustrative example of two cases of false discovery. Here $G^* = \{G_1, G_2, G_3, G_4\}$ and $s = s_0 = 4$. The green region represents the true support set S^* and the blue region represents the selected set S^t. The remaining elements in Figure 3 are the same as in Figure 1.

Remark 6 *Here we elaborate on why we split the analysis of false discovery into two cases. Subsequently, we present an example demonstrating that in the false discovery $S^t \cap (S^*)^c$, there does not exist a subset $\tilde{S}_t$ satisfying $\tilde{S}_t \subseteq S^c_{G^*}$ such that $ss_0\lambda^2_{t+1} \leq \sum_{i\in\tilde{S}_t}\{\mathcal{T}_{\lambda_{t+1},s_0}(H^{t+1})\}^2_i$.*

In Figure 3, it is easy to verify that $S^t \cap (S^*)^c$ has 8 entries and $S^t \cap (S^*)^c \notin \mathcal{S}^{m,d}(s, s_0)$ since it covers 5 groups. By the group-wise condition checking, $\|\{\mathcal{T}_{\lambda_{t+1},s_0}(H^{t+1})\}_{G_i}\|_2^2 \geq s_0\lambda^2_{t+1}$ for $i = 7, 8$. On the other hand, inside G^*, the absolute value of each element of $S^t \cap (S^*)^c$ is not less than λ_{t+1}. However, we cannot find a (s, s_0)-shaped subset such that $ss_0\lambda^2_{t+1} \leq \sum_{i\in\tilde{S}_t}\{\mathcal{T}_{\lambda_{t+1},s_0}(H^{t+1})\}^2_i$. Therefore, we consider covering the false discovery inside G^* and outside G^* by two (s, s_0)-shaped subsets, respectively.

2.4 Minimax lower bound for double sparse linear regression

In previous works, minimax rates for the high-dimensional sparse linear regression have been studied thoroughly. A number of papers focus on element-wise s-sparsity class (Raskutti et al., 2011; Verzelen, 2012; Bellec et al., 2018), and there is also some work devoted to group sparsity such as Huang and Zhang (2010) and Lounici et al. (2011). Recently, Cai et al. (2022) provided the non-asymptotic minimax lower bounds of double sparse linear regression. Here we prove it using a more concise technique. Consider parameter space

$\widetilde{\Theta}^{m,d}(s,s_0)$:

$$\widetilde{\Theta}^{m,d}(s,s_0) \coloneqq \{\beta \in \mathbb{R}^p : \|\beta\|_{0,2} \leq s \text{ and } \|\beta_{G_j}\|_0 \leq s_0, \forall j \in [m]\}.$$

Unlike $\Theta^{m,d}(s,s_0)$, $\widetilde{\Theta}^{m,d}(s,s_0)$ imposes an ℓ_0-ball constraint on each group with a radius of s_0. Additionally, the total sparsity of $\widetilde{\Theta}^{m,d}(s,s_0)$ is limited to ss_0. It can be easily observed that $\widetilde{\Theta}^{m,d}(s,s_0) \subseteq \Theta^{m,d}(s,s_0)$. Therefore,

$$\inf_{\hat{\beta}} \sup_{\beta^* \in \Theta^{m,d}(s,s_0)} \mathbf{E}_{\hat{\beta}} \|\hat{\beta} - \beta^*\|_2^2 \geq \inf_{\hat{\beta}} \sup_{\beta^* \in \widetilde{\Theta}^{m,d}(s,s_0)} \mathbf{E}_{\hat{\beta}} \|\hat{\beta} - \beta^*\|_2^2,$$

where $\mathbf{E}_{\hat{\beta}}$ represents the expectation with respect to $\hat{\beta}$.

Definition 7 (Packing Number) *A ρ-packing of a set $\mathcal{S}$ with repsect to a metric $\|\cdot\|_\psi$ is a collection $\{\beta^1, \ldots, \beta^M\} \subset \mathcal{S}$ such that $\|\beta^i - \beta^j\|_\psi > \rho$ for all distinct $i, j \in [M]$. The ρ-packing number $M(\delta; \mathcal{S}, \|\cdot\|_\psi)$ is the cardinality of the largest ρ-packing.*

Let $M(\rho; \widetilde{\Theta}^{m,d}(s,s_0), \|\cdot\|_H)$ be the cardinality of ρ-packing set of the parameter space $\widetilde{\Theta}^{m,d}(s,s_0)$ with repsect to Hamming metric $\|\cdot\|_H$. The lower bounds for the packing number of $\widetilde{\Theta}^{m,d}(s,s_0)$ are provided as follows.

Lemma 8 (Lower bounds for the packing number (Li et al., 2024)) *The cardinality of $\frac{ss_0}{4}$-packing set of $\widetilde{\Theta}^{m,d}(s,s_0)$ is lower bounded as*

$$\log\left(M(\frac{ss_0}{4}; \widetilde{\Theta}^{m,d}(s,s_0), \|\cdot\|_H)\right) \geq \frac{ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s}}{4}.$$

Li et al. (2024) leveraged the structures of double sparsity and combined multi-ary Gilbert-Varshamov bounds (Gilbert, 1952) to construct the packing set of $\widetilde{\Theta}^{m,d}(s,s_0)$ in a more concise way. By combining Lemma 8, we establish a minimax lower bound that is consistent with the results presented in Cai et al. (2022). This is stated in the following theorem.

Theorem 9 *Consider linear regression model $y = X\beta^* + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$. Denote the maximal $(2s, 2s_0)$-sparse eigenvalue as*

$$\vartheta_{\max} = \max_{u \in \Theta^{m,d}(2s,2s_0)} \frac{\|Xu\|_2}{\sqrt{n}\|u\|_2}.$$

Assume that $\vartheta_{\max} < \infty$. Then, we have

$$\inf_{\hat{\beta}} \sup_{\beta^* \in \Theta^{m,d}(s,s_0)} \mathbf{E}_{\hat{\beta}} \|\hat{\beta} - \beta^*\|_2^2 \geq \frac{\sigma^2}{512\vartheta_{\max}^2 n}\left(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s}\right). \tag{8}$$

Theorem 9 establishes the lower bounds for the ℓ_2 estimation errors, which are consistent with the results in Cai et al. (2022). The estimation error for $\beta^{t\infty}$ matches the minimax lower bound (8), demonstrating the optimality of our IHT-style procedure.

3. A fully adaptive optimal procedure

The procedure proposed in Section 2 relies on the unknown sparsity levels s, s_0, and noise level σ, which pose a challenge in practical applications. To address this, we adopt a data-driven approach to determine the initial threshold and the optimal stopping time of our procedure, making it more feasible for real-world settings. Given s_0, we introduce a procedure that is adaptive to the unknown s and σ in Section 3.1. In Section 3.2, we explore the trade-off between classical IHT and group IHT with respect to different values of s_0. Finally, we propose a data-adaptive tuning approach for s_0 and demonstrate its optimality, rendering our method a fully adaptive procedure.

3.1 Adaptation to unknown s and σ

In the remaining part of Section 3.1, we assume that sparsity level s_0 is given. Firstly, we introduce the adaptive choice of the initial threshold λ_0. The assumption of Theorem 4 provides a lower bound for the choice of λ_0. However, choosing a significantly large value of λ_0 may decrease the efficiency of the algorithm from an optimization perspective since it can result in more redundant iterations. In the rest of our paper, denote

$$M \coloneqq \frac{1}{n}X^\top y = \beta^* + \Phi\beta^* + \Xi \quad \text{and} \quad \sigma_t^2 \coloneqq \frac{1}{n}\|y - X\beta^t\|_2^2.$$

We provide an explicit form of λ_0 as

$$\lambda_0 \coloneqq \frac{100}{9}\sqrt{\frac{\sigma_0^2}{n}\left(\log\frac{ed}{s_0} + \frac{1}{s_0}\log em\right)} \vee \frac{19}{4}\|M\|_\infty, \tag{9}$$

where $\|M\|_\infty \coloneqq \max_i |M_i|$.

Theorem 10 *Assume that β^* is (s, s_0)-sparse and X satisfies DSRIP$(2s, \frac{3}{2}s_0, \delta)$. Assume that $\delta < 0.11$ and $n > 105^2(ss_0\log\frac{ed}{s_0} + s\log em)$. Then, with probability at least $1 - \exp\{-C(ss_0\log\frac{ed}{s_0} + s\log\frac{em}{s})\}$, we have $\|\beta^*\|_2 \le ss_0\lambda_0$.*

Theorem 10 states that the choice of (9) guarantees the satisfaction of the assumption in Theorem 4 with high probability. Next, we define three stopping times t_∞, t_0 and $\bar{t}$ as follows

$$\begin{aligned}
t_\infty &\coloneqq \inf\left\{t : \lambda_t \le 4\sqrt{\frac{\sigma^2}{n}\left(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s}\right)}\right\},\\
t_0 &\coloneqq \inf\left\{t : \lambda_t \le 12\sqrt{\frac{\sigma^2}{n}\left(\log\frac{ed}{s_0} + \frac{1}{s_0}\log em\right)}\right\},\\
\bar{t} &\coloneqq \inf\left\{t : \lambda_t \le 8\sqrt{\frac{\sigma_t^2}{n}\left(\log\frac{ed}{s_0} + \frac{1}{s_0}\log em\right)}\right\}.
\end{aligned} \tag{10}$$

t_∞ is the stopping time that hits the optimal threshold λ_∞. Obviously, $\bar{t}$ is an accessible stopping time that is independent of s and σ. On the other hand, t_0 and t_∞ are the theoretical stopping time that corresponds to the unknown parameters s and σ. We state the relationship among these three stopping times in the following theorem.

Theorem 11 *Assume all the conditions in Theorem 4 hold and sample size* $n > 105^2(ss_0 \log \frac{ed}{s_0} + s \log em)$. *Then, with probability at least* $1 - \exp\left\{-C(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s})\right\}$, *we have*

$$t_0 \leq \bar{t} \leq t_\infty.$$

Theorem 11 shows that $\bar{t}$ can be bounded by the theoretical stopping times t_0 and t_∞. In particular, since $\bar{t}$ is dominated by the optimal stopping time t_∞, the estimation error $\|\beta^{\bar{t}} - \beta^*\|$ can be upper bounded by (7). Additionally, Theorem 4 implies that β^{t_0} is sub-optimal in the minimax sense. More concretely, we can deduce that $\beta^{\bar{t}}$ achieves optimal statistical accuracy up to a logarithmic factor. We state this minimax sub-optimal result as Corollary 12.

Corollary 12 *Assume the conditions in Lemma 11 hold. Then, we have*

$$\sup_{S^* \in \mathcal{S}^{m,d}(s,s_0)} P\left(\|\beta^{\bar{t}} - \beta^*\|_2 \geq 50\sqrt{\frac{\sigma^2}{n}\left(ss_0 \log \frac{ed}{s_0} + s \log em\right)}\right) \leq e^{-C(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s})}.$$

Corollary 12 is a direct consequence of Theorem 11. It demonstrates that stopping at $\bar{t}$ is a minimax sub-optimal procedure. The next open question is whether we can improve this sub-optimal procedure to be minimax optimal. The following analysis answers the question positively under certain conditions. Denote

$$\Omega(\beta) \coloneqq s_0 \|\beta\|_G \log \frac{ed}{s_0} + \|\beta\|_G \log \frac{em}{\|\beta\|_G}, \tag{11}$$

where $\|\beta\|_G \coloneqq \|\beta\|_{0,2} \vee \frac{\|\beta\|_0}{s_0}$. We consider a variant of Birgé-Massart criterion (Birgé and Massart, 2001) :

$$\tilde{t} = \arg \min_{t \in [T] \setminus [\bar{t}-1]} \left\{ \frac{1}{n} \left\| y - X\beta^t \right\|_2^2 + \frac{1000\sigma_{\bar{t}}^2 \Omega(\beta^t)}{n} \right\}, \tag{12}$$

where $T \coloneqq \inf\{t : \lambda_t \leq 4\frac{\sigma_{\bar{t}}}{\sqrt{n}}\} + 1$. Here stopping time T takes a value larger than t_∞ to ensure a sufficiently large search domain. Once the iterations hit the sub-optimal stopping time $\bar{t}$, we begin to select the optimal iteration according to (12). Now we are ready to present the detailed pseudocode of our adaptive proposed procedure in Algorithm 2.

Algorithm 2 relies on the parameter s_0 and eliminates the dependence on the unknown values of s and σ. The optimal results of stopping time $\tilde{t}$ are presented as follows.

Theorem 13 *Assume that* β^* *is* (s, s_0)*-sparse and* X *satisfies DSRIP*$(5s, s_0, \delta)$. *Assume that* $\delta < 0.11 \wedge \kappa^{10}$ *and* $n > 1000^2(ss_0 \log \frac{ed}{s_0} + s \log em)$. *Then, we have*

$$\sup_{S^* \in \mathcal{S}^{m,d}(s,s_0)} P\left(\|\beta^{\tilde{t}} - \beta^*\|_2 \geq 150\sqrt{\frac{\sigma^2}{n}\left(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s}\right)}\right) \leq e^{-C(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s})},$$

and

$$\sup_{S^* \in \mathcal{S}^{m,d}(s,s_0)} P\left(\|\beta^{\tilde{t}}\|_G \geq 4s\right) \leq e^{-C(ss_0 \log \frac{ed}{s_0} + s \log \frac{em}{s})}.$$

Algorithm 2 Double **S**parse **IHT** (DSIHT) algorithm with known s_0

Require: $X,\ y,\ \{G_j\}_{j=1}^m,\ \kappa,\ s_0$.

Initialize $t=0$, $\beta^0=0$ and $\lambda_0=\frac{100}{9}\sqrt{\frac{\sigma_0^2}{n}(\log\frac{ed}{s_0}+\frac{1}{s_0}\log em)}\vee\frac{19}{4}\|M\|_\infty$.
while $\lambda_t\geq 8\sqrt{\frac{\sigma_t^2}{n}(\log\frac{ed}{s_0}+\frac{1}{s_0}\log em)}$, **do**
  $\beta^{t+1}=\mathcal{T}_{\lambda_t,s_0}\left(\beta^t+\frac{1}{n}X^\top(y-X\beta^t)\right)$.
  $\lambda_{t+1}=\kappa\lambda_t$.
  $t=t+1$.
end while
Compute $\sigma_{\tilde{t}}^2=\frac{1}{n}\|y-X\beta^t\|_2^2$.
while $\lambda_t\geq\frac{4\sigma_{\tilde{t}}}{\sqrt{n}}$, **do**
  Compute $\mathrm{C}_t=\frac{1}{n}\left\|y-X\beta^t\right\|_2^2+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^t)}{n}$.
  $\beta^{t+1}=\mathcal{T}_{\lambda_t,s_0}\left(\beta^t+\frac{1}{n}X^\top(y-X\beta^t)\right)$.
  $\lambda_{t+1}=\kappa\lambda_t$.
  $t=t+1$.
end while
$\tilde{t}=\underset{t}{\operatorname{argmin}}\,\mathrm{C}_t$.

Ensure: $\hat{\beta}=\beta^{\tilde{t}}$.

Theorem 13 establishes the upper bound for the estimation error of $\beta^{\tilde{t}}$, indicating that $\beta^{\tilde{t}}$ adaptively achieves the minimax optimal rate of convergence. Moreover, Theorem 13 demonstrates that our procedure can guarantee the sparsity of the estimator $\beta^{\tilde{t}}$ with high probability. Specifically, we can control the model size $\|\beta^{\tilde{t}}\|_0$ within the order of $O(ss_0)$ and the selected number of groups $\|\beta^{\tilde{t}}\|_{0,2}$ within the order of $O(s)$.

Corollary 14 *Assume that all conditions in Theorem 13 hold. For the stopping time T in (12), we have*

$$\sup_{S^*\in\mathcal{S}^{m,d}(s,s_0)}P\left(T\geq\log\left(6(\frac{\sqrt{n}\|\beta^*\|_2}{\sigma}\vee\sqrt{\log ep})\right)/\log(1/\kappa)+1\right)\leq e^{-C(ss_0\log\frac{ed}{s_0}+s\log\frac{em}{s})}.$$

Corollary 14 guarantees that our IHT procedure achieves optimal statistical accuracy with linear convergence with high probability, demonstrating the efficiency of our proposed method.

3.2 Adaptive trade-off between IHT and group-IHT

In this section, we investigate the problem of misspecification of s_0, which is typically unobservable in real-world applications. Let $\bar{s}_0$ be the input parameter in Algorithm 2. Notably, given the sample (X,y) and step size κ, estimator $\hat{\beta}$ is solely determined by $\bar{s}_0$ in Algorithm 2. Therefore, we introduce the following statistical measures derived from Algorithm 2 with the given $\bar{s}_0$:

- $\hat{\beta}(\bar{s}_0)$ denotes the estimator of Algorithm 2 given $\bar{s}_0$.

- $\hat{s}(\bar{s}_0)$ denotes the selected number of groups of $\hat{\beta}(\bar{s}_0)$.
- $\hat{A}(\bar{s}_0)$ denotes the number of nonzero entries of $\hat{\beta}(\bar{s}_0)$.

By the definition of (s, s_0)-sparsity and parameter space $\mathcal{S}^{m,d}(s, s_0)$, we establish the relationship

$$\mathcal{S}^{m,d}(s, s_0) \subseteq \begin{cases} \mathcal{S}^{m,d}(ss_0/\bar{s}_0, \bar{s}_0), & \bar{s}_0 \leq s_0. \\ \mathcal{S}^{m,d}(s, \bar{s}_0), & \bar{s}_0 > s_0. \end{cases}$$

On one hand, when $\bar{s}_0 > s_0$ in Algorithm 2, the design matrix X satisfies DSRIP$(5s, \bar{s}_0, \delta)$, and β^* is $(s, \bar{s}_0)$-sparse. Algorithm 2 can obtain a minimax optimal estimator concerning parameter space $\mathcal{S}^{m,d}(s, \bar{s}_0)$, preserving all the previous theoretical results from Theorem 4 to Corollary 14. On the other hand, given $\bar{s}_0 \leq s_0$ in Algorithm 2, if the design matrix X satisfies DSRIP$(5ss_0/\bar{s}_0, \bar{s}_0, \delta)$, and β^* is $(ss_0/\bar{s}_0, \bar{s}_0)$-sparse, Algorithm 2 can obtain a minimax optimal estimator with respect to parameter space $\mathcal{S}^{m,d}(ss_0/\bar{s}_0, \bar{s}_0)$, preserving all the previous theoretical results. We summarize these results in Table 1:

Table 1: Properties for s_0-mis-specified models.

Value	**Parameter space**	**Minimax Rate**	**Support Control**
$\bar{s}_0 < s_0$	$\mathcal{S}^{m,d}(ss_0/\bar{s}_0, \bar{s}_0)$	$\sqrt{\frac{\sigma^2}{n}\left(ss_0 \log\frac{ed}{\bar{s}_0} + \frac{ss_0}{\bar{s}_0}\log\frac{em\bar{s}_0}{ss_0}\right)}$	$\hat{A}(\bar{s}_0) \lesssim ss_0$
$\bar{s}_0 = s_0$	$\mathcal{S}^{m,d}(s, s_0)$	$\sqrt{\frac{\sigma^2}{n}\left(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s}\right)}$	$\hat{A}(\bar{s}_0) \lesssim ss_0$, $\hat{s}(\bar{s}_0) \lesssim s$
$\bar{s}_0 > s_0$	$\mathcal{S}^{m,d}(s, \bar{s}_0)$	$\sqrt{\frac{\sigma^2}{n}\left(s\bar{s}_0 \log\frac{ed}{\bar{s}_0} + s\log\frac{em}{s}\right)}$	$\hat{s}(\bar{s}_0) \lesssim s$

Table 1 indicates that the theoretical properties differ significantly between the cases $\bar{s}_0 > s_0$ and $\bar{s}_0 < s_0$. When $\bar{s}_0 < s_0$, the upper bound for estimation error is given as $\sqrt{\frac{\sigma^2}{n}\left(ss_0 \log\frac{ed}{\bar{s}_0} + \frac{ss_0}{\bar{s}_0}\log\frac{em\bar{s}_0}{ss_0}\right)}$, and model size can be controlled within an order of $O(ss_0)$. In the case of $\bar{s}_0 > s_0$, the upper bound is $\sqrt{\frac{\sigma^2}{n}\left(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s}\right)}$, and the selected groups can be controlled within an order of $O(s)$. Notably, whether $\bar{s}_0 < s_0$ or $\bar{s}_0 > s_0$, simultaneous control of sparsity at both the element and group levels is unattainable.

We illustrate the minimax rate with varying values of s_0 from 1 to d in Figure 4. As depicted in Figure 4, when $1 \leq \bar{s}_0 < s_0$, the minimax rate tends to be an inversely proportional function. On the other hand, when $s_0 < \bar{s}_0 \leq d$, the minimax rate exhibits a trend of near-linear growth. Notably, for $\bar{s}_0 = s_0$, the minimax rate attains the minimum among these values.

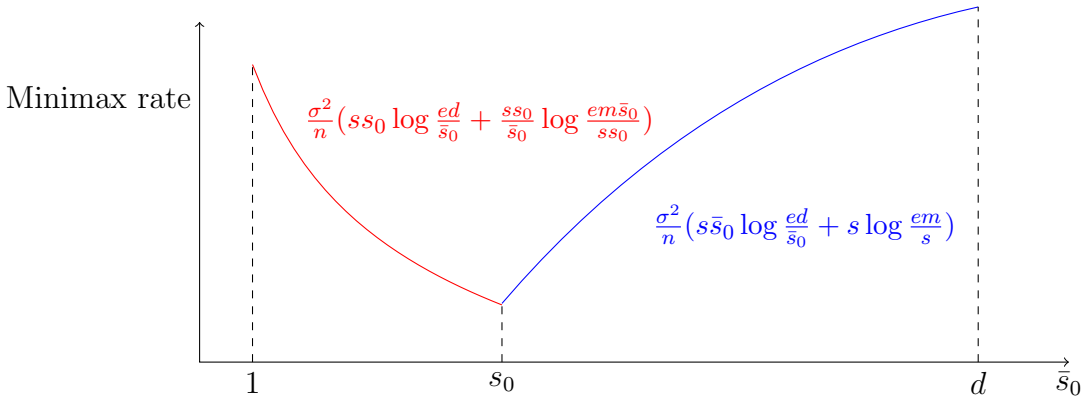

Figure 4: Minimax rate with metric $\|\cdot\|_2^2$ for different parameter spaces.

Remark 15 *Regardless of the value of $\bar{s}_0$, the above results provide the upper bound for estimation error and properties of sparsity control for Algorithm 2. In particular, when $\bar{s}_0 = 1$, the DSIHT algorithm reduces to the classical IHT algorithm (Ndaoud, 2020), and the results in Table 1 recover the minimax rate $O(\sqrt{\frac{\sigma^2}{n} ss_0 \log \frac{ep}{ss_0}})$ (Raskutti et al., 2011). When $\bar{s}_0 = d$, the results in Table 1 recover the minimax rate of group sparsity, namely, $O(\sqrt{\frac{\sigma^2}{n}(sd + s \log \frac{em}{s})})$ (Lounici et al., 2011). Therefore, DSIHT can be viewed as the trade-off between IHT (Ndaoud, 2020) and group IHT (Giraud, 2021) determined by the parameter s_0.*

3.3 Data-adaptive tuning for unknown s_0

Previous sections have introduced an adaptive procedure to address cases with unknown s and σ. In this section, we focus on constructing an adaptive estimator that achieves minimax optimality without prior knowledge of s_0, further demonstrating that our method (cf. Algorithm 3) is a fully adaptive algorithm.

Given a sequence $\{s_{0,l}\}_{l=1}^L$, an intuitive approach to determine the optimal choice involves treating s_0 as a tuning parameter. This entails running the DSIHT algorithm along the sequence and employing a model selection criterion to identify the optimal model size. Here, we utilize a variant of the Birgé-Massart criterion introduced by Verzelen (2012). This variant implicitly incorporates the knowledge of σ^2, rather than plugging in a same-order estimator of σ as demonstrated in criterion (12). Motivated by this, we propose a novel double sparse information criterion (DSIC) as follows, with $\hat{A}(\bar{s}_0)$ and $\hat{s}(\bar{s}_0)$ defined at the beginning of section 3.2:

$$\mathrm{DSIC}(\bar{s}_0) = \log\left(\frac{\|y - X\hat{\beta}(\bar{s}_0)\|_2^2}{n}\right) + \frac{K}{n}\left(\hat{A}(\bar{s}_0)\log ed + \hat{s}(\bar{s}_0)\log\frac{em}{\hat{s}(\bar{s}_0)}\right), \tag{13}$$

where K is a positive constant. The estimator $\hat{\beta}^{\bar{s}_0}$ minimizing (13) is the optimal solution of our procedure. The algorithm is summarized as follows:

Algorithm 3 **A**daptive **D**ouble **S**parse **IHT** (ADSIHT) algorithm

Require: $X,\ y,\ \{G_j\}_{j=1}^m,\ \kappa,\ \{s_{0,l}\}_{l=1}^L$.

1: **for** $l = 1, \ldots, L$, **do**
2: $\quad \hat{\beta}^l = \text{Algorithm 2}(X, y, \{G_j\}_{j=1}^m, \kappa, s_{0,l})$.
3: $\quad$ Compute the double sparse information criterion $\text{DSIC}(s_{0,l})$.
4: **end for**
5: $l^* = \underset{l \in [L]}{\text{argmin}}\{\text{DSIC}(s_{0,l})\}$.

Ensure: $\hat{\beta} = \hat{\beta}^{l^*}$.

Remark 16 *As discussed in Section 3.2, achieving optimal statistical performance necessitates that $\bar{s}_0$ is of the same order as s_0. Following the approach of Bellec et al. (2018), we set the candidate values of s_0 as an exponential sequence $\{s_{0,l}\}_{l=1}^L = \left\{2^{\frac{l-1}{2}}, 1 \le l \le L\right\}$, where $L := \max\left\{l \in \mathbb{N} : 2^{\frac{l-1}{2}} \le d\right\}$. This setting ensures that the candidate set includes a value of the same order as s_0. Recall that Cai et al. (2022) introduced candidate sets for the unknown parameters s and s_0, and employed a grid search technique for their tuning. In contrast, Algorithm 3 requires only a candidate set for s_0 with $O(\log d)$ elements, making it a much more computationally efficient tuning approach.*

Before presenting our theoretical results, we require some assumptions on the sample size and design matrix. First, we assume that there exists an interval $\mathcal{S}_0 := [s_{0,\min}, s_{0,\max}]$ such that $s_0 \in \mathcal{S}_0$.

Assumption 2 (Sample size assumption) *We assume that the sample size n satisfies $n \gtrsim \left\{\left(ss_0 \log ed + \frac{ss_0}{s_{0,\min}} \log em\right) \vee \left(ss_{0,\max} \log ed + s \log em\right)\right\}$.*

Assumption 2 is a necessary technical assumption for the minimax adaptation with an unknown noise level σ (Verzelen, 2012; Giraud et al., 2012). In addition, we require the DSRIP condition to satisfy each element of $\mathcal{S}_0$.

Assumption 3 (Adaptive DSRIP condition) *We assume that the design matrix X satisfies both DSRIP$(5s, s_{0,\max}, \delta)$ and DSRIP$(5ss_0/s_{0,\min}, s_{0,\min}, \delta)$.*

Remark 17 *In particular, when $s_{0,\min}$ is relatively small, especially for $s_{0,\min} = 1$, we observe that DSRIP$(5ss_0/s_{0,\min}, s_{0,\min}, \delta)$ reduces to the classical RIP condition (Candes and Tao, 2005). Conversely, when $s_{0,\max} = d$, DSRIP$(5s, s_{0,\max}, \delta)$ becomes the group RIP condition (Eldar and Mishali, 2009).*

Now we give the minimax adaptive result in the following theorem:

Theorem 18 *Assume that β^* is (s, s_0)-sparse. Given interval $\mathcal{S}_0$, assume that Assumption 2 and 3 hold and $\delta < 0.11 \wedge \kappa^{10}$. Let $\hat{s}_0 = \arg\min_{\bar{s}_0 \in \mathcal{S}_0}$ DSIC$(\bar{s}_0)$ with a sufficiently large K. Then, with probability greater than $1 - \exp\big\{-C_1(ss_0 \log(ed/s_0) + s\log(em/s))\big\}$, we have*

$$\left\|\hat{\beta}(\hat{s}_0) - \beta^*\right\|_2 \le C_2\sigma\sqrt{\frac{ss_0 \log ed + s\log(em/s)}{n}}. \tag{14}$$

Theorem 18 shows that our adaptive procedure, i.e., Algorithm 3, is an optimal fully adaptive procedure. Importantly, Algorithm 3 obtains the minimax adaptive solution without the knowledge of s, s_0 and σ.

Remark 19 *The significance of adapting to s_0 lies in achieving an optimal trade-off between classical IHT (Ndaoud, 2020) and group IHT (Giraud, 2021). If both Assumptions 2 and 3 are satisfied, this optimal trade-off can be attained. It is important to emphasize that when s_0 is unknown, simultaneous control of element-wise sparsity and group-wise sparsity is unattainable. Consequently, we derive near-optimal estimation error bounds for our adaptive estimator. Further details are provided in the proof of Theorem 18.*

4. Oracle estimation rate with beta-min condition

As is well-known, the ordinary least-squares (OLS) estimator supported on the true support set S^* can achieve the oracle estimation rate of $O(\sigma\sqrt{\frac{ss_0}{n}})$. In this section, under the beta-min condition, we demonstrate that the DSIHT algorithm can also attain the oracle estimation rate. This implies that the estimator obtained by DSIHT performs as well as the oracle OLS estimator. Furthermore, DSIHT exhibits almost full recovery (Butucea et al., 2018) of the true support set S^* under the beta-min condition.

Denote

$$\tilde{\lambda}_a := a\sqrt{\frac{8\sigma^2}{n}\left(\log\frac{ed}{s_0}+\frac{1}{s_0}\log\frac{em}{s}\right)},\ a>0. \tag{15}$$

Given an initial estimator $\tilde{\beta}^0$, we update the estimator by using a fixed threshold $\tilde{\lambda}_2$ in the DSIHT operator $\mathcal{T}_{\tilde{\lambda}_2,s_0}$. In specific, we update the coefficient by

$$\tilde{\beta}^{t+1}=\mathcal{T}_{\tilde{\lambda}_2,s_0}\left(\tilde{\beta}^t+\frac{1}{n}X^\top(y-X\tilde{\beta}^t)\right). \tag{16}$$

Denote $\tilde{S}^t$ as the support set of $\tilde{\beta}^t$. The following theorem investigates the theoretical guarantees of the iteration procedure with a fixed threshold.

Theorem 20 *Assume $\min_{i\in S^*}|\beta_i^*|\geq(\sqrt{2}+\epsilon)\tilde{\lambda}_2$ and $\min_{j\in G^*}\|\beta^*_{G_j}\|_2\geq(\sqrt{2}+\epsilon)\sqrt{s_0}\tilde{\lambda}_2$ for any constant $\epsilon>0$. Assume that X satisfies $DSRIP(3s,\frac{5}{3}s_0,\delta)$ and $\delta\leq\epsilon^4\wedge 0.05$. Let $\tilde{\beta}^0$ be an initial estimator satisfying* (5)-(7) *in Theorem 4. We run* (16) *and obtain the corresponding solution sequence $\{\tilde{\beta}^t\}$. Then, for $\forall t\geq 0$, as $\min\left\{\log\frac{ed}{s_0}+\frac{1}{s_0}\log\frac{em}{s},\ \frac{ss_0}{\log\frac{ed}{s_0}+\frac{1}{s_0}\log\frac{em}{s}}\right\}\to\infty$, with probability tending to 1‡, we have*

(i) $S^c_{G^*}\cap\tilde{S}^t\in\mathcal{S}^{m,d}(s,s_0)$.

(ii) $S_{G^*}\cap(S^*)^c\cap\tilde{S}^t\in\mathcal{S}^{m,d}(s,s_0)$.

‡. In specific, when $\Delta:=\frac{1}{s_0}\log(em/s)+\log(ed/s_0)$ is sufficiently large, this probability is greater than $1-C_1\exp(-C_2ss_0/\Delta)-C_3\Delta^2\exp(-C_4\Delta)$. And the tail probability of Theorem 21 is the same case.

(iii) The upper bound for estimation error satisfies

$$\left\|\tilde{\beta}^t - \beta^*\right\|_2 < \ 16\left(\frac{3}{4}\right)^t \sqrt{\frac{\sigma^2}{n}\left(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s}\right)} + 16\sqrt{\frac{\sigma^2 ss_0}{n}}. \tag{17}$$

The fixed iteration procedure preserves the results of false discoveries control, as shown in Theorem 20. Specifically, under the beta-min conditions, result (17) indicates that the upper bound for estimation error can be decomposed into two components: a diminishing optimization error $16\left(\frac{3}{4}\right)^t \sqrt{\frac{\sigma^2}{n}(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s})}$ that approaches zero as $t \to \infty$, and a statistical error $16\sqrt{\frac{\sigma^2 ss_0}{n}}$. When the optimization error becomes smaller than the statistical error, the term $O\left(\sqrt{\frac{\sigma^2 ss_0}{n}}\right)$ dominates the estimation error.

As a consequence of Theorem 20, for a sufficiently large t, the estimator $\tilde{\beta}^t$ can achieve the oracle estimation rate and almost recover the true support set at both the element and group levels. To clarify this property, we denote the element-wise decoder $\eta^* \in \{0,1\}^p$ as $\eta_i^* = \mathrm{I}(\beta_i^* \neq 0)$, and the group-wise decoder $\eta_G^* \in \{0,1\}^m$ as $(\eta_G^*)_j = \mathrm{I}(\beta_{G_j}^* \neq \mathbf{0})$. For $\tilde{\beta}^t$, denote $\tilde{\eta}^t \in \{0,1\}^p$ as $\tilde{\eta}_i^t = \mathrm{I}(\tilde{\beta}_i^t \neq 0)$, and the group-wise decoder $\tilde{\eta}_G^t \in \{0,1\}^m$ as $(\tilde{\eta}_G^t)_j = \mathrm{I}(\tilde{\beta}_{G_j}^t \neq \mathbf{0})$.

Theorem 21 *Assume that all the conditions in Theorem 20 hold. For* $\forall t > 2\log\big(256(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s})\big)$, *as* $\min\left\{\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s},\ \frac{ss_0}{\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s}}\right\} \to \infty$, *with a probability tending to 1, we have:*

(i) The estimator $\tilde{\beta}^t$ satisfies

$$\left\|\tilde{\beta}^t - \beta^*\right\|_2 \leq \ 17\sigma\sqrt{\frac{ss_0}{n}}. \tag{18}$$

(ii) The estimator $\tilde{\beta}^t$ achieves group-wise almost full recovery, that is,

$$\|\tilde{\eta}_G^t - \eta_G^*\|_0 = o\left(s\right). \tag{19}$$

(iii) The estimator $\tilde{\beta}^t$ achieves element-wise almost full recovery, that is,

$$\|\tilde{\eta}^t - \eta^*\|_0 = o\left(ss_0\right). \tag{20}$$

Theorem 21 affirms that, for a sufficiently large number of iterations, $\tilde{\beta}^t$ achieves the oracle estimation rate. Crucially, Bellec (2018) demonstrated that convex estimators cannot achieve the oracle estimation rate even when the beta-min conditions are satisfied. This highlights the superiority of our DSIHT algorithm over sparse group Lasso. Moreover, the beta-min conditions also ensure almost full recovery (Butucea et al., 2018) at both the element-wise and group-wise levels. Specifically, we can control both type-I and type-II errors within the order of $o(ss_0)$ and $o(s)$ at the element-wise and group-wise levels, respectively.

Table 2: Properties for s_0-mis-specified models under the beta-min conditions.

Value	**Parameter space**	**Order of $\tilde{\lambda}_2$**	**Oracle Estimation Rate**	**Almost Full Recovery**
$\bar{s}_0 < s_0$	$\mathcal{S}^{m,d}\left(\frac{ss_0}{\bar{s}_0}, \bar{s}_0\right)$	$\sqrt{\frac{\sigma^2}{n}\left(\log\frac{ed}{\bar{s}_0} + \frac{1}{\bar{s}_0}\log\frac{em\bar{s}_0}{ss_0}\right)}$	$\sqrt{\frac{\sigma^2}{n}ss_0}$	element-wise
$\bar{s}_0 = s_0$	$\mathcal{S}^{m,d}(s, s_0)$	$\sqrt{\frac{\sigma^2}{n}\left(\log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s}\right)}$	$\sqrt{\frac{\sigma^2}{n}ss_0}$	element-wise and group-wise
$\bar{s}_0 > s_0$	$\mathcal{S}^{m,d}(s, \bar{s}_0)$	$\sqrt{\frac{\sigma^2}{n}\left(\log\frac{ed}{\bar{s}_0} + \frac{1}{\bar{s}_0}\log\frac{em}{s}\right)}$	$\sqrt{\frac{\sigma^2}{n}s\bar{s}_0}$	group-wise

Similar to the approach in Table 1, when $\bar{s}_0 > s_0$ or $\bar{s}_0 < s_0$, we can utilize alternative parametric spaces, i.e., $\mathcal{S}^{m,d}(s, \bar{s}_0)$ or $\mathcal{S}^{m,d}(ss_0/\bar{s}_0, \bar{s}_0)$, and obtain the corresponding oracle estimation rates. These outcomes are illustrated in Table 2.

Table 2 reveals that when $\bar{s}_0 \leq s_0$, the oracle estimation rate is $\sqrt{\frac{\sigma^2}{n}ss_0}$, showing insensitivity to the variations in $\bar{s}_0$. Conversely, for $\bar{s}_0 > s_0$, the oracle estimation rate increases to $\sqrt{\frac{\sigma^2}{n}s\bar{s}_0}$, further emphasizing the role of s_0 as a trade-off between IHT and group IHT as discussed in Section 3.2.

Remark 22 *When $\bar{s}_0 = 1$, our results align with the assumptions and findings of element-wise IHT (Ndaoud, 2020). While both IHT and DSIHT attain the oracle estimation rate $O(\sqrt{\frac{\sigma^2}{n}ss_0})$ for (s, s_0)-sparse vectors with the beta-min conditions, Theorem 21 demonstrates that DSIHT not only achieves almost full recovery at the element level, as indicated by* (20)*, but also at the group level, as indicated by* (19)*. This underscores the superiority of DSIHT over IHT.*

5. Numerical experiments

In this section, we present numerical experiments that shed light on the empirical performances of our proposals using both synthetic and real-world data sets. Our algorithms are implemented in R package `ADSIHT`. We compare against several state-of-the-art methods: sparse group Lasso (SGLasso, Simon et al. (2013)), which is fitted by R package `sparsegl` (Liang et al., 2024), group bridge (GBridge, Huang et al. (2009)), group exponential Lasso (GEL, Breheny (2015)) and composite minimax concave penalty (CMCP, Breheny and Huang (2009)), which are computed by R package `grpreg` (Breheny, 2015). For SGLasso, we determine the tuning parameter by five-fold cross-validation. For the other comparison methods, we select the optimal solution using EBIC (Chen and Chen, 2008). For ADSIHT, we use our proposed DSIC with $K = 5$ to select the optimal model. Moreover, we leave the

remaining hyper-parameters to their default values in `sparsegl` and `grpreg`. All numerical experiments are conducted in R and executed on a personal laptop (AMD Ryzen 9 5900HX, 3.30 GHz, 16.00GB of RAM).

5.1 Analysis on Synthetic Data

Synthetic data sets are generated from the underlying model $y = X\beta^* + \xi$, where $\beta^* \in \mathbb{R}^p$ has m groups with equal group size, namely, $p_1 = \cdots = p_m = d$. The design matrix X is generated from a multivariate Gaussian distribution $\mathcal{MVN}(0, \Sigma)$. The covariance matrix Σ is considered as the auto-regressive structure, that is, $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$. Next, the coefficients β^* are generated under the following two scenarios:

- Homogeneous signal: β^* is randomly chosen from $\{1, -1\}$.
- Heterogeneous signal: β^* is randomly chosen from $\mathcal{N}(0, 1)$.

Finally, the random error ξ_i is generated independently from $N(0, \sigma^2)$, and σ is chosen to achieve a desired signal-to-noise ratio (SNR). All simulation results are based on 100 repetitions. Given an output $(\hat{S}, \hat{\beta})$, we use the following measures to assess the accuracy of variable selection and parameter estimation:

- **Sparsity Error (SE)**: $|\hat{S}| - |S^*|$.
- **Group-wise Sparsity Error (GSE)**: $\|\hat{\beta}\|_{0,2} - \|\beta^*\|_{0,2}$.
- **Mathew's Correlation Coefficient (MCC)**:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}},$$

  where TP= $\hat{S} \cap S^*$ and TN= $\hat{S}^c \cap (S^*)^c$ stand for true positives/negatives, respectively. FP= $\hat{S} \cap (S^*)^c$ and FN= $\hat{S}^c \cap S^*$ stand for false positives/negatives, respectively.
- **Estimation Error (EE)**: $\|\hat{\beta} - \beta^*\|_2$.

Here SE or GSE close to zero means better estimation results on the support set. MCC ranges in $[-1, 1]$, and a larger MCC means a better variable selection performance.

5.1.1 Statistical performance for varying SNR

In this section, we study the effect of varying the SNR of model on the performance of ADSIHT and other state-of-the-art methods. We consider the generating model contains 50 nonzero coefficients, distributed evenly into 10 groups. We set sample size $n = 300$, group size $d = 10$, number of group $m = 100$. The SNR increases from to 20 with an increment equal to 2. Figure 5 shows the computational results of the homogeneous scenario and heterogeneous scenario in sub-figure **A** and **B**, respectively.

Figure 5 shows that with the increase of SNR, all methods tend to perform better. Our method exhibits excellent performances in terms of all measures across the whole SNR range. For the homogeneous signal setup, our method is able to achieve full support recovery for high SNR. On the other hand, although none of the considered methods can identify all the true variables accurately even for high SNR, our method still shows its superiority in terms of variable selection and parameter estimation.

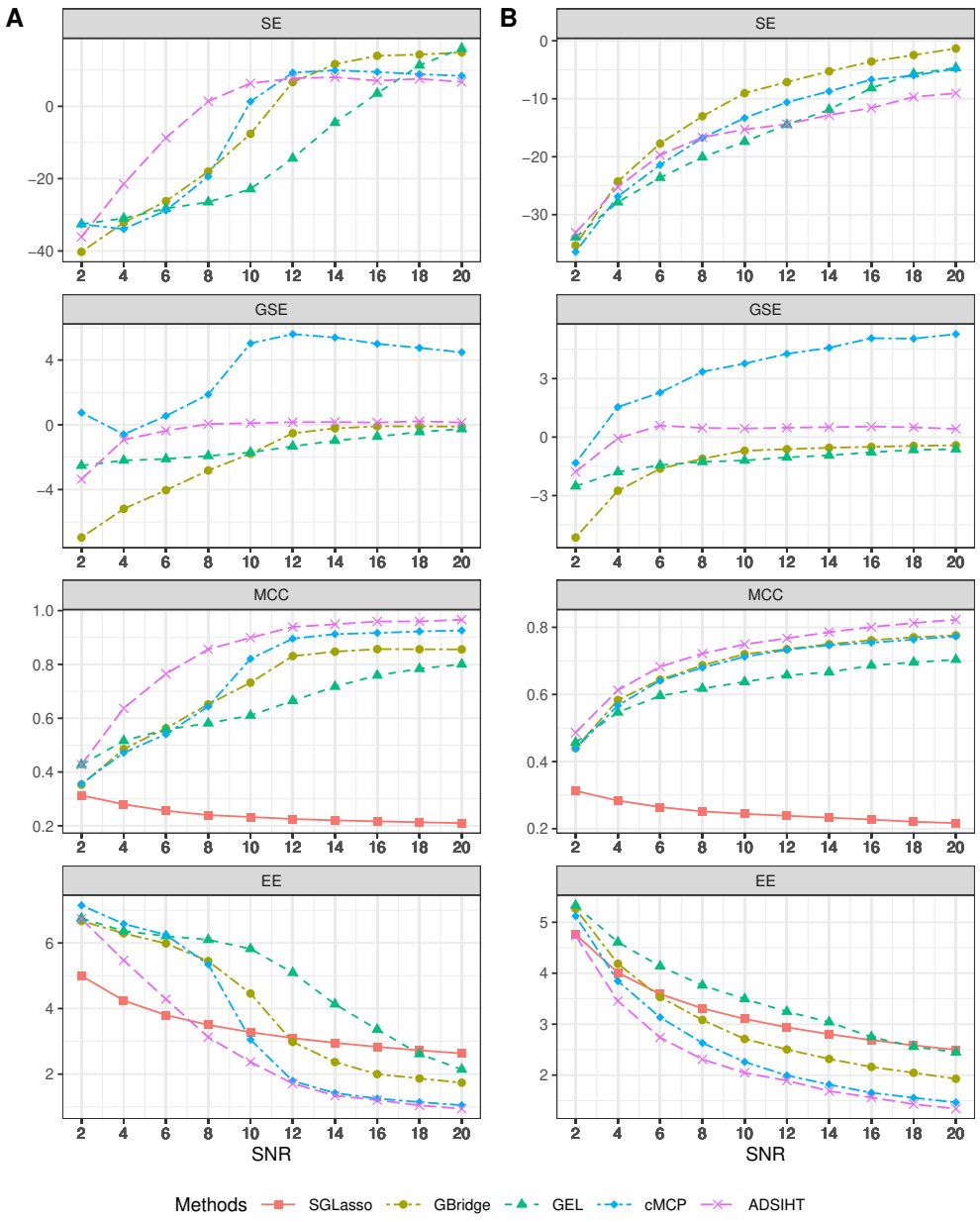

Figure 5: Performance measures as the signal-to-noise ratio (SNR) increases from 1 to 10. (A) Computational results with homogeneous signal. (B) Computational results with heterogeneous signal.

5.1.2 Statistical performance for varying number of groups

Here we study how the statistical metrics change with the number of groups. We consider the generating model contains 50 nonzero coefficients, distributed evenly into 10 groups. We set sample size $n = 500$, group size $d = 10$ and SNR $= 5$. The number of groups increases from 50 to 500 with an increment equal to 50. We show the results in figure 6.

From Figure 6, we see that our method is more robust in the high-dimensional settings. In terms of variable selection and parameter estimation, our method appears to outperform the other considered methods, with the differences being most pronounced in the high-dimensional settings. As the number of groups increases, the performances of other methods, especially for GBridge, decrease significantly.

5.1.3 Statistical performance for varying sample size

Here we investigate the effect of varying the sample size on the performances while keeping the other parameters fixed. We consider the generating model contains 50 nonzero coefficients, distributed evenly into 5 groups. We set group size $d = 20$, number of group $m = 200$ and SNR$= 5$. The sample size increases from 300 to 1000 with an increment equal to 100.

As shown in Figure 7, the performances of all methods improve significantly as the sample size increases. Our method notably outperforms the other methods across different statistical metrics. For the homogeneous signal setup, our method perfectly recovers the support set when the sample size exceeds 800. In comparison, other methods cannot achieve full support recovery even for a sufficiently large sample size. In particular, for both setups of signals, our method can estimate the coefficients accurately, which aligns with the minimax optimality of our method in the sense of parameter estimation.

5.2 Analysis on Real-world Data

The TRIM32 dataset, which pertains to the Bardet-Biedl syndrome gene expression, was initially presented by Scheetz et al. (2006) and has been extensively studied in various statistical works (Huang et al., 2010; Fan et al., 2011; Zhang et al., 2023). In this study, 120 twelve-week-old male rats were gathered for tissue harvesting from the eyes and for micro-array analysis. For this data set, TRIM32, a gene that has been associated with causing Bardet-Biedl syndrome (Chiang et al., 2006), serves as the response variable, while the remaining 18,975 gene probes that have the potential to impact TRIM32 expression are treated as covariates.

In this paper, we aim to identify the genes which are statistically significantly related to gene TRIM32 and build an accurate prediction model. Of the 18,975 probes, the top 300 probes with the highest marginal ball correlation (Pan et al., 2019) are considered. Then, for each gene, we utilize a ten-term natural cubic spline basis expansion to form a group with 10 variables. This technique, which is commonly employed in scientific research (Huang et al., 2010; Breheny and Huang, 2015; Zhang et al., 2023), allows us to analyze the data more effectively. After performing the aforementioned operations, this problem can be described as a high-dimensional variable selection problem with $n = 120$, $m = 300$, and $d = 10$. In our analysis, the 120 rats are randomly split into a training set with 100 samples and a test set with the remaining 20 samples. We repeat these random splitting procedures

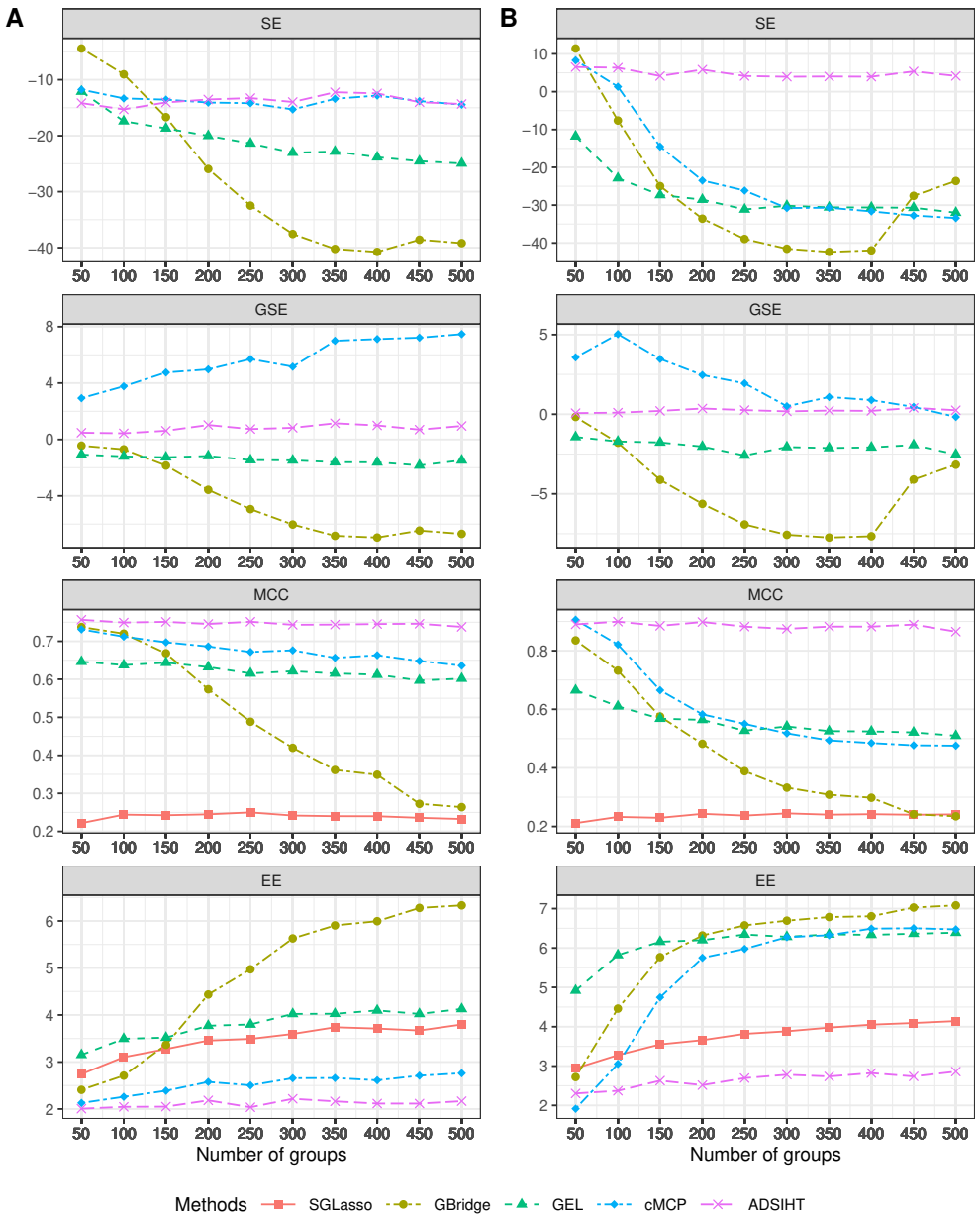

Figure 6: Performance measures as the number of groups increases from 100 to 1200. (A) Computational results with homogeneous signal. (B) Computational results with heterogeneous signal.

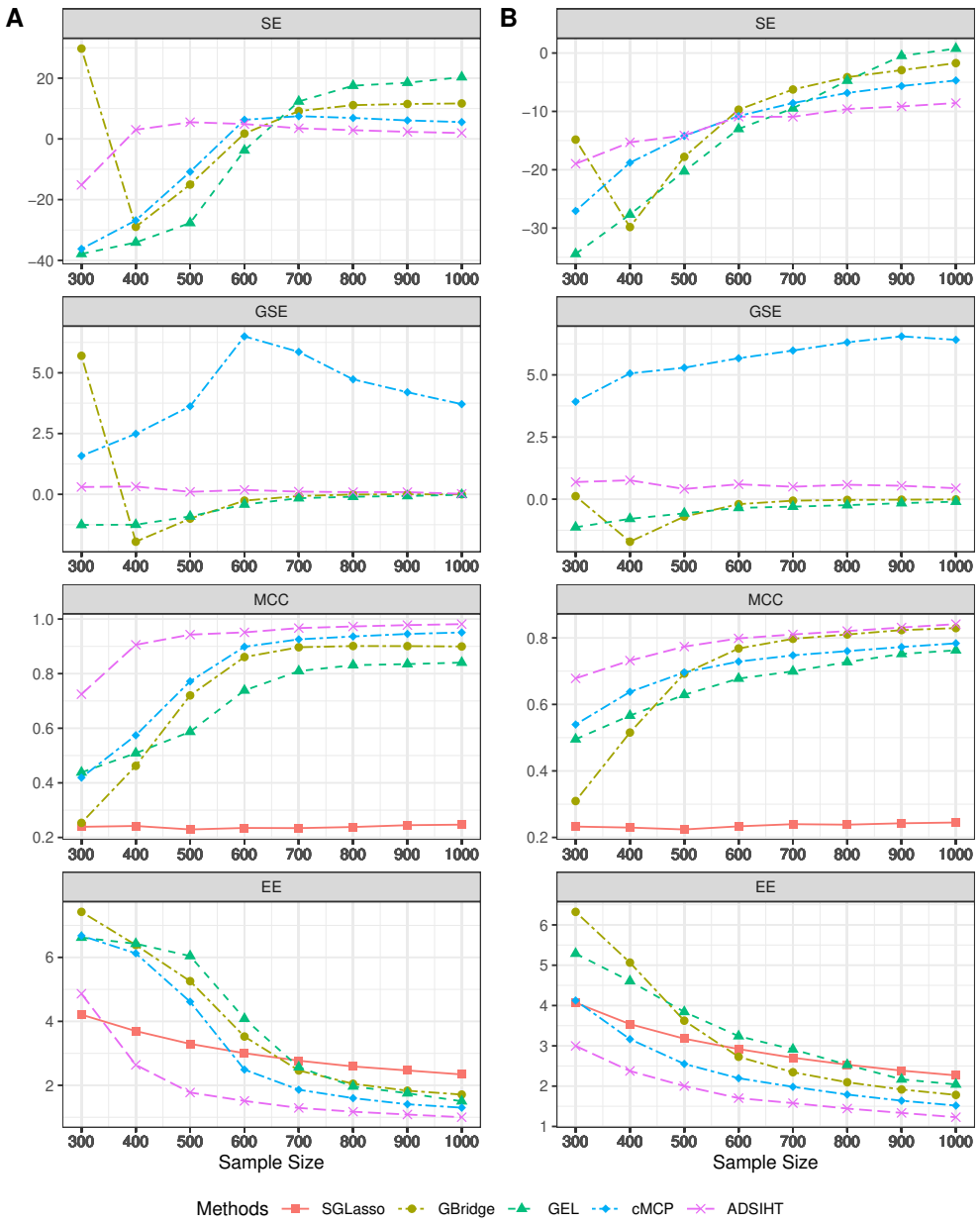

Figure 7: Performance measures as the sample size increases from 300 to 1000. (A) Computational results with homogeneous signal. (B) Computational results with heterogeneous signal.

200 times and compute the average of the numbers of selected variables and groups and the prediction mean square error (PMSE) in the test set. The computational results and the box plot of the PMSE are shown in Table 3 and Figure 8, respectively.

Table 3: Computational results for TRIM32 dataset. The standard deviations are shown in parentheses.

Method	Number of variables	Number of groups	100×PMSE
SGLasso	139.26 (68.74)	26.95 (12.45)	1.71 (1.84)
GBridge	2.95 (0.81)	1.04 (0.18)	2.01 (2.00)
GEL	35.78 (28.03)	7.07 (3.69)	2.55 (2.70)
cMCP	21.60 (3.37)	20.95 (3.08)	1.92 (2.12)
ADSIHT	29.06 (11.93)	9.20 (4.09)	**1.70 (1.85)**

Table 3 demonstrates that SGLasso identifies significantly more variables and groups than other methods. However, this does not lead to the best prediction performance on the test set. On the other hand, our proposed method delivers the highest statistical accuracy in predicting outcomes, despite using fewer variables and groups. Furthermore, Figure 8 illustrates that our approach is both accurate and robust in its predictive performance, demonstrating the superiority of our method over other methods.

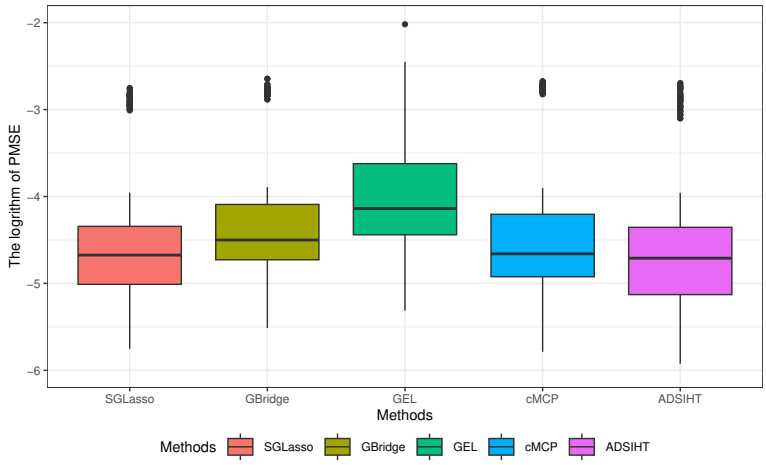

Figure 8: Boxplot of the PMSE.

To perform further investigation, we consider the entire set of 120 samples to learn a double sparse linear model for TRIM32 expression. Figure 9 displays QQ-plots of the residuals estimated from our proposed method and comparative methods. The sub-figures of cMCP and ADSIHT have points that roughly lie on the diagonal line, which indicates the satisfaction of the normality assumption. In contrast, Figure 9 reveals that the residual distributions of SGLasso, GBridge, and GEL have longer tails on the left side, which implies that analyzing this dataset using the fitted linear models may be unconvincing. Moreover,

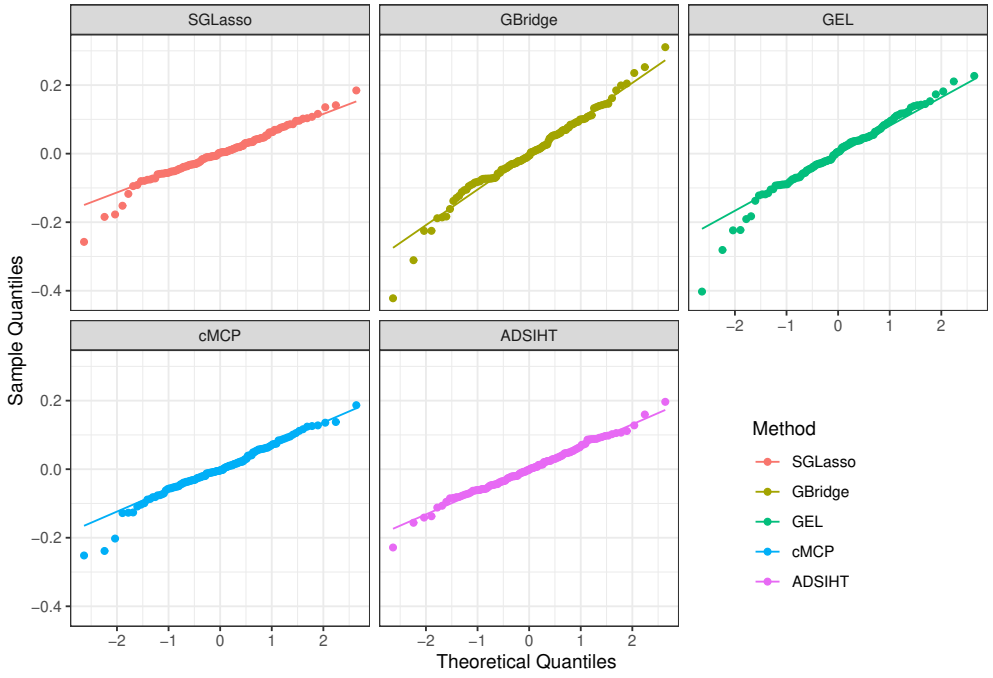

Figure 9: QQ-plots of the residuals.

we calculate the R^2 and adjusted R^2 for each method, as outlined in Table 4. The computational results in Table 4 demonstrate the favorable fitting performance of ADSIHT. Specifically, ADSIHT effectively identifies 14 important groups and 31 significant variables within these groups, collectively explaining 79% of the variance in TRIM32 expression. While SGLasso achieves the highest variance explanation in TRIM32 expression, there is a potential concern of overfitting, as it selects an excessively large model.

Table 4: The R^2 and adjusted R^2 for each method. Adjusted R^2 is omitted for SGLasso due to the excessively large model size selected by SGLasso.

	SGLasso	GBridge	GEL	cMCP	ADSIHT
R^2	88%	44%	54%	75%	79%
Adjusted R^2	×	42%	52%	68%	71%

6. Conclusion

In our work, we propose a minimax optimal IHT-style procedure for high-dimensional double sparse linear regression. In specific, we introduce a novel double sparse iterative hard thresholding (DSIHT) operator. To effectively control false discoveries, we iteratively decrease the threshold in the DSIHT operator until it reaches the optimal threshold. Un-

der certain conditions, we prove that our DSIHT algorithm obtains a minimax optimal estimator.

Notably, for the (s, s_0)-sparse structure, we devise a fully adaptive optimal procedure that enables our algorithm to derive a minimax optimal estimator with unknown sparsity levels s, s_0, and variance σ^2. Initially, given s_0, we introduce an adaptive procedure that determines the optimal stopping time using a variant of the Birgé-Massart criterion, which is independent of s and σ. Importantly, we highlight the role of sparsity level s_0 as the trade-off between IHT and group IHT. Building on this result, we propose a novel double sparse information criterion to select the optimal s_0, making our method a fully adaptive procedure. In theory, we demonstrate that our two-step adaptive procedure achieves optimal statistical accuracy with fast convergence. More importantly, to illustrate why our algorithm outperforms sparse group Lasso, we prove that under the beta-min conditions, our algorithm can attain the oracle estimation rate, which is unachievable for convex estimators, and achieve almost full recovery of the true support set. Finally, numerical experiments show that our methods exhibit more accurate and robust statistical performance than other state-of-the-art methods.

In this paper, we consider the double sparse structure in linear regression, and a similar approach can be explored in generalized linear models or single-index models. Moreover, our technical results can be applicable to various other problems with simultaneous sparsity structures, such as sparse additive models (Raskutti et al., 2012; Yuan and Zhou, 2016) and high-dimensional change point problems (Liu et al., 2021). We identify these avenues as potential future lines of research.

Acknowledgments

The authors thank the editor, action editor and anonymous referees for their many helpful comments that have resulted in significant improvements in the article. This research is supported by the Beijing Natural Science Foundation (L242104), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001) and National Key Research and Development Program of China (No. 2020YFC2004900).

Appendix

The Appendix contains the technical proofs of all Theorems and Corollaries. The proofs of the main results are presented in Appendix A. Appendix B contains the proofs of the auxiliary lemmas. Appendix C provides an example of DSRIP condition under sub-Gaussian random design. To simplify the notations of the appendix, we denote

$$\Delta \coloneqq \log\frac{ed}{s_0} + \frac{1}{s_0}\log\frac{em}{s} \quad \text{and} \quad \Delta' \coloneqq \log\frac{ed}{s_0} + \frac{1}{s_0}\log em.$$

Given a p-dimensional vector β with $\|\beta\|_0 = \hat{A}$ and $\|\beta\|_{0,2} = \hat{s}$, denote

$$\Omega^*(\beta) := (s+\hat{s})\log\frac{em}{s+\hat{s}} + (ss_0 + \hat{A})\log\frac{ed(s+\hat{s})}{ss_0+\hat{A}}.$$

Appendix A : Proofs of main results

Proof of Lemma 3

From Theorem 2.1 of Hsu et al. (2012), $\forall S \in \mathcal{S}$, we have

$$P\left(\frac{\|X_S^\top \xi\|_2^2}{\sigma^2} \geq Tr(X_S^\top X_S) + 2\|X_S^\top X_S\|_F\sqrt{t} + 2\lambda_{\max}(X_S^\top X_S)t\right) \leq e^{-t}, \tag{21}$$

where constant $t \geq 0$. Since $X_S \in \mathbb{R}^{n\times ss_0}$ and $\|X_j\|_2 = \sqrt{n}, j \in [p]$, we have

$$Tr(X_S^\top X_S) = \left(\sum_{j=1}^{ss_0}\sum_{i=1}^{n} X_{ij}^2\right) \leq ss_0 n. \tag{22}$$

On one hand, we have

$$\lambda_{\max}(X_S^\top X_S) \leq n(1+\delta). \tag{23}$$

On the other hand, from (22) and (23), we have

$$\|X_S^\top X_S\|_F = \sqrt{Tr(X_S^\top X_S X_S^\top X_S)} \leq (1+\delta)\sqrt{ss_0}n, \tag{24}$$

Substituting (22) - (24) into (21), we have

$$P\left(\frac{1}{n\sigma^2}\|X_S^\top \xi\|_2^2 \geq 2(1+\delta)\left[\sqrt{t} + \frac{\sqrt{ss_0}}{2}\right]^2 + \frac{1-\delta}{2}ss_0\right) \leq e^{-t}.$$

Note that $\delta < 1$ and $\Delta \gg 1$. For some positive constant C, let $t = (1+C)ss_0\Delta$, and we have $2(1+\delta)\left[\sqrt{t} + \frac{\sqrt{ss_0}}{2}\right]^2 + \frac{1-\delta}{2}ss_0 < 4ss_0\Delta$. Consequently, we have

$$P\left(\frac{1}{n}\|X_S^\top \xi\|_2^2 \geq 4\sigma^2 ss_0\Delta\right) \leq e^{-(1+C)ss_0\Delta}. \tag{25}$$

Note that

$$|\mathcal{S}^{m,d}(s,s_0)| \leq \binom{m}{s} \times \binom{sd}{ss_0} \leq \left(\frac{em}{s}\right)^s \times \left(\frac{ed}{s_0}\right)^{ss_0} \leq e^{ss_0\Delta}. \tag{26}$$

Therefore, combining (25) and (26), we have

$$\begin{aligned}
&P\left(\forall S \in \mathcal{S}^{m,d}(s,s_0), \sum_{i\in S}\Xi_i^2 \le \frac{4\sigma^2 s s_0 \Delta}{n}\right)\\
=&1-P\left(\exists S \in \mathcal{S}^{m,d}(s,s_0), \sum_{i\in S}\Xi_i^2 > \frac{4\sigma^2 s s_0 \Delta}{n}\right)\\
\ge&1-|\mathcal{S}^{m,d}(s,s_0)|P\left(\sum_{i\in S}\Xi_i^2 > \frac{4\sigma^2 s s_0 \Delta}{n}\right)\\
\ge&1-e^{-Css_0\Delta},
\end{aligned}$$

where the first inequality follows from the union bound. This completes the proof of Lemma 3.

Proof of Theorem 4

We proceed with the proof of Theorem 4 under the assumption that event $\mathcal{E}$ holds. Initially, it's straightforward to confirm that the results are trivial for $t = 0$. Then, we assume that the results are true for step t, and prove them for step $t + 1$.

We first prove (5) and (6) by contradiction. Assume that (5) and (6) are wrong for $t+1$, i.e., $S_{G^*} \cap S^{t+1} \cap (S^*)^c \notin \mathcal{S}^{m,d}(s,s_0)$ and $S^c_{G^*} \cap S^{t+1} \notin \mathcal{S}^{m,d}(s,s_0)$.

Step 1

For result (5), note that $S_{G^*} \cap (S^*)^c$ covers no more than s groups. According to the **Case 1** in Section 2, it holds that there exists a (s,s_0)-shaped subset $\tilde{S}_{1,t+1}$ of $S_{G^*} \cap (S^*)^c$ with cardinality ss_0 such that

$$ss_0\lambda_{t+1}^2 \le \sum_{i\in\tilde{S}_{1,t+1}} \left\{\mathcal{T}_{\lambda_{t+1,s_0}}(H^{t+1})\right\}_i^2.$$

Note that $\beta_i^* = 0$ for $i \in \tilde{S}_{1,t+1} \subseteq (S^*)^c$. Then, using equation (4) and the triangle inequality, we obtain

$$\sqrt{ss_0}\lambda_{t+1} \le \sqrt{\sum_{i\in\tilde{S}_{1,t+1}} \langle \Phi_i^\top, \beta^t - \beta^*\rangle^2} + \sqrt{\sum_{i\in\tilde{S}_{1,t+1}} \Xi_i^2}.$$

Recall that β^* is (s,s_0)-sparse, and both (5) and (6) hold for t by assumption. Then, we have $\beta^t - \beta^*$ is $(2s, \frac{3}{2}s_0)$-sparse. Consequently, using the DSRIP condition and Lemma 3,

we have

$$\begin{aligned}\sqrt{ss_0}\lambda_{t+1} \leq & \delta\|\beta^* - \beta^t\|_2 + 2\sigma\sqrt{\frac{ss_0\Delta}{n}} \\ \leq & \frac{3}{2}(1+\sqrt{2})\delta\sqrt{ss_0}\lambda_t + 2\sigma\sqrt{\frac{ss_0\Delta}{n}} \\ \leq & \frac{3}{2}(1+\sqrt{2})\delta^{\frac{9}{10}}\sqrt{ss_0}\lambda_{t+1} + \frac{1}{2}\sqrt{ss_0}\lambda_\infty \\ \leq & \left(\frac{1}{2} + \frac{3}{2}(1+\sqrt{2})\delta^{\frac{9}{10}}\right)\sqrt{ss_0}\lambda_{t+1} \\ < & \sqrt{ss_0}\lambda_{t+1},\end{aligned}$$

which leads to a contradiction. Since we have assumed that (7) holds for t, the second inequality holds based on it, and the last inequality follows from $\delta < 0.11 \wedge \kappa^{10}$. Therefore, we have $S_{G^*} \cap S^{t+1} \cap (S^*)^c \in \mathcal{S}^{m,d}(s, s_0)$, indicating that (5) holds for $t+1$.

Step 2

For result (6), if $S^c_{G^*} \cap S^{t+1}$ covers no more than s groups, the analysis of result (6) is the same as **Step 1**. Otherwise, according to **Case 2** in Section 2, there exists a (s, s_0)-shaped subset $\tilde{S}_{2,t+1}$ of $S^c_{G^*}$ such that

$$ss_0\lambda^2_{t+1} \leq \sum_{i\in\tilde{S}_{2,t+1}} \left\{\mathcal{T}_{\lambda_{t+1,s_0}}(H^{t+1})\right\}^2_i.$$

The remaining proof of (6) is similar to **Step 1**. Therefore, (6) holds for $t+1$.

Step 3

We now turn to the proof of (7). Note that results (5) and (6) hold for $t+1$, which imply that $\beta^{t+1} - \beta^*$ is $(2s, \frac{3}{2}s_0)$-sparse. Observe that for any $i \in [p]$,

$$\beta^{t+1}_i - \beta^*_i = -H^{t+1}_i \mathrm{I}(i \notin S^{t+1}) + \langle \Phi_i^\top, \beta^* - \beta^t\rangle + \Xi_i. \tag{27}$$

On one hand, summing both sides of (27) over set $S^{t+1} \cap (S^*)^c$, we have

$$\begin{aligned}\|\beta^{t+1}_{(S^*)^c}\|_2 \leq & \sqrt{\sum_{i\in S^{t+1}\cap(S^*)^c} \langle \Phi_i^\top, \beta^* - \beta^t\rangle^2} + \sqrt{\sum_{i\in S^{t+1}\cap(S^*)^c} \Xi_i^2} \\ \leq & \delta\|\beta^* - \beta^t\|_2 + 2\sigma\sqrt{\frac{2ss_0\Delta}{n}},\end{aligned} \tag{28}$$

where the right-hand side of the second inequality comes from the accumulation of two parts of random errors corresponding to (5) and (6). On the other hand, summing both sides of (27) over support set S^*, we have

$$\begin{aligned}\|(\beta^{t+1} - \beta^*)_{S^*}\|_2 \leq & \sqrt{\sum_{i\in S^*} (H^{t+1}_i)^2 \mathrm{I}(i \notin S^{t+1})} + \sqrt{\sum_{i\in S^*} \langle \Phi_i^\top, \beta^* - \beta^t\rangle^2} + \sqrt{\sum_{i\in S^*} \Xi_i^2} \\ \leq & \sqrt{2ss_0}\lambda_{t+1} + \delta\|\beta^* - \beta^t\|_2 + 2\sigma\sqrt{\frac{ss_0\Delta}{n}}.\end{aligned} \tag{29}$$

Since the procedure of operator $\mathcal{T}_{\lambda,s_0}(\cdot)$ has two steps, the term $\sum_{i\in S^*}(H_i^{t+1})^2$ in (29) is upper bounded by $2ss_0\lambda_{t+1}^2$. Combining (28) and (29), we conclude that

$$\begin{aligned}\|\beta^{t+1}-\beta^*\|_2 &\le \|\beta^{t+1}_{(S^*)^c}\|_2 + \|(\beta^{t+1}-\beta^*)_{S^*}\|_2\\ &\le \sqrt{2ss_0}\lambda_{t+1} + 2\delta\|\beta^*-\beta^t\|_2 + 2(1+\sqrt{2})\sigma\sqrt{\frac{ss_0\Delta}{n}}\\ &\le \left(\sqrt{2}+3(1+\sqrt{2})\delta^{\frac{9}{10}}+\frac{1+\sqrt{2}}{2}\right)\sqrt{ss_0}\lambda_{t+1}\\ &\le \frac{3}{2}(1+\sqrt{2})\sqrt{ss_0}\lambda_{t+1},\end{aligned}$$

where the third and the last inequalities follow from $\delta < 0.11\wedge\kappa^{10}$. We prove that (7) holds for $t+1$.

Finally, we have proved that the results in Theorem 4 hold for $t+1$ under the induction hypothesis. This completes the proof of Theorem 4.

Proof of Theorem 9

Consider the $\frac{ss_0}{4}$-packing set $\{\beta^1,\ldots,\beta^M\}$, where M is the shorthand for the packing number $M(\frac{ss_0}{4};\widetilde{\Theta}^{m,d}(s,s_0),\|\cdot\|_H)$. We set all the non-zero elements of $\beta\in\{\beta^1,\ldots,\beta^M\}$ equal to δ, where δ is a parameter that need to be determined below. For any $\beta^i\neq\beta^j$, since each β^i has at most ss_0 nonzero elements, we have

$$\|\beta^i-\beta^j\|_2^2 \le 2ss_0\delta^2,\ \forall i,j\in[M]. \tag{30}$$

On the other hand, since $\{\beta^1,\ldots,\beta^M\}$ is a $\frac{ss_0}{4}$-packing set of $\widetilde{\Theta}^{m,d}(s,s_0)$, we have

$$\|\beta^i-\beta^j\|_2^2 \ge \frac{1}{4}ss_0\delta^2,\ \forall i,j\in[M]. \tag{31}$$

Given design matrix X, denote $y^i = X\beta^i+\xi, \forall i\in[M]$. We consider the Kullback-Leibler divergence between different distribution pairs as

$$\begin{aligned}KL\left(y^i||y^j\right) &= \frac{1}{2\sigma^2}\|X(\beta^i-\beta^j)\|_2^2\\ &\le \frac{n\vartheta_{\max}^2}{2\sigma^2}\|\beta^i-\beta^j\|_2^2,\end{aligned}$$

where the last inequality follows from the eigenvalue value condition of X and $\beta^i-\beta^j\in\widetilde{\Theta}^{m,d}(2s,2s_0)$. Denote B as the random vector uniformly distributed over the packing set. Observe that

$$I(y;B) \le \frac{1}{\binom{M}{2}}\sum_{i\neq j}KL(y^i||y^j) \tag{32}$$

$$\le \frac{1}{\binom{M}{2}}\sum_{i\neq j}\frac{n\vartheta_{\max}^2}{2\sigma^2}\|\beta^i-\beta^j\|_2^2 \tag{33}$$

$$\le \frac{n\vartheta_{\max}^2}{\sigma^2}ss_0\delta^2, \tag{34}$$

where the last inequality uses (30). Combining the generalized Fano's Lemma (Cover and Thomas, 2006) and (32), we have

$$\begin{aligned} P(B \neq \widetilde{\beta}) &\geq 1 - \frac{I(y;B) + \log 2}{\log M} \\ &\geq 1 - \frac{\frac{n\vartheta_{\max}^2}{\sigma^2} ss_0\delta^2 + \log 2}{\log M}, \end{aligned}$$

where $\widetilde{\beta}$ takes value in the packing set. To guarantee $P(B \neq \widetilde{\beta}) \geq \frac{1}{2}$, it suffices to choose $\delta = \frac{1}{2}\sqrt{\frac{\sigma^2 \log M}{n\vartheta_{\max}^2 ss_0}}$. Substituting it into equation (31) and from Lemma 8, we have

$$\begin{aligned} &\inf_{\hat{\beta}} \sup_{\beta^* \in \Theta^{m,d}(s,s_0)} \mathbf{E}_{\hat{\beta}} \|\hat{\beta} - \beta^*\|_2^2 \\ \geq &\frac{1}{16} ss_0\delta^2 \cdot \inf_B P\left(B \neq \tilde{\beta}\right) \\ \geq &\frac{\sigma^2 \log M}{128 n\vartheta_{\max}^2} \\ \geq &\frac{\sigma^2}{512 n\vartheta_{\max}^2}\left(ss_0 \log\frac{ed}{s_0} + s\log\frac{em}{s}\right), \end{aligned}$$

which completes the proof of Theorem 9.

Proof of Theorem 10

Using Lemma 23, with probability at least $1 - \exp\{-Css_0\Delta\}$, we have

$$\sigma_0 \geq \frac{19}{20}\sigma - \sqrt{1+\delta}\|\beta^*\|_2. \tag{35}$$

With probability at least $1 - \exp\{-Css_0\Delta\}$, we have

$$\begin{aligned} \|M_{S^*}\|_2 &= \|\left(\beta^* + \Phi\beta^* + \Xi\right)_{S^*}\|_2 \\ &\geq \|\beta^*\|_2 - \|\Phi\beta^*\|_2 - \|\Xi_{S^*}\|_2 \\ &\geq (1-\delta)\|\beta^*\|_2 - 2\sigma\sqrt{\frac{ss_0\Delta}{n}}. \end{aligned} \tag{36}$$

Note that $\sqrt{ss_0}\|M\|_\infty \geq \|M_{S^*}\|_2$. Combining (35) and (36), with probability at least $1 - \exp\{-Css_0\Delta\}$, we have

$$\begin{aligned} \sqrt{ss_0}\lambda_0 &\geq \frac{100}{9}\sigma_0\sqrt{\frac{ss_0\Delta'}{n}} \vee \frac{19}{4}\sqrt{ss_0}\|M\|_\infty \\ &\geq \frac{9}{19}\times\frac{100}{9}\sigma_0\sqrt{\frac{ss_0\Delta'}{n}} + \frac{10}{19}\times\frac{19}{4}\|M_{S^*}\|_2 \\ &\geq \frac{100}{19}\left(\frac{19}{20}\sigma - \sqrt{1+\delta}\|\beta^*\|_2\right)\sqrt{\frac{ss_0\Delta'}{n}} + \frac{5}{2}\left((1-\delta)\|\beta^*\|_2 - 2\sigma\sqrt{\frac{ss_0\Delta}{n}}\right) \\ &\geq \left(\frac{5}{2}(1-\delta) - \frac{100}{19}\sqrt{1+\delta}\sqrt{\frac{ss_0\Delta'}{n}}\right)\|\beta^*\|_2 \\ &\geq \|\beta^*\|_2 \end{aligned}$$

where the fourth inequality uses the fact that $\Delta' \geq \Delta$, and the last inequality uses the fact that $n > 105^2 ss_0\Delta'$ and $\delta < 0.11$. We complete the proof of Theorem 10.

Proof of Theorem 11

Note that $t_0 \leq t_\infty$ holds since $\Delta' \geq \Delta$. We first claim that $t_0 \geq \bar{t}$. For any $t \leq t_0$, according to the definition of t_0, we have

$$\sigma\sqrt{\frac{\Delta'}{n}} \leq \frac{1}{12}\lambda_t. \tag{37}$$

From Lemma 23 and Theorem 4, with probability at least $1 - \exp\{-Css_0\Delta\}$, we have

$$\begin{aligned} \sigma_t &\leq \sqrt{1+\delta}\|\beta^* - \beta^t\|_2 + \frac{21}{20}\sigma \\ &\leq \sqrt{1+\delta}\frac{3}{2}(1+\sqrt{2})\sqrt{ss_0}\lambda_t + \frac{21}{20}\sigma. \end{aligned} \tag{38}$$

From (38), it comes out that

$$\begin{aligned} 8\sigma_t\sqrt{\frac{\Delta'}{n}} &\leq 12(1+\sqrt{2})\sqrt{1+\delta}\lambda_t\sqrt{\frac{ss_0\Delta'}{n}} + \frac{42}{5}\sigma\sqrt{\frac{\Delta'}{n}} \\ &\leq 12(1+\sqrt{2})\sqrt{1+\delta}\lambda_t\sqrt{\frac{ss_0\Delta'}{n}} + \frac{7}{10}\lambda_t \\ &\leq \lambda_t, \end{aligned} \tag{39}$$

where the first inequality uses (37), and the second inequality follows from $\delta < 0.11$ and $n > 105^2 ss_0\Delta'$. (39) leads to the fact that $t \leq \bar{t}$, which deduces that $t_0 \leq \bar{t}$ holds with high probability.

Next, we turn to the proof of $\bar{t} \leq t_\infty$. Since $t_0 \leq t_\infty$, Theorem 4 shows us that

$$\|\beta^{t_0} - \beta^*\|_2 \leq 18(1+\sqrt{2})\sigma\sqrt{\frac{ss_0\Delta'}{n}}. \tag{40}$$

From Lemma 23, for any $t_0 \leq t \leq t_\infty$, it holds with probability at least $1 - \exp\{-Css_0\Delta\}$ that

$$\begin{aligned} |\sigma_t - \sigma| &\leq \sqrt{1+\delta}\|\beta^* - \beta^t\|_2 + \frac{1}{20}\sigma \\ &\leq 18(1+\sqrt{2})\sqrt{1+\delta}\sigma\sqrt{\frac{ss_0\Delta'}{n}} + \frac{1}{20}\sigma \\ &\leq (\frac{9}{20} + \frac{1}{20})\sigma = \frac{1}{2}\sigma, \end{aligned}$$

where the second inequality follows from (40), and the last inequality follows from $\delta < 0.11$ and $n > 105^2 ss_0\Delta'$. Combining the above inequalities, we have

$$\frac{8\sigma_t}{\sqrt{n}}\sqrt{\Delta'} \geq \frac{4\sigma}{\sqrt{n}}\sqrt{\Delta'} \geq \frac{4\sigma}{\sqrt{n}}\sqrt{\Delta}.$$

This result implies that $\bar{t} \leq t_\infty$, which completes the proof of Theorem 11.

Proof of Theorem 13

From Lemma 23 and (40), with probability at least $1-\exp\{-Css_0\Delta\}$, we have

$$\begin{aligned}|\sigma_{\bar{t}}-\sigma| &\le \sqrt{1+\delta}\|\beta^{\bar{t}}-\beta^*\|_2+\frac{1}{20}\sigma\\ &\le \sigma\left(18(1+\sqrt{2})\sqrt{1+\delta}\sqrt{\frac{ss_0\Delta'}{n}}+\frac{1}{20}\right)\\ &\le \frac{1}{10}\sigma,\end{aligned} \tag{41}$$

where the last inequality uses $\delta<0.11$ and $n>1000^2ss_0\Delta'$.

First, we prove $\|\beta^{\tilde{t}}\|_G\le 4s$ by contradiction. Let us assume that $\|\beta^{\tilde{t}}\|_G>4s$. According to the definition of $\tilde{t}$, we have

$$\frac{1}{n}\left\|y-X\beta^{\tilde{t}}\right\|_2^2+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^{\tilde{t}})}{n}\le\frac{1}{n}\left\|y-X\beta^{t_\infty}\right\|_2^2+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^{t_\infty})}{n}. \tag{42}$$

On one hand, we have

$$\begin{aligned}\left\|y-X\beta^{\tilde{t}}\right\|_2^2 &\ge \|\xi\|_2^2+\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2^2-2\left|\left\langle\xi,X(\beta^{\tilde{t}}-\beta^*)\right\rangle\right|\\ &\ge \|\xi\|_2^2+\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2^2-2\sigma\sqrt{3\Omega^*(\beta^{\tilde{t}})}\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2\\ &\ge \|\xi\|_2^2+\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2^2-2\sigma\sqrt{3\times\frac{5}{4}\Omega(\beta^{\tilde{t}})}\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2\\ &\ge \|\xi\|^2-\frac{15}{4}\sigma^2\Omega(\beta^{\tilde{t}}),\end{aligned}$$

where the second inequality follows from Lemma 24, and the third inequality uses the fact that $\frac{5}{4}\Omega(\beta^{\tilde{t}})\ge\Omega^*(\beta^{\tilde{t}})$ when $\|\beta^{\tilde{t}}\|_G>4s$. The definition of $\Omega^*(\beta)$ is given at the beginning of the Appendix. By some simple algebras, it comes out that

$$\begin{aligned}\frac{1}{n}\left\|y-X\beta^{\tilde{t}}\right\|_2^2+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^{\tilde{t}})}{n} &\ge \frac{\|\xi\|_2^2}{n}-\frac{15\sigma^2\Omega(\beta^{\tilde{t}})}{4n}+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^{\tilde{t}})}{n}\\ &\ge \frac{\|\xi\|_2^2}{n}+\frac{950\sigma_{\tilde{t}}^2\Omega(\beta^{\tilde{t}})}{n}.\end{aligned} \tag{43}$$

On the other hand, we have

$$\begin{aligned}\left\|y-X\beta^{t_\infty}\right\|_2^2 &\le \|\xi\|_2^2+\left\|X(\beta^{t_\infty}-\beta^*)\right\|_2^2+2\left|\langle\xi,X(\beta^{t_\infty}-\beta^*)\rangle\right|\\ &\le \|\xi\|_2^2+\left\|X(\beta^{t_\infty}-\beta^*)\right\|_2^2+6\sigma\sqrt{\Omega(\beta^*)}\left\|X(\beta^{t_\infty}-\beta^*)\right\|_2\\ &\le \|\xi\|_2^2+2\left\|X(\beta^{t_\infty}-\beta^*)\right\|_2^2+9\sigma^2\Omega(\beta^*),\end{aligned}$$

where the second inequality follows from (89) in Lemma 24 since $\beta^{t_\infty}-\beta^*$ is $(2s,\frac{3}{2}s_0)$-sparse. By some simple algebras, it comes out that

$$\begin{aligned}\frac{1}{n}\left\|y-X\beta^{t_\infty}\right\|_2^2+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^{t_\infty})}{n}\leq&\frac{\|\xi\|_2^2}{n}+\frac{2}{n}\left\|X(\beta^{t_\infty}-\beta^*)\right\|_2^2+\frac{9\sigma^2\Omega(\beta^*)}{n}\\&+\frac{1000\sigma_{\tilde{t}}^2\Omega(\beta^{t_\infty})}{n}\\\leq&\frac{\|\xi\|_2^2}{n}+2(1+\delta)\left(\frac{3}{2}(1+\sqrt{2})\right)^2\times\frac{16\sigma^2ss_0\Delta}{n}\\&+\frac{9\sigma^2ss_0\Delta}{n}+\frac{1000\sigma_{\tilde{t}}^2}{n}\left(2s\log\frac{em}{2s}+3ss_0\log\frac{ed}{s_0}\right)\\\leq&\frac{\|\xi\|_2^2}{n}+\frac{480\sigma^2ss_0\Delta}{n}+\frac{3000\sigma_{\tilde{t}}^2ss_0\Delta}{n}\\\leq&\frac{\|\xi\|_2^2}{n}+\frac{3600\sigma_{\tilde{t}}^2ss_0\Delta}{n},\end{aligned}\tag{44}$$

where the third inequality holds for $\delta<\frac{1}{10}$ and the last inequality follows from (41). Combining (42)-(44), from the triangle relationship, we have

$$950\Omega(\beta^{\tilde{t}})\leq 3600ss_0\Delta.$$

Recall $\Omega(\beta^{\tilde{t}})>4ss_0\Delta$ under the assumption $\|\beta^{\tilde{t}}\|_G>4s$. Then, we have

$$3800ss_0\Delta<950\Omega(\beta^{\tilde{t}})\leq 3600ss_0\Delta.$$

which contradicts the assumption of $\|\beta^{\tilde{t}}\|_G>4s$. Therefore, we must have $\|\beta^{\tilde{t}}\|_G\leq 4s$.

Next, we show the upper bounds for the estimation error $\|\beta^{\tilde{t}}-\beta^*\|$. From (42), we have

$$\begin{aligned}\left\|y-X\beta^{t_\infty}\right\|_2^2+1000\sigma_{\tilde{t}}^2\Omega(\beta^{t_\infty})\geq&\left\|y-X\beta^{\tilde{t}}\right\|_2^2+1000\sigma_{\tilde{t}}^2\Omega(\beta^{\tilde{t}})\\\geq&\left\|y-X\beta^{\tilde{t}}\right\|_2^2\\\geq&\|\xi\|_2^2+\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2^2-2\sigma\sqrt{3\Omega^*(\beta^{\tilde{t}})}\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2.\end{aligned}$$

Combining the above inequalities and (44), we have

$$\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2^2-2\sigma\sqrt{3\Omega^*(\beta^{\tilde{t}})}\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2\leq 3600\sigma_{\tilde{t}}^2ss_0\Delta.\tag{45}$$

By solving the quadratic inequalities (45), we have

$$\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2\leq 140\sigma\sqrt{ss_0\Delta}.$$

Note that $\beta^{\tilde{t}}$ is $(4s,s_0)$-sparse. Then, we conclude that by DSRIP condition, we have

$$\left\|\beta^{\tilde{t}}-\beta^*\right\|_2\leq\frac{\left\|X(\beta^{\tilde{t}}-\beta^*)\right\|_2}{\sqrt{n(1-\delta)}}\leq 150\sigma\sqrt{\frac{ss_0\Delta}{n}},$$

which completes the proof of Theorem 13.

Proof of Corollary 14

Note that with probability at least $1 - p^{-C}$,

$$\begin{aligned} \|M\|_\infty \leq & \|\beta^* + \Phi\beta^*\|_\infty + \|\Xi\|_\infty \\ \leq & \|\beta^* + \Phi\beta^*\|_2 + \|\Xi\|_\infty \\ \leq & (1+\delta)\|\beta^*\|_2 + 2\sigma\sqrt{\frac{\log ep}{n}} \\ \leq & 4\left(\|\beta^*\|_2 \vee \sigma\sqrt{\frac{\log ep}{n}}\right), \end{aligned} \tag{46}$$

where the last inequality uses $\delta \leq 1$. Substituting (46) into the definition of λ_0, we have

$$\begin{aligned} \lambda_0 = & \frac{100}{9}\sigma_0\sqrt{\frac{\Delta'}{n}} \vee \frac{19}{4}\|M\|_\infty \\ \leq & 19\left(\|\beta^*\|_2 \vee \sigma\sqrt{\frac{\log ep}{n}}\right), \end{aligned} \tag{47}$$

where the last inequality uses $\Delta' \leq \log(ep)$. Observe that $\kappa^T\lambda_0 \leq 4\frac{\sigma_{\bar{t}}}{\sqrt{n}}$. By some simple algebras, with probability at least $1 - \exp\{-Css_0\Delta\}$, we have

$$\begin{aligned} T \leq & \log\left(\frac{\lambda_0\sqrt{n}}{4\sigma_{\bar{t}}}\right) / \log(\frac{1}{\kappa}) \\ \leq & \log\left(\frac{5\lambda_0\sqrt{n}}{18\sigma}\right) / \log(\frac{1}{\kappa}) \\ \leq & \log\left(6(\frac{\sqrt{n}\|\beta^*\|_2}{\sigma} \vee \sqrt{\log ep})\right) / \log(\frac{1}{\kappa}), \end{aligned}$$

where the second inequality follows from 41 and the last inequality uses (47).

Therefore,

$$\sup_{S^* \in \mathcal{S}^{m,d}(s,s_0)} P\left(T \geq \log\left(6(\frac{\sqrt{n}\|\beta^*\|_2}{\sigma} \vee \sqrt{\log ep})\right) / \log(\frac{1}{\kappa}) + 1\right) \leq e^{-Css_0\Delta}.$$

Proof of Theorem 18

Our technique for tuning s_0 is notably distinct and more complex than that of Verzelen (2012). As discussed in Section 3.2, the theoretical properties differ significantly between the cases $\bar{s}_0 > s_0$ and $\bar{s}_0 < s_0$. We can control the sparsity at either the element-wise or group-wise level, but not both simultaneously. Additionally, as illustrated in Figure 4, the minimax rates for different values of $\bar{s}_0$ exhibit a"U-shaped" curve, rather than the monotonically increasing trend observed under element-wise sparsity (Raskutti et al., 2011; Verzelen, 2012). Therefore, we must separately analyze the cases $\bar{s}_0 > s_0$ and $\bar{s}_0 < s_0$.

Basic inequality of Verzelen's procedure

In this part, we give the basic comparable inequality used in the proof. This part is similar to Theorem 5.2 in Verzelen (2012). Denote

$$\mathrm{pen}\left(\bar{s}_0\right)=\frac{K}{n}\left(\hat{A}(\bar{s}_0)\log ed+\hat{s}(\bar{s}_0)\log\frac{em}{\hat{s}(\bar{s}_0)}\right),\quad \mathrm{pen}\,'\left(\bar{s}_0\right)=-1+\exp\left(\mathrm{pen}\left(\bar{s}_0\right)\right).$$

By the definition of $\hat{s}_0$, we have

$$\frac{1}{n}\left\|y-X\hat{\beta}\left(\hat{s}_0\right)\right\|_2^2\cdot\left(1+\mathrm{pen}'(\hat{s}_0)\right)\leq\frac{1}{n}\left\|y-X\hat{\beta}\left(s_0\right)\right\|_2^2\cdot\left(1+\mathrm{pen}\,'\left(s_0\right)\right). \tag{48}$$

For the right-hand side of (48), a strategy similar to (44) leads that

$$\frac{1}{n}\left\|y-X\hat{\beta}\left(s_0\right)\right\|_2^2\leq\frac{1}{n}\|\xi\|_2^2+C_1\frac{\sigma^2 ss_0\Delta}{n}. \tag{49}$$

Recall that we assume n is large enough so that

$$\mathrm{pen}(s_0)\leq\frac{4K}{n}\left(ss_0\log ed+s\log(em/s)\right)<0.1,$$

which implies that $1+\mathrm{pen}'\left(s_0\right)=\exp\left\{\mathrm{pen}\left(s_0\right)\right\}\leqslant e$. Combining (48) and (49), we have

$$\frac{1}{n}\left\|y-X\hat{\beta}\left(\hat{s}_0\right)\right\|_2^2\cdot\left(1+\mathrm{pen}\,'\left(\hat{s}_0\right)\right)\leq\frac{1}{n}\|\xi\|_2^2\left(1+\mathrm{pen}'\left(s_0\right)\right)+C_2\frac{\sigma^2 ss_0\Delta}{n}. \tag{50}$$

For the left-hand side of (48), a strategy similar to (43) leads that

$$\begin{aligned}\frac{1}{n}\left\|y-X\hat{\beta}\left(\hat{s}_0\right)\right\|_2^2\geq\frac{1}{n}\|\xi\|_2^2+\frac{1}{n}\left\|X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)\right\|_2^2\\ -\frac{2}{n}\left\|X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)\right\|_2\cdot\left|\left\langle\xi,\frac{X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)}{\left\|X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)\right\|_2}\right\rangle\right|.\end{aligned} \tag{51}$$

Then, we can also upper bound the inner product by Lemma 24 as

$$\begin{aligned}\left|\left\langle\frac{\xi}{\sigma},\frac{X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)}{\left\|X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)\right\|_2}\right\rangle\right|^2&\precsim\left(ss_0+\hat{A}(\hat{s}_0)\right)\log\frac{ed(s+\hat{s}(\hat{s}_0))}{ss_0+\hat{A}(\hat{s}_0)}+(s+\hat{s}(\hat{s}_0))\log\frac{em}{s+\hat{s}(\hat{s}_0)}\\&\leq\left(ss_0+\hat{A}(\hat{s}_0)\right)\log ed+(s+\hat{s}(\hat{s}_0))\log\frac{em}{s+\hat{s}(\hat{s}_0)}.\end{aligned} \tag{52}$$

Denote $\mathcal{L}:=\left\|X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)\right\|_2$ and $\hat{\Gamma}:=\left(ss_0+\hat{A}(\hat{s}_0)\right)\log ed+(s+\hat{s}(\hat{s}_0))\log\frac{em}{s+\hat{s}(\hat{s}_0)}$. Therefore, by (48)-(52), we obtain

$$\mathcal{L}^2-C_2'\sigma\sqrt{\hat{\Gamma}}\mathcal{L}\leq\mathrm{pen}'(s_0)\|\xi\|_2^2+C_2\sigma^2 ss_0\Delta. \tag{53}$$

By Lemma 1 of Laurent and Massart (2000), we conclude that $0.9\sigma^2 \leq \frac{1}{n}\|\xi\|_2^2 \leq 1.1\sigma^2$ holds with probability at least $1-\exp(-C_4 n)$. Besides, by $\mathrm{pen}(s_0) < 0.1$, we derive that $\mathrm{pen}'(s_0) = \exp(\mathrm{pen}(s_0)) - 1 \leq 2\,\mathrm{pen}(s_0)$, therefore

$$\mathrm{pen}'(s_0)\|\xi\|_2^2 + C_2\sigma^2 s s_0 \Delta \leq C_3\sigma^2\big(s s_0 \log d + s\log(em/s)\big) \leq C_3\sigma^2\hat{\Gamma},$$

which leads to

$$\mathcal{L}^2 - C_2'\sigma\sqrt{\hat{\Gamma}}\mathcal{L} \leq C_3\sigma^2\hat{\Gamma}. \tag{54}$$

By solving inequality (54), we obtain the upper bound $\mathcal{L}^2 \leq C_4\sigma^2\hat{\Gamma}$. Therefore, to get the optimal upper bound for estimation error, we just need to prove that

$$\hat{\Gamma} \precsim s s_0 \log ed + s\log(em/s). \tag{55}$$

By far, based on table 1 we know that $\hat{A}(\hat{s}_0) \leq 4 s s_0$ for $\hat{s}_0 \leq s_0$, and $\hat{s}(\hat{s}_0) \leq 4s$ for $\hat{s}_0 \geq s_0$. Therefore, with high probability, we conclude that

$$\hat{\Gamma} \leq \begin{cases} 5 s s_0 \log ed + (s + \hat{s}(\hat{s}_0))\log\frac{em}{s+\hat{s}(\hat{s}_0)}, & \hat{s}_0 \leq s_0. \\ \Big(s s_0 + \hat{A}(\hat{s}_0)\Big)\log ed + 5s\log\frac{em}{s}, & \hat{s}_0 > s_0. \end{cases} \tag{56}$$

For convenience, we divide the next proof into two cases: $s s_0 \log ed \leq s\log\frac{em}{s}$ or $s s_0 \log ed \geq s\log\frac{em}{s}$.

Assumption A: $s s_0 \log ed \geq s\log\frac{em}{s}$**.**

CASE 1: $\hat{s}_0 \geq s_0$**.**

By (56), we need to prove $\hat{A}(\hat{s}_0) \leq 9 s s_0$. By using contradiction, we assume $\hat{A}(\hat{s}_0) > 9 s s_0$ holds at first and obtain:

$$\begin{aligned} \frac{1+\mathrm{pen}'(\hat{s}_0)}{1+\mathrm{pen}'(s_0)} &\geq \exp\left\{\frac{K}{n}\left(\hat{A}(\hat{s}_0)\log ed + \hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}\right) - 4\frac{K}{n}\left(s s_0\log ed + s\log\frac{em}{s}\right)\right\} \\ &\geq \exp\left\{\frac{K}{n}\hat{A}(\hat{s}_0)\log ed - 4\frac{K}{n}\left(s s_0\log ed + s\log\frac{em}{s}\right)\right\} \\ &\geq \exp\left\{\frac{K}{n}\hat{A}(\hat{s}_0)\log ed - 8\frac{K}{n}s s_0\log ed\right\} \\ &\geq \exp\left\{\frac{K}{9n}\hat{A}(\hat{s}_0)\log ed\right\}. \end{aligned} \tag{57}$$

Combining with (49) we have

$$\frac{1}{n}\left\|y - X\hat{\beta}\,(s_0)\right\|_2^2 \leq \frac{1}{n}\|\xi\|_2^2 + C_1\frac{\sigma^2 s s_0\Delta}{n} \leq \frac{1}{n}\|\xi\|_2^2 + C_6\sigma^2\frac{\hat{A}(\hat{s}_0)\log ed}{n}. \tag{58}$$

Besides, by (51), we also have

$$\frac{1}{n}\left\|y - X\hat{\beta}\,(\hat{s}_0)\right\|_2^2 \geq \frac{1}{n}\|\xi\|_2^2 - \frac{1}{n}C_7\sigma^2\left(\hat{A}(\hat{s}_0)\log ed\right), \tag{59}$$

where we use $a^2-2ab\geq -b^2$, and the inner product term of (51) is upper bounded by (52). Therefore, combining (48) and (57)-(59), we have

$$\left(\frac{1}{n}\|\xi\|_2^2-\frac{C_7\sigma^2}{n}\hat{A}(\hat{s}_0)\log ed\right)\exp\left\{\frac{K}{9n}\hat{A}(\hat{s}_0)\log ed\right\}\leq\frac{1}{n}\|\xi\|_2^2+\frac{C_6\sigma^2}{n}\hat{A}(\hat{s}_0)\log ed. \quad (60)$$

Let $t=\frac{1}{9n}\hat{A}(\hat{s}_0)\log ed$. Note that n is large enough such that $t\in(0,\frac{1}{K})$ by Assumption 2. To establish a contradiction, we need to verify that for $t\in(0,\frac{1}{K})$,

$$F(t)=\left(1-\frac{9C_7t}{\|\xi\|_n^2/\sigma^2}\right)\exp(Kt)-\left(1+\frac{9C_6t}{\|\xi\|_n^2/\sigma^2}\right)>0 \quad (61)$$

always holds. Note that $F(0)=0$ and $F'(t)=\exp(Kt)\left\{K-\frac{9C_7(Kt+1)}{\|\xi\|_n^2/\sigma^2}\right\}-\frac{9C_6}{\|\xi\|_n^2/\sigma^2}$. By $0.9\sigma^2\leq\frac{1}{n}\|\xi\|_2^2\leq 1.1\sigma^2$ and $Kt\in(0,1)$, we could select a sufficiently large $K\geq 10C_6+20C_7$. Hence, we verify that $F'(t)>0$ for $\forall t\in(0,\frac{1}{K})$, which leads to an absurd to (60) with high probability.

Therefore, we prove $\hat{A}(\hat{s}_0)\leq 9ss_0$ holds with high probability. Then, based on (56) we derive that

$$\hat{\Gamma}\leq 10\left(ss_0\log ed+s\log(em/s)\right),\ \forall\hat{s}_0\geq s_0, \quad (62)$$

which proves (55) holds with high probability.

CASE 2: $\hat{s}_0<s_0$.

We divide this case into two subcases and analyse them respectively.

(a) If $9ss_0\log ed>\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}$, we just bound the inner product in (52) by:

$$\begin{aligned}\left|\left\langle\xi,\frac{X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)}{\left\|X\left(\beta^*-\hat{\beta}\left(\hat{s}_0\right)\right)\right\|_2}\right\rangle\right|^2&\leq 5ss_0\log ed+(s+\hat{s}(\hat{s}_0))\log\frac{em}{(s+\hat{s}(\hat{s}_0))}\\&\leq 5ss_0\log ed+s\log(em/s)+9ss_0\log ed\\&\leq 14\left(ss_0\log ed+s\log(em/s)\right),\end{aligned}$$

where the first inequality uses $\hat{A}(\hat{s}_0)<4ss_0$ for $\hat{s}_0<s_0$, and the second inequality uses $\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}<9ss_0\log ed$. By solving (53), we derive an upper bound for $\mathcal{L}$ as

$$\mathcal{L}^2\precsim\frac{\sigma^2}{n}\left(ss_0\log ed+s\log\frac{em}{s}\right), \quad (63)$$

therefore by DSRIP condition we prove (14).

(b) If $9ss_0\log ed\leq\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}$. Then, similar to case 1, we will find an absurd with high probability. First, we obtain

$$\begin{aligned}\frac{1+\mathrm{pen}'\left(\hat{s}_0\right)}{1+\mathrm{pen}'\left(s_0\right)}&\geq\exp\left\{\frac{K}{n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}-8\frac{K}{n}ss_0\log ed\right\}\\&\geq\exp\left\{\frac{K}{9n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}\right\}.\end{aligned}$$

Then, use similar techniques in (58) and (59),we obtain the following inequalities:

$$\begin{aligned}\frac{1}{n}\left\|y-X\hat{\beta}\left(s_0\right)\right\|_2^2 &\leq \frac{1}{n}\|\xi\|_2^2+C_1\frac{\sigma^2 s s_0\Delta}{n}\\ &\leq \frac{1}{n}\|\xi\|_2^2+\frac{2C_1}{9n}\sigma^2\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)},\\ \frac{1}{n}\left\|y-X\hat{\beta}\left(\hat{s}_0\right)\right\|_2^2 &\geq \frac{1}{n}\|\xi\|_2^2-\frac{3\sigma^2}{n}\left(\left(s+\hat{s}(\hat{s}_0)\right)\log\frac{em}{s+\hat{s}(\hat{s}_0)}+4ss_0\log ed\right)\\ &\geq \frac{1}{n}\|\xi\|_2^2-\frac{14\sigma^2}{3n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)},\end{aligned} \tag{64}$$

and thus

$$\begin{aligned}&\left(\frac{1}{n}\|\xi\|_2^2-\frac{14\sigma^2}{3n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}\right)\exp\left\{\frac{K}{9n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}\right\}\\ \leq&\frac{1}{n}\|\xi\|_2^2+\frac{2C_1}{9n}\sigma^2\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}.\end{aligned} \tag{65}$$

Now let $t=\frac{1}{9n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}$, and using the same technique corresponding to (61), with a sufficiently large $K\geq\frac{20C_1+840}{9}$ and we get an absurd. Therefore, we prove that with high probability, $9ss_0\log ed>\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}$ holds, which leads to (63) and completes the proof in case 2.

By far, we have finished the proof in **Assumption A:** $ss_0\log ed\geq s\log\frac{em}{s}$. When $ss_0\log ed<s\log\frac{em}{s}$, the proof strategy is similar and we just give a proof sketch below.

Assumption B: $ss_0\log ed<s\log\frac{em}{s}$**.**

CASE 3: $\hat{s}_0\geq s_0$**.**
Similar to case 2, we continue to divide this case into two subcases:

(a) If $\hat{A}(\hat{s}_0)\log ed\leq 9s\log(em/s)$, we obtain

$$\begin{aligned}\hat{\Gamma}&=\left(ss_0+\hat{A}(\hat{s}_0)\right)\log ed+(s+\hat{s}(\hat{s}_0))\log\frac{em}{s+\hat{s}(\hat{s}_0)}\\ &\leq ss_0\log ed+9s\log(em/s)+5s\log(em/s)\\ &\leq 14\left(ss_0\log ed+s\log\frac{em}{s}\right).\end{aligned} \tag{66}$$

Therefore, we prove that (55) holds.

(b) If $\hat{A}(\hat{s}_0)\log ed>9s\log(em/s)$, we show that

$$\frac{1+\mathrm{pen}'\left(\hat{s}_0\right)}{1+\mathrm{pen}'\left(s_0\right)}\geq\exp\left\{\frac{K}{9n}\hat{A}(\hat{s}_0)\log ed\right\}.$$

Hence by using a strategy similar to (64) and (65), we show that $\hat{A}(\hat{s}_0)\log ed>9s\log(em/s)$ can not hold with high probability. Therefore, in case 3, (66) holds with high probability, which prove that (55) holds.

CASE 4: $\hat{s}_0 < s_0$**.**

In this case, we just need to control $\hat{s}(\hat{s}_0)$. By using a contradiction similar to case 1, at first, we assume $\hat{s}(\hat{s}_0) \geq 9s$ holds, which leads that

$$\begin{aligned}\frac{1+\text{pen}'(\hat{s}_0)}{1+\text{pen}'(s_0)} &\geq \exp\left\{\frac{K}{n}\left(\hat{A}(\hat{s}_0)\log ed + \hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}\right) - 4\frac{K}{n}\left(ss_0\log ed + s\log\frac{em}{s}\right)\right\} \\ &\geq \exp\left\{\frac{K}{9n}\hat{s}(\hat{s}_0)\log\frac{em}{\hat{s}(\hat{s}_0)}\right\},\end{aligned} \tag{67}$$

and by using a strategy similar to (58)-(61) we prove that $\hat{s}(\hat{s}_0) \geq 9s$ can not hold with high probability. Therefore we obtain

$$\begin{aligned}\hat{\Gamma} &= \left(ss_0 + \hat{A}(\hat{s}_0)\right)\log ed + (s+\hat{s}(\hat{s}_0))\log\frac{em}{s+\hat{s}(\hat{s}_0)} \\ &\leq 5ss_0\log ed + 10s\log(em/s) \\ &\leq 10\left(ss_0\log ed + s\log\frac{em}{s}\right),\ \forall \hat{s}_0 < s_0,\end{aligned} \tag{68}$$

which leads to (55).

Overall, combining these 4 cases, we derive the upper bound for $\mathcal{L}$ as

$$\mathcal{L}^2 \precsim \frac{\sigma^2\left(ss_0\log ed + s\log\frac{em}{s}\right)}{n},$$

and by DSRIP condition, we finally complete the proof of Theorem 18.

Proof of Theorem 20

We use a strategy similar to Theorem 4 to prove these results. For ease to display, we define $\Upsilon(A,\tilde{\beta}^t) := \sum_{(i,j)\in A}\langle\Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle^2$. In specific, if $A\cup\text{supp}(\beta^*-\tilde{\beta}^t) \in \mathcal{S}^{m,d}(3s, \frac{5s_0}{3})$, by DSRIP$(3s, \frac{5}{3}s_0, \delta)$ condition, we have $\Upsilon(A,\tilde{\beta}^t) \leq \delta^2\|\beta^*-\tilde{\beta}^t\|_2^2$. In the proof of Theorem 20 and 21 (and also in Lemma 25-27), we use double index (i,j) to denote the i-th entry (variable) of the j-th group G_j. Firstly, we provide the probability inequalities used

frequently in this proof:

$$
\begin{aligned}
P\left\{\forall S\in\mathcal{S}(s',s_0),\ \|\Xi_S\|_2^2>\frac{6\sigma^2 s's_0}{n}\Delta(s',s_0)\right\}&=o(1),\quad \text{where } s'=\frac{s}{8\Delta^2};\\
P\left\{\sum_{(i,j)\in S_{G^*}}\Xi_{ij}^2\mathrm{I}\Big\{|\Xi_{ij}|\geq\tilde{\lambda}_1\Big\}\geq\frac{\sigma^2 ss_0}{n\Delta}\right\}&=o(1);\\
P\left(\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot\mathrm{I}\Big(|\Xi_{ij}|>\frac{\epsilon}{2}\tilde{\lambda}_2\Big)\geq\frac{\sigma^2 ss_0}{n\Delta}\right)&=o(1);\\
P\left\{\sum_{j\in G^*}s_0\tilde{\lambda}_2^2\cdot\mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2>\frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)\geq\frac{\sigma^2 ss_0}{n\Delta}\right\}&=o(1);\\
P\left(\|\Xi_{S^*}\|_2^2\geq\frac{2\sigma^2 ss_0}{n}\right)&=o(1),
\end{aligned}
\tag{69}
$$

as $\min\{\Delta,\ ss_0/\Delta\}\to\infty$. Define $\Delta(s',s_0):=\frac{1}{s_0}\log\frac{em}{s'}+\log\frac{ed}{s_0}$. We provide the proof of the above inequalities in Appendix B.

Here we prove Theorem 20 by mathematical induction. From the assumption, the initial estimator $\tilde{\beta}^0=\hat{\beta}$ is $(2s,\frac{3}{2}s_0)$-sparse and minimax optimal. It is easy to check that the three results in Theorem 20 hold for $t=0$. Now for $\forall t\geq 0$, assume the conclusions in Theorem 20 hold for the t-th iteration, we will prove these hold for the $(t+1)$-th iteration by contradiction and induction.

Step 1 (Control falsely discovered groups).

Assume that more than s groups are falsely discovered in the $(t+1)$-th iteration. Then, we can always choose arbitrary s falsely discovered groups and construct a (s,s_0)-sparse set $S'_{OG}\in\tilde{S}^{t+1}\cap S^c_{G^*}$. The details of the selection process can be described as follows:

For any selected group $j\notin G^*$, if it has more than s_0 falsely discovered entries, then choose arbitrarily s_0 non-zero entries of these falsely discovered entries into S'_{OG}; if it has less than s_0, then we choose all these falsely discovered entries into S'_{OG}. We repeat this operation s times for any s falsely discovered groups, and we obtain a (s,s_0)-sparse set S'_{OG}.

Then, based on the definition of DSIHT operator $\mathcal{T}_{s_0,\tilde{\lambda}_2}(\cdot)$, for any falsely discovered group j selected into set S'_{OG}, we have $\|\tilde{\beta}^{t+1}_{G_j\cap S'_{OG}}\|_2^2\geq s_0\tilde{\lambda}_2^2$, which yields that

$$
\begin{aligned}
\sqrt{ss_0}\tilde{\lambda}_2\leq&\sqrt{\Upsilon\left(S'_{OG},\tilde{\beta}^t\right)}+\sqrt{\sum_{(i,j)\in S'_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{\mathcal{T}_{\lambda_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}\\
\overset{(i)}{\leq}&\frac{7}{2}\delta\|\tilde{\beta}^t-\beta^*\|_2+\sqrt{\frac{\sigma^2 ss_0}{n\Delta}},
\end{aligned}
\tag{70}
$$

where inequality (i) follows Lemma 25. From the assumption of mathematical induction, since (17) holds for t-iteration, we have $\left\|\tilde{\beta}^t-\beta^*\right\|_2<\ 16\sigma\sqrt{\frac{ss_0\Delta}{n}}+16\sqrt{\frac{\sigma^2 ss_0}{n}}$. Combining

with (70), we obtain

$$\sqrt{ss_0}\tilde{\lambda}_2 = \sqrt{\frac{32\sigma^2 ss_0\Delta}{n}} < 2.8\sqrt{\frac{\sigma^2 ss_0\Delta}{n}} + 3.8\sqrt{\frac{\sigma^2 ss_0}{n}}, \tag{71}$$

which can not hold when $\Delta > 2.5$. Thus we find the absurd.

We have proved that no more than s groups are falsely discovered in the $(t+1)$-th iteration. Next, we will prove that no more than ss_0 entries will be falsely discovered outside true groups G^*. If not so, we can construct an (s, s_0)-sparse set $S''_{OG} \in \tilde{S}^t \cap S^c_{G^*}$. Then we obtain

$$\begin{aligned}\sqrt{ss_0}\tilde{\lambda}_2 \leq & \sqrt{\Upsilon\left(S''_{OG}, \tilde{\beta}^t\right)} + \sqrt{\sum_{(i,j)\in S''_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{\mathcal{T}_{\lambda_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij} \neq 0\Big\}} \\ \overset{(i)}{\leq} & \frac{7}{2}\delta\|\tilde{\beta}^t - \beta^*\|_2 + \sqrt{\frac{\sigma^2 ss_0}{n\Delta}},\end{aligned} \tag{72}$$

where inequality (i) follows Lemma 25. This leads to a contradiction as (71).

Step 2 (Control falsely discovered entries in S_{G^*}).

Assume that there are more than ss_0 falsely discovered entries within the true groups G^*. Then, we can construct a (s, s_0)-sparse set $S_{IG} \in S_{G^*} \cap \tilde{S}^t \cap (S^*)^c$ such that for each entry in S_{IG}, $|\tilde{\beta}_{ij}^{t+1}| = |\Xi_{ij} + \langle \Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle| \geq \tilde{\lambda}_2$ always holds, which yields that

$$\begin{aligned}\sqrt{ss_0}\tilde{\lambda}_2 \leq & \sqrt{\Upsilon\left(S_{IG}, \tilde{\beta}^t\right)} + \sqrt{\sum_{(i,j)\in S_{IG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij} + \langle \Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle| \geq \tilde{\lambda}_2\Big\}} \\ \leq & \delta\|\tilde{\beta}^t - \beta^*\|_2 + \sqrt{\sum_{(i,j)\in S_{G^*}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| \geq \tilde{\lambda}_1\Big\}} \\ & + \sqrt{\sum_{(i,j)\in S_{IG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| < \tilde{\lambda}_1 < |\langle \Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle|\Big\}} \\ \overset{(i)}{<} & 2\delta\|\tilde{\beta}^t - \beta^*\|_2 + \sqrt{\frac{\sigma^2 ss_0}{n\Delta}},\end{aligned} \tag{73}$$

where inequality (i) follows Lemma 26. Since (17) holds for t-th iteration, it leads to a contradiction as (71) again.

Step 3 (ℓ_2 estimation error of $\tilde{\beta}^{t+1}$).

Now we have already proved the first two conclusions in Theorem 20 still hold in the $(t+1)$-th iteration, and then we will prove the third one also holds for $(t+1)$-th iteration. Note that

$$\tilde{\beta}_{ij}^{t+1} - \beta_{ij}^* = -\tilde{H}_i^{t+1} \cdot \mathrm{I}\left((i,j) \notin \tilde{S}^{t+1}\right) + \langle \Phi_{kj}^\top, \beta^* - \tilde{\beta}^t\rangle + \Xi_{ij} \tag{74}$$

We now focus on the estimation error on S^* and $\tilde{S}^{t+1} \cap (S^*)^c$ respectively. On S^*, we have

$$\begin{aligned}
\|\tilde{\beta}^{t+1}_{S^*} - \beta^*_{S^*}\|_2 \leq & \sqrt{\sum_{(i,j)\in S^*} \left(\tilde{H}^{t+1}_{ij}\right)^2 \mathrm{I}\left((i,j) \notin \tilde{S}^{t+1}\right)} + \sqrt{\Upsilon\left(S^*, \tilde{\beta}^t\right)} + \sqrt{\sum_{(i,j)\in S^*} \Xi^2_{ij}} \\
\overset{(i)}{\leq} & \frac{4}{\epsilon}\delta \left\|\tilde{\beta}^t - \beta^*\right\|_2 + 2\sqrt{\frac{\sigma^2 s s_0}{n\Delta}} + \sqrt{\Upsilon\left(S^*, \tilde{\beta}^t\right)} + \sqrt{\sum_{(i,j)\in S^*} \Xi^2_{ij}} \\
\overset{(ii)}{\leq} & \left(\frac{4}{\epsilon} + 1\right)\delta \left\|\tilde{\beta}^t - \beta^*\right\|_2 + 2\sqrt{\frac{\sigma^2 s s_0}{n\Delta}} + \sqrt{\frac{2\sigma^2 s s_0}{n}},
\end{aligned} \tag{75}$$

where inequality (i) uses the result of Lemma 27. Inequality (ii) uses Lemma 3 and Theorem 2.1 in Hsu et al. (2012), that is, for a fixed set $S^* \in \mathcal{S}^{m,d}(s, s_0)$ and every $t > 0$, we have

$$P\left(\frac{n}{\sigma^2}\|\Xi_{S^*}\|_2^2 \geq s s_0 + 2(1+\delta)\sqrt{s s_0 t} + 2(1+\delta)t\right) \leq e^{-t}. \tag{76}$$

Let $t = \frac{s s_0}{10}$. Based on $\delta < \frac{1}{5}$, we obtain that $P\left(\|\Xi_{S^*}\|_2^2 \geq \frac{2\sigma^2 s s_0}{n}\right) \leq \exp\left(-\frac{s s_0}{10}\right) = o(1)$ as $s s_0 \to \infty$.

On $\tilde{S}^{t+1} \cap (S^*)^c$, when X satisfies DSRIP$(3s, \frac{5}{3}s_0, \delta)$, we have

$$\begin{aligned}
& \left\|\tilde{\beta}^{t+1}_{\tilde{S}^{t+1}\cap(S^*)^c} - \beta^*_{\tilde{S}^{t+1}\cap(S^*)^c}\right\|_2 = \left\|\tilde{\beta}^{t+1}_{\tilde{S}^{t+1}\cap(S^*)^c}\right\|_2 \\
\leq & \sqrt{\Upsilon\left(\tilde{S}^{t+1} \cap (S^*)^c, \tilde{\beta}^t\right)} + \sqrt{\sum_{(i,j)\in\tilde{S}^{t+1}\cap(S^*)^c} \Xi^2_{ij} \mathrm{I}\left(T_{\lambda_2, s_0}\left(\tilde{H}^{t+1}\right)_{ij} \neq 0\right)} \\
\leq & \delta \left\|\tilde{\beta}^t - \beta^*\right\|_2 + \sqrt{\sum_{(i,j)\in S_{OG}} \Xi^2_{ij} \mathrm{I}\left(T_{\lambda_2, s_0}\left(\tilde{H}^{t+1}\right)_{ij} \neq 0\right)} + \sqrt{\sum_{(i,j)\in S_{IG}} \Xi^2_{ij} \mathrm{I}\left(T_{\lambda_2, s_0}\left(\tilde{H}^{t+1}\right)_{ij} \neq 0\right)} \\
\overset{(i)}{\leq} & \frac{9}{2}\delta \left\|\tilde{\beta}^t - \beta^*\right\|_2 + 2\sqrt{\frac{\sigma^2 s s_0}{n\Delta}},
\end{aligned} \tag{77}$$

where inequality (i) follows from Lemma 25, (73) and Lemma 26.

Finally, based on (75) and (77), we have

$$\begin{aligned}
\left\|\tilde{\beta}^{t+1} - \beta^*\right\|_2 \leq & \left\|\tilde{\beta}^{t+1}_{S^*} - \beta^*_{S^*}\right\|_2 + \left\|\tilde{\beta}^{t+1}_{\tilde{S}^{t+1}\cap(S^*)^c} - \beta^*_{\tilde{S}^{t+1}\cap(S^*)^c}\right\|_2 \\
\leq & \left(\frac{4}{\epsilon} + \frac{11}{2}\right)\delta \left\|\tilde{\beta}^t - \beta^*\right\|_2 + 4\sqrt{\frac{\sigma^2 s s_0}{n\Delta}} + \sqrt{\frac{2\sigma^2 s s_0}{n}} \\
\overset{(i)}{<} & \left(\frac{4}{\epsilon} + \frac{11}{2}\right)\delta \left\|\tilde{\beta}^t - \beta^*\right\|_2 + 4\sqrt{\frac{\sigma^2 s s_0}{n}},
\end{aligned} \tag{78}$$

where in inequality (i) we assume $\Delta > 2.5$, which leads $\frac{4}{\sqrt{\Delta}} + \sqrt{2} < 4$.

Then, based on the initialized inequality $\left\|\tilde{\beta}^0 - \beta^*\right\|_2 \leq 16\sigma\sqrt{\frac{s s_0 \Delta}{n}}$ and $\delta \leq \epsilon^4 \wedge 0.05$, we have $\delta\left(\frac{4}{\epsilon} + \frac{11}{2}\right) \leq \frac{3}{4}$, which leads

$$\left\|\tilde{\beta}^{t+1} - \beta^*\right\|_2 < 16\left(\frac{3}{4}\right)^{t+1}\sigma\sqrt{\frac{s s_0 \Delta}{n}} + 16\sqrt{\frac{\sigma^2 s s_0}{n}}. \tag{79}$$

Consequently, we prove that the conclusions in Theorem 20 hold for the $(t+1)$-th iteration, which completes the proof.

proof of Theorem 21

Under the conditions of Theorem 20, note that the probability inequalities in (69) still hold.

Step 1 (Sharp upper bound).

Let $t > 2\log(256\Delta)$ and we have

$$16\left(\frac{3}{4}\right)^t \sigma\sqrt{\frac{ss_0\Delta}{n}} < \sqrt{\frac{\sigma^2 ss_0}{n}}. \tag{80}$$

From (17), we have

$$\left\|\tilde{\beta}^t - \beta^*\right\|_2 < \sqrt{\frac{\sigma^2 ss_0}{n}} + 16\sqrt{\frac{\sigma^2 ss_0}{n}} = 17\sqrt{\frac{\sigma^2 ss_0}{n}}. \tag{81}$$

Step 2 (Group-wise almost full recovery).

Note that based on the first conclusion of Theorem 20, no more than s groups are falsely discovered in the $(t+1)$-th iteration. Denote G_{FD}^{t+1} as the falsely discovered group index set in the $(t+1)$-th iteration, which satisfies $\left|G_{FD}^{t+1}\right| < s$. Then, we have

$$\begin{aligned}
\|\tilde{\eta}_G^{t+1} - \eta_G^*\|_0 &= \sum_{j=1}^{m} |(\tilde{\eta}_G^{t+1})_j - (\eta_G^*)_j| \\
&= \sum_{j\in G^*} |(\tilde{\eta}_G^{t+1})_j - 1| + \sum_{j\in G_{FD}^{t+1}} |(\tilde{\eta}_G^{t+1})_j - 0| \\
&= \sum_{j\in G^*} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_2,s_0}(\tilde{H}_{G_j}^{t+1}) = \mathbf{0}\right) + \sum_{j\in G_{FD}^{t+1}} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_2,s_0}(\tilde{H}_{G_j}^{t+1}) \neq \mathbf{0}\right).
\end{aligned} \tag{82}$$

For the first term in (82), based on Lemma 27, we have

$$
\begin{aligned}
&\sum_{j\in G^*} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_2,s_0}(\tilde{H}^{t+1}_{G_j}) = \mathbf{0}\right) \\
\leq &\sum_{j\in G^*} \mathrm{I}\left(\sum_{k\in S_{G_j}\cap S^*} \left(\tilde{H}^{t+1}_{kj}\right)^2 \mathrm{I}\left(|\tilde{H}^{t+1}_{kj}| \geq \tilde{\lambda}_2\right) < s_0\tilde{\lambda}_2^2\right) \\
\leq &\sum_{j\in G^*} \mathrm{I}\left(\sum_{k\in S_{G_j}\cap S^*} \left(\tilde{H}^{t+1}_{kj}\right)^2 < (s_0+s_j)\tilde{\lambda}_2^2\right) \\
\overset{(i)}{\leq} &\sum_{j\in G^*} \mathrm{I}\left(\Upsilon\left(S_{G_j}\cap S^*, \tilde{\beta}^t\right) > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right) + \sum_{j\in G^*} \mathrm{I}\left(\sum_{k\in S_{G_j}\cap S^*} \Xi^2_{kj} > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right) \\
\overset{(ii)}{\leq} &\frac{4\delta^2\left\|\tilde{\beta}^t-\beta^*\right\|_2^2}{\epsilon^2 s_0\tilde{\lambda}_2^2} + \frac{\sigma^2 s}{n\tilde{\lambda}_2^2\Delta} \\
\precsim &\frac{s}{\Delta} + \frac{s}{\Delta^2} \\
= &O\left(\frac{s}{\Delta}\right), \quad \text{as } \Delta\to\infty,
\end{aligned}
\tag{83}
$$

where inequality (i) follows from (110) and inequality (ii) follows from (114).

For the second term in (82), based on Lemma 25, we have

$$
\begin{aligned}
&\sum_{j\in G^{t+1}_{FD}} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_2,s_0}(\tilde{H}^{t+1}_{G_j}) \neq \mathbf{0}\right) \\
\leq &\sum_{j\in G^{t+1}_{FD}} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_1,s_0}(\Xi_{\tilde{S}^{t+1}\cap S_{G_j}}) \neq \mathbf{0}\right) + \sum_{j\in G^{t+1}_{FD}} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_1,s_0}(\Xi_{\tilde{S}^{t+1}\cap S_{G_j}}) = \mathbf{0},\ \mathcal{T}_{\tilde{\lambda}_2,s_0}(\tilde{H}^{t+1}_{G_j}) \neq \mathbf{0}\right) \\
\overset{(i)}{\leq} &\frac{s}{8\Delta^2} + \sum_{j\in G^{t+1}_{FD}} \mathrm{I}\left(\sum_{k\in\tilde{S}^{t+1}\cap S_{G_j}} \Xi^2_{kj}\mathrm{I}(|\Xi_{kj}|\geq\tilde{\lambda}_1) < s_0\tilde{\lambda}_1^2,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}(\tilde{H}^{t+1}_{G_j}) \neq \mathbf{0}\right) \\
\overset{(ii)}{\leq} &\frac{s}{8\Delta^2} + \sum_{j\in G^{t+1}_{FD}} \mathrm{I}\left(s_0\tilde{\lambda}_1^2 \leq 2\Upsilon\left(\tilde{S}^{t+1}\cap S_{G_j}, \tilde{\beta}^t\right)\right) \\
\precsim &\frac{s}{\Delta^2} + \frac{s}{\Delta} = O\left(\frac{s}{\Delta}\right), \quad \text{as } \Delta\to\infty,
\end{aligned}
\tag{84}
$$

where inequality (i) follows from a similar contradiction in the proof of the first term in Lemma 25, and inequality (ii) follows from the result of (99).

Combining (82), (83) and (84) together, we prove that $\|\tilde{\eta}^{t+1}_G - \eta^*_G\|_0 = O\left(\frac{s}{\Delta}\right)$.

Step 3 (Element-wise almost full recovery).

Based on the first two conclusions of Theorem 20, we have

$$\begin{aligned}\|\tilde{\eta}^{t+1}-\eta^*\|_0 =& \sum_{j=1}^{m}\sum_{i=1}^{d}|\tilde{\eta}_{ij}^{t+1}-\eta_{ij}^*| \\ =& \sum_{(i,j)\in S^*}|\tilde{\eta}_{ij}^{t+1}-1|+\sum_{(i,j)\in S_{G^*}\cap(S^*)^c\cap\tilde{S}^{t+1}}|\tilde{\eta}_{ij}^{t+1}-0|+\sum_{(i,j)\in S_{G^*}^c\cap\tilde{S}^{t+1}}|\tilde{\eta}_{ij}^{t+1}-0|.\end{aligned} \tag{85}$$

We can just analyze these three terms respectively. For the first one, note that

$$\begin{aligned}\sum_{(i,j)\in S^*}|\tilde{\eta}_{ij}^{t+1}-1| =& \sum_{(i,j)\in S^*}\mathrm{I}\left((i,j)\notin\tilde{S}^{t+1}\right) \\ \overset{(i)}{\leq}& \sum_{(i,j)\in S^*}\mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|<\tilde{\lambda}_2\right) \\ &+\frac{1}{\tilde{\lambda}_2^2}\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2\mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|\geq\tilde{\lambda}_2,\ \sum_{k\in S_{G_j}\cap S^*}\left(\tilde{H}_{kj}^{t+1}\right)^2<(s_0+s_j)\tilde{\lambda}_2^2\right) \\ \overset{(ii)}{\leq}& \frac{16\delta^2\left\|\tilde{\beta}^t-\beta^*\right\|_2^2}{\epsilon^2\tilde{\lambda}_2^2}+\frac{4\sigma^2 ss_0}{n\tilde{\lambda}_2^2\Delta}\precsim\frac{ss_0}{\Delta}+\frac{ss_0}{\Delta^2} \\ =& O\left(\frac{ss_0}{\Delta}\right),\ \text{ as } \Delta\to\infty,\end{aligned} \tag{86}$$

where inequality (i) follows from the first inequality of (106) in Lemma 27, and inequality (ii) follows from the last inequality of (106), (109) and (115).

For the second term, we obtain

$$\begin{aligned}\sum_{(i,j)\in S_{G^*}\cap(S^*)^c\cap\tilde{S}^{t+1}}|\tilde{\eta}_{ij}^{t+1}-0| \leq& \sum_{(i,j)\in S_{G^*}\cap(S^*)^c\cap\tilde{S}^{t+1}}\mathrm{I}\left(\left|\tilde{H}_{ij}^{t+1}\right|\geq\tilde{\lambda}_2\right) \\ \overset{(i)}{\leq}& \sum_{(i,j)\in S_{G^*}}\mathrm{I}\left(|\Xi_{ij}|\geq\tilde{\lambda}_1\right) \\ &+\sum_{(i,j)\in S_{G^*}\cap(S^*)^c\cap\tilde{S}^{t+1}}\mathrm{I}\big(|\Xi_{ij}|<\tilde{\lambda}_1<|\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle|\big) \\ \overset{(ii)}{\leq}& \frac{\sigma^2 ss_0}{n\tilde{\lambda}_1^2\Delta}+\frac{\delta^2\left\|\tilde{\beta}^t-\beta^*\right\|_2^2}{\tilde{\lambda}_1^2}\precsim\frac{ss_0}{\Delta^2(s,s_0)}+\frac{ss_0}{\Delta} \\ =& O\left(\frac{ss_0}{\Delta}\right),\ \text{ as } \Delta\to\infty,\end{aligned} \tag{87}$$

where inequality (i) follows from (73) in Theorem 20, inequality (ii) follows from the probability inequality (104) in Lemma 26 and $\sum_{(i,j)\in S_{G^*}}\tilde{\lambda}_1^2\mathrm{I}\Big\{|\Xi_{ij}|\geq\tilde{\lambda}_1\Big\}\leq\sum_{(i,j)\in S_{G^*}}\Xi_{ij}^2\mathrm{I}\Big\{|\Xi_{ij}|\geq\tilde{\lambda}_1\Big\}$.

For the third term, we obtain

$$\begin{aligned}
&\sum_{(i,j)\in S^c_{G^*}\cap\tilde{S}^{t+1}} |\tilde{\eta}^{t+1}_{ij}-0| = \sum_{(i,j)\in S^c_{G^*}\cap\tilde{S}^{t+1}} \mathrm{I}\left(\mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\right)\\
&\overset{(i)}{\leq} \sum_{(i,j)\in S^c_{G^*}\cap\tilde{S}^{t+1}} \mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij}\neq 0\Big\} + \sum_{(i,j)\in S^c_{G^*}\cap\tilde{S}^{t+1}} \mathrm{I}\Big\{|\Xi_{ij}|<\tilde{\lambda}_1,\ |\Xi_{ij}+\langle\Phi^\top_{ij},\beta^*-\tilde{\beta}^t\rangle|\geq\tilde{\lambda}_2\Big\}\\
&\quad + \frac{1}{\tilde{\lambda}_1^2}\sum_{(i,j)\in S^c_{G^*}\cap\tilde{S}^{t+1}} \Xi^2_{ij}\mathrm{I}\left\{|\Xi_{ij}|\geq\tilde{\lambda}_1,\ \sum_{k\in S_{G_j}\cap\tilde{S}^{t+1}}\Xi^2_{kj}\mathrm{I}(|\Xi_{kj}|\geq\tilde{\lambda}_1)<s_0\tilde{\lambda}_1^2,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\right\}\\
&\overset{(ii)}{<} s's_0 + \frac{\delta^2\left\|\tilde{\beta}^t-\beta^*\right\|_2^2}{\tilde{\lambda}_1^2} + \frac{2\delta^2\left\|\tilde{\beta}^t-\beta^*\right\|_2^2}{\tilde{\lambda}_1^2}\\
&\precsim \frac{ss_0}{\Delta^2}+\frac{ss_0}{\Delta} = O\left(\frac{ss_0}{\Delta}\right),\quad \text{as } \Delta\to\infty,
\end{aligned} \tag{88}$$

where inequality (i) follows from (94) in Lemma 25, and inequality (ii) follows from the framework of the first term in Lemma 25, (99) and (100), and recall $s'=\frac{s}{8\Delta^2}$.

Combining (85), (86), (87) and (88) together, we prove that $\|\tilde{\eta}^{t+1}-\eta^*\|_0 = O\left(\frac{ss_0}{\Delta}\right)$.

Appendix B : Auxiliary lemmas

Lemma 23 *Assume that X satisfies DSRIP$(2s,\frac{3}{2}s_0,\frac{\delta}{2})$. Then, with probability at least $1-\exp\{-Css_0\Delta\}$, we have*

$$|\sigma_t-\sigma|\leq\sqrt{1+\delta}\|\beta^*-\beta^t\|_2+\frac{1}{20}\sigma.$$

Proof

Denote event $\mathcal{A}=\{|\frac{\|\xi\|_2}{\sigma}-\sqrt{n}|\leq\frac{1}{20}\sqrt{n}\}$. From Hanson-Wright inequality (Rudelson and Vershynin, 2013), it holds that $P(\mathcal{A})\geq 1-e^{-Cn}\geq 1-e^{-Css_0\Delta}$. Therefore,

$$\begin{aligned}
|\sigma_t-\sigma| &\leq |\sigma_t-\frac{\|\xi\|_2}{\sqrt{n}}|+|\frac{\|\xi\|_2}{\sqrt{n}}-\sigma|\\
&\leq \frac{1}{\sqrt{n}}\left|\|X(\beta^*-\beta^t)+\xi\|_2-\|\xi\|_2\right|+\frac{1}{20}\sigma\\
&\leq \sqrt{1+\delta}\|\beta^*-\beta^t\|_2+\frac{1}{20}\sigma,
\end{aligned}$$

where the second inequality follows from event $\mathcal{A}$, and the last inequality follows from DSRIP condition.

■

To control the inner product between ξ and $X\left(\hat{\beta}-\beta^*\right)$, we provide a useful lemma.

Lemma 24 *Given integers $v_1, v_2 > 0$, and assume that β is a $(v_1, v_2/v_1)$-sparse vector. we have*

$$P\left(\sup_{\beta\in\Theta^{m,d}(v_1,\frac{v_2}{v_1})}\left|\left\langle\xi,\frac{X\beta}{\|X\beta\|_2}\right\rangle\right|^2\geq 3\sigma^2(v_1\log\frac{em}{v_1}+v_2\log\frac{edv_1}{v_2})\right)\leq e^{-C\left(v_1\log\frac{em}{v_1}+v_2\log\frac{edv_1}{v_2}\right)}. \tag{89}$$

In specific, if β^* is a (s, s_0)-sparse vector and $\hat{\beta}\in\Theta^{m,d}(\hat{s},\frac{\hat{A}}{\hat{s}})$, i.e., $\|\hat{\beta}\|_0\leq\hat{A}$ and $\|\hat{\beta}\|_{0,2}\leq\hat{s}$, we have

$$P\left(\sup_{\hat{\beta}\in\Theta^{m,d}(\hat{s},\frac{\hat{A}}{\hat{s}})}\left|\left\langle\xi,\frac{X(\hat{\beta}-\beta^*)}{\|X(\hat{\beta}-\beta^*)\|_2}\right\rangle\right|^2\geq 3\sigma^2\Omega^*(\hat{\beta})\right)\leq e^{-C\Omega^*(\hat{\beta})}, \tag{90}$$

where $\Omega^*(\hat{\beta})$ is defined at the beginning of the Appendix.

Proof For a fixed set S satisfies $\mathrm{supp}(\beta)\subseteq S$, denote X_S as the span space of columns of X indexed by S, thus we have $\langle\xi, X\beta\rangle=\langle\xi, X_S\beta_S\rangle$. Denote $\pi_S=X_S(X_S^\top X_S)^{-1}X_S^\top\in\mathbb{R}^{n\times n}$, which is an orthogonal matrix of rank no more than $|S|$. Therefore, for $\forall\beta\in\Theta^{m,d}(v_1,\frac{v_2}{v_1})$, we obtain the following by Cauchy-Schwartz inequality:

$$\left|\left\langle\xi,\frac{X\beta}{\|X\beta\|_2}\right\rangle\right|=\left|\left\langle\pi_S\xi,\frac{X_S\beta_S}{\|X_S\beta_S\|_2}\right\rangle\right|\leq\|\pi_S\xi\|_2\leq\sup_{S\in\mathcal{S}^{m,d}\left(v_1,\frac{v_2}{v_1}\right)}\|\pi_S\xi\|_2\,, \tag{91}$$

Note that for $\forall S\in\mathcal{S}^{m,d}\left(v_1,\frac{v_2}{v_1}\right)$, we have $rank\big(\pi_S\big)\leq v_2$, so that $Tr(\pi_S)\leq rank\big(\pi_S\big)\cdot\|\pi_S\|_2\leq v_2$. Thus by Theorem 2.1 of Hsu et al. (2012), for $\forall t>0$, we have

$$P\left(\frac{\|\pi_S\xi\|_2^2}{\sigma^2}\geq\frac{5}{2}t\right)\leq P\left(\frac{\|\pi_S\xi\|_2^2}{\sigma^2}\geq v_2+2\sqrt{v_2t}+2t\right)\leq e^{-t}, \tag{92}$$

where the first inequality holds when $t\gg v_2$.

Similarly to (26), we have $\left|S^{m,d}\left(v_1,\frac{v_2}{v_1}\right)\right|\leq\left(\frac{em}{v_1}\right)^{v_1}\times\left(\frac{edv_1}{v_2}\right)^{v_2}$, thus by (92) we get a union bound as:

$$P\left(\sup_{S\in\mathcal{S}^{m,d}\left(v_1,\frac{v_2}{v_1}\right)}\|\pi_S\xi\|_2^2\geq 3\sigma^2\left(v_1\log\frac{em}{v_1}+v_2\log\frac{edv_1}{v_2}\right)\right)\leq e^{-C\left(v_1\log\frac{em}{v_1}+v_2\log\frac{edv_1}{v_2}\right)}.$$

Let $t=(1+C)(v_1\log\frac{em}{v_2}+v_1+\log\frac{edv_1}{v_2})$ for some constant $0<C<\frac{1}{5}$, which satisfies $t\gg v_2$. We complete (89).

For (90), for any $\hat{\beta}\in\Theta^{m,d}(\hat{s},\frac{\hat{A}}{\hat{s}})$, combined with $\beta^*\in\Theta^{m,d}(s,s_0)$, so we have:

$$\hat{\beta}-\beta^*\in\Theta^{m,d}\left(\hat{s}+s,\frac{ss_0+\hat{A}}{\hat{s}+s}\right).$$

We let $v_1=\hat{s}+s$ and $v_2=ss_0+\hat{A}$, we obtain the (90) directly by (89). ■

For ease of display, in the next three lemmas, we use double index (i,j) to denote the i-th entry (variable) in the j-th group G_j. Besides, we recall the abbreviation $\Upsilon(A,\tilde{\beta}^t)=\sum_{(i,j)\in A}\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle^2$, $\Delta=\frac{1}{s_0}\log(em/s)+\log(ed/s_0)$ and $\tilde{\lambda}_a=a\sqrt{\frac{8\sigma^2}{n}\left(\log\frac{ed}{s_0}+\frac{1}{s_0}\log\frac{em}{s}\right)}$. Denote $S_{OG}:=\tilde{S}^{t+1}\cap S_{G^*}^c$

Firstly, to bound the ℓ_2 norm of the selected entries of Ξ in S_{OG}, we give the following lemma.

Lemma 25 *Assume all the conditions in Theorem 20 hold. For $\forall t\geq 0$, as $\Delta,\frac{ss_0}{\Delta}\to\infty$, we have*

$$P\left(\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_i^2\mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}<\sqrt{\frac{\sigma^2 ss_0}{n\Delta}}+\frac{5}{2}\delta\|\tilde{\beta}^t-\beta^*\|_2\right)\to 1. \tag{93}$$

Proof Note that

$$\begin{aligned}
&\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}\\
\leq&\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij}\neq 0,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}\\
&+\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij}=0,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}\\
\leq&\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij}\neq 0,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}\\
&+\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{|\Xi_{ij}|<\tilde{\lambda}_1,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\Big\}}\\
&+\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\left\{|\Xi_{ij}|\geq\tilde{\lambda}_1,\sum_{k\in S_{G_j}\cap S_{OG}}\Xi_{kj}^2\mathrm{I}(|\Xi_{kj}|\geq\tilde{\lambda}_1)<s_0\tilde{\lambda}_1^2,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij}\neq 0\right\}}\\
\leq&\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij}\neq 0\Big\}}+\sqrt{\sum_{(i,j)\in S_{OG}}\Xi_{ij}^2\mathrm{I}\Big\{|\Xi_{ij}|<\tilde{\lambda}_1,\ |\Xi_{ij}+\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle|\geq\tilde{\lambda}_2\Big\}}\\
&+\sqrt{\sum_{j\in\tilde{G}^{t+1}\cap(G^*)^c}s_0\tilde{\lambda}_1^2\cdot\mathrm{I}\left\{\sum_{k\in S_{G_j}\cap S_{OG}}\Xi_{kj}^2\mathrm{I}(|\Xi_{kj}|\geq\tilde{\lambda}_1)<s_0\tilde{\lambda}_1^2,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{G_j}\neq 0\right\}}.
\end{aligned}\tag{94}$$

Next, we bound the three terms in the last inequality respectively.

First term. Let $s'=\frac{s}{8\Delta^2}$, $s_0'=s_0$. Then, we show that under the event $\mathcal{E}(s',s_0)$ in Lemma 3, only less than s' groups in $\Xi_{S_{OG}}$ could be discovered by $\mathcal{T}_{\tilde{\lambda}_1,s_0}$. If not so, choose

any s' discovered groups and construct an $S' \subset S_{OG}$ and $S' \in \mathcal{S}^{m,d}(s', s_0)$, which satisfies

$$\sum_{(i,j)\in S'} \Xi_{ij}^2 \geq \sum_{(i,j)\in S'} \Xi_{ij}^2 \mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij} \neq 0\Big\} \geq \ s's_0\tilde{\lambda}_1^2 \geq \ \frac{s}{8\Delta^2}\cdot s_0 \cdot \frac{8\sigma^2}{n}\Delta = \frac{ss_0\sigma^2}{n\Delta}.$$

When Δ is sufficiently large, we can show that $\log\left(8\Delta^2\right) < \frac{s_0}{3}\Delta$, which leads that

$$\Delta < \Delta(s', s_0) := \frac{1}{s_0}\log\frac{em}{s'} + \log\frac{ed}{s_0} < \frac{4}{3}\Delta.$$

Thus we have

$$\sum_{(i,j)\in S'} \Xi_{ij}^2 \geq \frac{ss_0\sigma^2}{n\Delta} = \frac{8s's_0\sigma^2\Delta}{n} > \frac{8s's_0\sigma^2}{n}\cdot\frac{3}{4}\Delta(s', s_0) = \frac{6s's_0\sigma^2\Delta(s', s_0)}{n},$$

which contradicts the event $\mathcal{E}(s', s_0)$ in Lemma 3 with high probability. Thus we show only less than s' groups in $\Xi_{S_{OG}}$ are discovered. Similarly, we can show only less than $s's_0$ entries are discovered in $\Xi_{S_{OG}}$. If not so, take $S_2 \in \mathcal{S}^{m,d}(s', s_0)$ and $S_2 \subset S_{OG}$, whose entries are all falsely discovered in S_{OG}, which leads

$$\sum_{(i,j)\in S_2} \Xi_{ij}^2 \geq s's_0\tilde{\lambda}_1^2 \geq \frac{ss_0\sigma^2}{n\Delta} \geq \frac{6s's_0\sigma^2\Delta(s', s_0)}{n}. \tag{95}$$

Under the event $\mathcal{E}(s', s_0)$ in Lemma 3, (95) leads to an absurd again. Thus we can bound the first term in (94) by

$$\sum_{(i,j)\in S_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_1,s_0}\big(\Xi_{S_{OG}}\big)_{ij} \neq 0\Big\} \leq \sup_{S_2\in\mathcal{S}(s',s_0)} \sum_{(i,j)\in S_2} \Xi_{ij}^2 \leq \frac{6\sigma^2 s's_0\Delta(s', s_0)}{n} \leq \frac{\sigma^2 ss_0}{n\Delta}. \tag{96}$$

Second term. Note that

$$\begin{aligned}
&\sum_{(i,j)\in S_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| < \tilde{\lambda}_1,\ |\Xi_{ij} + \langle\Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle| \geq \tilde{\lambda}_2\Big\} \\
\leq &\sum_{(i,j)\in S_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| < \tilde{\lambda}_1,\ |\Xi_{ij}| + |\langle\Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle| \geq 2\tilde{\lambda}_1\Big\} \\
\leq &\sum_{(i,j)\in S_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| < \tilde{\lambda}_1 \leq |\langle\Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle|\Big\} \\
\leq &\sum_{(i,j)\in S_{OG}} \langle\Phi_{(i,j)}^\top, \beta^* - \tilde{\beta}^t\rangle^2 = \left\|\Phi_{S_{OG}}\left(\beta^* - \tilde{\beta}^t\right)\right\|_2^2.
\end{aligned} \tag{97}$$

Thus we can bound the second term in (94) by

$$\sqrt{\sum_{(i,j)\in S_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| < \tilde{\lambda}_1,\ |\Xi_{ij} + \langle\Phi_{ij}^\top, \beta^* - \tilde{\beta}^t\rangle| \geq \tilde{\lambda}_2\Big\}} \leq \left\|\Phi_{S_{OG}}\left(\beta^* - \tilde{\beta}^t\right)\right\|_2 \leq \delta\left\|\tilde{\beta}^t - \beta^*\right\|_2. \tag{98}$$

Third term. For any group $j \notin G^*$ such that $\sum_{k\in S_{G_j}\cap S_{OG}} \Xi_{kj}^2 \mathrm{I}(|\Xi_{kj}| \geq \tilde{\lambda}_1) < s_0\tilde{\lambda}_1^2$ and $\mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{G_j} \neq \mathbf{0}$ (where the index ranges over S_{OG}), we have

$$\begin{aligned}
s_0\tilde{\lambda}_2^2 \leq & \sum_{k\in S_{G_j}\cap S_{OG}} \Big(\underbrace{\Xi_{kj} + \langle \Phi_{kj}^\top, \beta^* - \tilde{\beta}^t\rangle}_{\tilde{H}_{kj}^{t+1}}\Big)^2 \mathrm{I}\Big(|\Xi_{kj} + \langle \Phi_{kj}^\top, \beta^* - \tilde{\beta}^t\rangle| \geq \tilde{\lambda}_2\Big) \\
\leq & \sum_{k\in S_{G_j}\cap S_{OG}} 2\Xi_{kj}^2 \mathrm{I}\Big(|\tilde{H}_{kj}^{t+1}| \geq \tilde{\lambda}_2\Big) + \sum_{k\in S_{G_j}\cap S_{OG}} 2\langle \Phi_{kj}^\top, \beta^* - \tilde{\beta}^t\rangle^2 \mathrm{I}\Big(|\tilde{H}_{kj}^{t+1}| \geq \tilde{\lambda}_2\Big) \\
\leq & \sum_{k\in S_{G_j}\cap S_{OG}} 2\Xi_{kj}^2 \mathrm{I}\Big(|\Xi_{kj}| \geq \tilde{\lambda}_1\Big) + \sum_{k\in S_{G_j}\cap S_{OG}} 2\Xi_{kj}^2 \mathrm{I}\Big(|\Xi_{kj}| < \tilde{\lambda}_1 \leq |\langle \Phi_{kj}^\top, \beta^* - \tilde{\beta}^t\rangle|\Big) \\
& + 2\sum_{k\in S_{G_j}\cap S_{OG}} \langle \Phi_{kj}^\top, \beta^* - \tilde{\beta}^t\rangle^2 \\
\leq & 2s_0\tilde{\lambda}_1^2 + 4\Upsilon\left(S_{G_j}\cap S_{OG}, \tilde{\beta}^t\right),
\end{aligned} \tag{99}$$

which leads to $s_0\tilde{\lambda}_1^2 \leq 2\Upsilon\left(S_{G_j}\cap S_{OG}, \tilde{\beta}^t\right)$. Thus we can bound the third term in (94) as

$$\begin{aligned}
& \sqrt{\sum_{(i,j)\in S_{OG}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| \geq \tilde{\lambda}_1, \sum_{k\in S_{G_j}\cap S_{OG}} \Xi_{kj}^2 \mathrm{I}(|\Xi_{kj}| \geq \tilde{\lambda}_1) < s_0\tilde{\lambda}_1^2,\ \mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij} \neq 0\Big\}} \\
\leq & \sqrt{\sum_{j\in \tilde{G}^{t+1}\cap (G^*)^c} s_0\tilde{\lambda}_1^2 \mathrm{I}\Big\{s_0\tilde{\lambda}_1^2 \leq 2\Upsilon\left(S_{G_j}\cap S_{OG}, \tilde{\beta}^t\right)\Big\}} \\
\leq & \sqrt{2\Upsilon\left(S_{OG}, \tilde{\beta}^t\right)} \\
< & \frac{3}{2}\delta\left\|\tilde{\beta}^t - \beta^*\right\|_2.
\end{aligned} \tag{100}$$

Combining these three terms (96), (98) and (100) together, we finally get that

$$P\left(\sqrt{\sum_{(i,j)\in S_{OG}} \Xi_i^2 \mathrm{I}\Big\{\mathcal{T}_{\tilde{\lambda}_2,s_0}\big(\tilde{H}^{t+1}\big)_{ij} \neq 0\Big\}} < \sqrt{\frac{\sigma^2 s s_0}{n\Delta}} + \frac{5}{2}\delta\|\tilde{\beta}^t - \beta^*\|_2\right) \to 1, \tag{101}$$

as $\Delta, \frac{ss_0}{\Delta} \to \infty$.

■

Similarly, we can bound the ℓ_2-norm of the selected entries of Ξ within the true groups G^*, which can be expressed in the following lemma.

Lemma 26 *Assume all the conditions in Theorem 20 hold. As $\Delta \to \infty$, we have*

$$P\left(\sqrt{\sum_{(i,j)\in S_{G^*}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| \geq \tilde{\lambda}_1\Big\}} < \sqrt{\frac{\sigma^2 s s_0}{n\Delta}}\right) \to 1. \tag{102}$$

Proof Since for any $(i,j) \in S_{G^*}$, Ξ_{ij} is sub-Gaussian with parameter $\frac{\sigma^2}{n}$, we conclude that

$$\begin{aligned}
\mathbf{E}\Big(\Xi_{ij}^2 \mathrm{I}\left\{|\Xi_{ij}| \geq \tilde{\lambda}_1\right\}\Big) =& \int_0^\infty P\Big(\Xi_{ij}^2 \mathrm{I}\left\{|\Xi_{ij}| \geq \tilde{\lambda}_1\right\} > u\Big)\mathrm{d}u \\
=& \int_0^{\tilde{\lambda}_1^2} P\Big(|\Xi_{ij}| \geq \tilde{\lambda}_1\Big)\mathrm{d}u + \int_{\tilde{\lambda}_1^2}^\infty P\Big(|\Xi_{ij}| \geq \sqrt{u}\Big)\mathrm{d}u \\
\leq& 2\tilde{\lambda}_1^2 \exp\left(-\frac{n\tilde{\lambda}_1^2}{2\sigma^2}\right) + \int_{\tilde{\lambda}_1^2}^\infty 2\exp\left(-\frac{nu}{2\sigma^2}\right)\mathrm{d}u \\
=& \left(2\tilde{\lambda}_1^2 + \frac{4\sigma^2}{n}\right)\exp\left(-\frac{n\tilde{\lambda}_1^2}{2\sigma^2}\right) \\
\leq& 3\tilde{\lambda}_1^2 \exp\left(-\frac{n\tilde{\lambda}_1^2}{2\sigma^2}\right),
\end{aligned} \tag{103}$$

where the last inequality follows from $\tilde{\lambda}_1^2 = \frac{8\sigma^2}{n}\Delta \geq \frac{4\sigma^2}{n}$. Thus, based on Markov inequality we have

$$\begin{aligned}
P\left(\sum_{(i,j)\in S_{G^*}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| \geq \tilde{\lambda}_1\Big\} \geq \frac{\sigma^2 s s_0}{n\Delta}\right) \leq& \frac{n\Delta}{\sigma^2 s s_0} \cdot \mathbf{E}\left(\sum_{(i,j)\in S_{G^*}} \Xi_{ij}^2 \mathrm{I}\Big\{|\Xi_{ij}| \geq \tilde{\lambda}_1\Big\}\right) \\
\leq& 3\Delta \cdot \frac{\tilde{\lambda}_1^2 n}{\sigma^2} \cdot \frac{d}{s_0} \cdot \exp\left(-\frac{n\tilde{\lambda}_1^2}{2\sigma^2}\right) \\
\leq& \frac{3}{4}\left(\frac{n\tilde{\lambda}_1^2}{\sigma^2}\right)^2 \exp\left(-\frac{n\tilde{\lambda}_1^2}{4\sigma^2}\right) \\
=& o(1), \quad \text{as } \Delta \to \infty,
\end{aligned} \tag{104}$$

where the last inequality uses $\tilde{\lambda}_1^2 \geq \frac{4\sigma^2\Delta}{n}$ and $\frac{d}{s_0} < \exp(\Delta) \leq \exp\left(\frac{n\tilde{\lambda}_1^2}{4\sigma^2}\right)$. ■

Now we turn to analyze the term of the estimation error on S^*. Under proper beta-min conditions, we can bound $\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2 \mathrm{I}\left((i,j) \notin \tilde{S}^{t+1}\right)$ by the following lemma.

Lemma 27 *Assume all the conditions in Theorem 20 hold. Then, for any $\epsilon > 0$, we have*

$$P\left(\sqrt{\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2 \mathrm{I}\left((i,j) \notin \tilde{S}^{t+1}\right)} < \frac{4}{\epsilon}\delta\left\|\tilde{\beta}^t - \beta^*\right\|_2 + 2\sqrt{\frac{\sigma^2 s s_0}{n\Delta}}\right) \to 1, \tag{105}$$

as $\Delta \to \infty$.

Proof Note that

$$
\begin{aligned}
&\sqrt{\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2 \mathrm{I}\left((i,j)\notin \tilde{S}^{t+1}\right)}\\
\leq&\sqrt{\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2 \mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|<\tilde{\lambda}_2\right)}\\
&+\sqrt{\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2 \mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|\geq\tilde{\lambda}_2,\ \sum_{k\in S^*\cap S_{G_j}}\left(\tilde{H}_{kj}^{t+1}\right)^2 \mathrm{I}(|\tilde{H}_{kj}^{t+1}|\geq\tilde{\lambda}_2)<s_0\tilde{\lambda}_2^2\right)}\\
<&\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|<\tilde{\lambda}_2\right)}\\
&+\sqrt{\sum_{j\in G^*}s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\left(\tilde{H}_{kj}^{t+1}\right)^2 \mathrm{I}(|\tilde{H}_{kj}^{t+1}|\geq\tilde{\lambda}_2)<s_0\tilde{\lambda}_2^2\right)}\\
\leq&\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|<\tilde{\lambda}_2\right)}+\sqrt{\sum_{j\in G^*}s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\left(\tilde{H}_{kj}^{t+1}\right)^2<(s_j+s_0)\tilde{\lambda}_2^2\right)}.
\end{aligned}
\tag{106}
$$

Next, we analyze these two terms respectively.

First term. Recall that $\tilde{H}_{ij}^{t+1}=\beta_{ij}^*+\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle+\Xi_{ij}$ and $|\beta_{ij}^*|\geq(1+\epsilon)\tilde{\lambda}_2$ holds for every support entry. Therefore, we have

$$
\begin{aligned}
\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|<\tilde{\lambda}_2\right)}\leq&\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\Big(|\beta_{ij}|-|\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle+\Xi_{ij}|<\tilde{\lambda}_2\Big)}\\
\leq&\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\Big(\epsilon\tilde{\lambda}_2<|\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle|+|\Xi_{ij}|\Big)}\\
\leq&\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\Big(|\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle|>\frac{\epsilon}{2}\tilde{\lambda}_2\Big)}\\
&+\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\Big(|\Xi_{ij}|>\frac{\epsilon}{2}\tilde{\lambda}_2\Big)}\\
<&\frac{2}{\epsilon}\sqrt{\Upsilon\left(S^*,\tilde{\beta}^t\right)}+\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\Big(|\Xi_{ij}|>\frac{\epsilon}{2}\tilde{\lambda}_2\Big)}.
\end{aligned}
\tag{107}
$$

Under the fixed S^* and based on Markov inequality, we have

$$\begin{aligned} P\left(\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\Big(|\Xi_{ij}|>\frac{\epsilon}{2}\tilde{\lambda}_2\Big)\geq\frac{\sigma^2 ss_0}{n\Delta}\right) \leq & \frac{n\Delta}{\sigma^2 ss_0}\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot P\Big(|\Xi_{ij}|>\frac{\epsilon}{2}\tilde{\lambda}_2\Big) \\ \leq & \frac{1}{16}(\frac{n\tilde{\lambda}_2^2}{\sigma^2})^2\exp(-\frac{\epsilon^2}{8}\cdot\frac{n\tilde{\lambda}_2^2}{\sigma^2}) \\ = & o(1), \quad \text{as } \Delta\to\infty, \end{aligned} \tag{108}$$

where recall that $\tilde{\lambda}_2=2\sqrt{\frac{8\sigma^2}{n}\Delta}$. Thus the first term in (106) is bounded by

$$\sqrt{\sum_{(i,j)\in S^*}\tilde{\lambda}_2^2\cdot \mathrm{I}\left(|\tilde{H}_{ij}^{t+1}|<\tilde{\lambda}_2\right)}<\frac{2}{\epsilon}\delta\left\|\tilde{\beta}^t-\beta^*\right\|_2+\sqrt{\frac{\sigma^2 ss_0}{n\Delta}}. \tag{109}$$

Second term. Let $s_j=\|\beta^*_{G_j}\|_0$ for $j\in G^*$. For $\forall j\in G^*$, by element-wise beta-min condition $\min\limits_{(i,j)\in S^*}|\beta^*_{ij}|\geq(\sqrt{2}+\epsilon)\tilde{\lambda}_2$ and group-wise beta-min condition $\min\limits_{j\in G^*}\|\beta^*_{G_j}\|_2\geq(\sqrt{2}+\epsilon)\sqrt{s_0}\tilde{\lambda}_2$, we conclude that

$$\|\beta^*_{S_{G_j}}\|_2\geq(\sqrt{2}+\epsilon)\sqrt{s_0\vee s_j}\tilde{\lambda}_2\geq\sqrt{s_j+s_0}\tilde{\lambda}_2+\epsilon\sqrt{s_j\vee s_0}\tilde{\lambda}_2.$$

Therefore, we have

$$\begin{aligned} & \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\left(\tilde{H}_{kj}^{t+1}\right)^2<(s_j+s_0)\tilde{\lambda}_2^2\right) \\ \leq & \mathrm{I}\left(\sqrt{\sum_{k\in S^*\cap S_{G_j}}\left(\beta^*_{kj}\right)^2}-\sqrt{\sum_{k\in S^*\cap S_{G_j}}\left(\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle+\Xi_{kj}\right)^2}<\sqrt{s_j+s_0}\tilde{\lambda}_2\right) \\ \leq & \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\left(\langle\Phi_{ij}^\top,\beta^*-\tilde{\beta}^t\rangle+\Xi_{kj}\right)^2>\epsilon^2(s_j\vee s_0)\tilde{\lambda}_2^2\right) \\ \leq & \mathrm{I}\left(\Upsilon\left(S^*\cap S_{G_j},\tilde{\beta}^t\right)+\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2>\frac{\epsilon^2}{2}(s_j\vee s_0)\tilde{\lambda}_2^2\right) \\ \leq & \mathrm{I}\left(\Upsilon\left(S^*\cap S_{G_j},\tilde{\beta}^t\right)>\frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)+\mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2>\frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right), \end{aligned} \tag{110}$$

which yields that

$$
\begin{aligned}
&\sqrt{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\left(\tilde{H}_{kj}^{t+1}\right)^2 < (s_j+s_0)\tilde{\lambda}_2^2\right)}\\
\leq&\sqrt{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\Upsilon\left(S^*\cap S_{G_j},\tilde{\beta}^t\right) > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)}\\
&+\sqrt{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2 > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)} \qquad (111)\\
\leq&\frac{2}{\epsilon}\sqrt{\Upsilon\left(S^*,\tilde{\beta}^t\right)}+\sqrt{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2 > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)}\\
\leq&\frac{2}{\epsilon}\delta\left\|\tilde{\beta}^t-\beta^*\right\|_2+\sqrt{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2 > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)}.
\end{aligned}
$$

Now, based on Lemma 3 and Theorem 2.1 in Hsu et al. (2012), for every $t>0$ and every support group G_j, we obtain that

$$
P\left(\frac{n}{\sigma^2}\|\Xi_{S^*\cap S_{G_j}}\|_2^2 \geq s_j+2(1+\delta)\sqrt{s_jt}+2(1+\delta)t\right)\leq e^{-t}. \qquad (112)
$$

Let $t=\frac{n\epsilon^2(s_j\vee s_0)}{24\sigma^2}\tilde{\lambda}_2^2$. We can show $t>s_j$, as $\Delta\to\infty$. From $\delta\leq\frac{1}{4}$ we obtain

$$
s_j+2(1+\delta)\sqrt{s_jt}+2(1+\delta)t \;\leq\; 6t=\frac{n\epsilon^2(s_j\vee s_0)}{4\sigma^2}\tilde{\lambda}_2^2, \qquad (113)
$$

which implies that $P\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2>\frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)\leq\exp\left(-\frac{n\epsilon^2(s_j\vee s_0)}{24\sigma^2}\tilde{\lambda}_2^2\right)$. Therefore, by Markov inequality, we have

$$
\begin{aligned}
&P\left\{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2 > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)\geq\frac{\sigma^2 ss_0}{n\Delta}\right\}\\
\leq&\frac{n\Delta}{\sigma^2 ss_0}\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot P\left(\sum_{k\in S^*\cap S_{G_j}}\Xi_{kj}^2 > \frac{\epsilon^2}{4}(s_j\vee s_0)\tilde{\lambda}_2^2\right)\\
\leq&\frac{n\Delta}{\sigma^2 ss_0}\sum_{j\in G^*} s_0\tilde{\lambda}_2^2\cdot \exp\left(-\frac{n\epsilon^2(s_j\vee s_0)}{24\sigma^2}\tilde{\lambda}_2^2\right) \qquad (114)\\
\leq&\frac{n\Delta}{\sigma^2}\tilde{\lambda}_2^2\cdot\exp(-\frac{n\epsilon^2 s_0}{24\sigma^2}\tilde{\lambda}_2^2)\\
=&\frac{1}{32}(\frac{n\tilde{\lambda}_2^2}{\sigma^2})^2\exp(-\frac{\epsilon^2 s_0}{24}\cdot\frac{n\tilde{\lambda}_2^2}{\sigma^2})\\
=&o(1),\quad\text{as }\Delta\to\infty.
\end{aligned}
$$

Combining (111) and (114), we bound the second term by

$$\sqrt{\sum_{j\in G^*} s_0\tilde{\lambda}_2^2 \cdot \mathrm{I}\left(\sum_{k\in S^*\cap S_{G_j}} \left(\tilde{H}_{kj}^{t+1}\right)^2 < (s_j+s_0)\tilde{\lambda}_2^2\right)} \leq \frac{2}{\epsilon}\delta\left\|\tilde{\beta}^t-\beta^*\right\|_2 + \sqrt{\frac{\sigma^2 s s_0}{n\Delta}}. \quad (115)$$

Finally, based on (109) and (115), we have

$$P\left(\sqrt{\sum_{(i,j)\in S^*}\left(\tilde{H}_{ij}^{t+1}\right)^2 \mathrm{I}\left((i,j)\notin \tilde{S}^{t+1}\right)} < \frac{4}{\epsilon}\delta\left\|\tilde{\beta}^t-\beta^*\right\|_2 + 2\sqrt{\frac{\sigma^2 s s_0}{n\Delta}}\right) \to 1, \quad (116)$$

as $\Delta\to\infty$.

■

Appendix C: Example of sub-Gaussian random design

Assume $\zeta_1,\zeta_2,\cdots,\zeta_n$ are independent and identically distributed p-dimensional isotropic, sub-Gaussian random vectors, forming a random matrix $Z\in\mathbb{R}^{n\times p}$, whose i-th row Z_i is denoted by ζ_i. In this paper, we consider a random design matrix X, which is generated as follows:

$$X = Z\Sigma^{\frac{1}{2}}, \quad (117)$$

where Σ is the covariance matrix.

According to the theoretical framework of Zhou (2009) and Mendelson et al. (2008), given the vector space $\mathcal{V}\in\mathbb{R}^p$, the key point is to construct the restricted isometric properties between Xv and $\Sigma^{\frac{1}{2}}v$ for $v\in\mathcal{V}$. The empirical process technique plays an important role, and we define Gaussian complexity first:

Definition 28 (Gaussian complexity) *Given a subset $\mathcal{V}\subseteq\mathbb{R}^p$, we define the Gaussian complexity of $\mathcal{V}$ as follows:*

$$\ell^*(\mathcal{V}) \coloneqq \mathbf{E}_g \sup_{\theta\in\mathcal{V}}\left|\sum_{i=1}^p g_i\theta_i\right|,$$

where θ_i is each component of vector θ, and $g_1,g_2,\cdots,g_p$ are independently drawn from $\mathcal{N}(0,1)$ distributions. In particular, given a non-negative definite matrix Σ, we define

$$\tilde{\ell}^*(\mathcal{V}) \coloneqq \ell^*(\Sigma^{\frac{1}{2}}\mathcal{V}) = \mathbf{E}_g\sup_{v\in\mathcal{V}}\left|\langle\Sigma^{\frac{1}{2}}v,g\rangle\right| = \mathbf{E}_g\sup_{v\in\mathcal{V}}\left|\langle v,\Sigma^{\frac{1}{2}}g\rangle\right|.$$

According to the homogeneity of the norm, we only need to consider the subset of the unit ball sphere S^{p-1}, which is defined as:

$$S^{p-1} \coloneqq \{v\in\mathbb{R}^p : \|v\|_2 = 1\}.$$

The main technique we use is the following empirical process result:

Lemma 29 (Theorem 2.1 in Mendelson et al. (2008)) *Let $1 \leq n \leq p$ and $0 < \delta < 1$. Let $\zeta \in \mathbb{R}^p$ be an isotropic sub-Gaussian random vector with parameter α. Let $\zeta_1, \ldots, \zeta_n$ be the independent copies of ζ. Define X as the random matrix in (117), and let $\mathcal{V}$ satisfy $\Sigma^{\frac{1}{2}} v \in S^{p-1}$ for all $v \in \mathcal{V}$. If sample size n satisfies $n > c'\alpha^4\delta^2\tilde{\ell}^*(\mathcal{V})^2$, then with probability of at least $1 - \exp(-\bar{c}\delta^2 n/\alpha^4)$, for all $v \in \mathcal{V}$, we have*

$$1 - \delta \leq \|Xv\|_2/\sqrt{n} \leq 1 + \delta,$$

where $c', \bar{c} > 0$ are some absolute constants.

Denote parameter space $\mathcal{V} := \Theta^{m,d}(s, s_0) \cap \{v : \Sigma^{\frac{1}{2}} v \in S^{p-1}\}$. Then, given any $v \in \mathcal{V}$, we assume that

$$\rho_{\min} \leq \frac{\left\|\Sigma^{1/2} v\right\|_2}{\|v\|_2} \leq \rho_{\max}.$$

Next, we derive the Gaussian complexity $\tilde{\ell}^*(\mathcal{V})$ for the double sparse structure. We denote

$$U := \Theta^{m,d}(s, s_0) \cap \{v : \|\Sigma^{\frac{1}{2}} v\|_2 \leq 1\}.$$

Recall that $m \times d = p$. Then, we have

$$\begin{aligned}
\tilde{\ell}^*(\mathcal{V}) \leq \widetilde{\ell}^*(U) &= \mathbf{E}_g \sup_{t \in U} \left| \left\langle t, \Sigma^{1/2} g \right\rangle \right| \\
&\leq 3\sqrt{\log |\Theta^{m,d}(s, s_0)|} \sup_{t \in U} \sqrt{\mathbf{E}_g \left| \langle t, \Sigma^{1/2} g \rangle \right|^2} \\
&\leq C\sqrt{s s_0 \log(ed/s_0) + s \log(em/s)} \sup_{t \in U} \left\| \Sigma^{1/2} t \right\|_2 \\
&\leq C\sqrt{s s_0 \log(ed/s_0) + s \log(em/s)},
\end{aligned}$$

where the first inequality follows from Chapter 3 in Ledoux and Talagrand (1991). Note that

$$\frac{\|Xv\|_2}{\sqrt{n}\|v\|_2} = \frac{\|Xv\|_2}{\sqrt{n}\|\Sigma_S^{\frac{1}{2}} v\|_2} \cdot \frac{\|\Sigma_S^{\frac{1}{2}} v\|_2}{\|v\|_2}.$$

Therefore, by Lemma 29, for $n > C'\alpha^4\delta^2 \cdot (s s_0 \log(ed/s_0) + s \log(em/s))$, we have

$$(1 - \delta)\rho_{\min} \leq \frac{\|Xv\|_2}{\sqrt{n}\|v\|_2} \leq (1 + \delta)\rho_{\max}.$$

This proves the satisfaction of the DSRIP condition under the sub-Gaussian random design.

References

Pierre C Bellec. The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv preprint arXiv:1804.01230*, 2018.

Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets Lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603 – 3642, 2018. doi: 10.1214/17-AOS1670. URL `https://doi.org/10.1214/17-AOS1670`.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 04 2016. doi: 10.1214/15-AOS1388. URL https://doi.org/10.1214/15-AOS1388.

Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Thomas Blumensath and Mike E Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of selected topics in signal processing*, 4 (2):298–309, 2010.

Patrick Breheny. The group exponential lasso for bi-level variable selection. *Biometrics*, 71 (3):731–740, 2015.

Patrick Breheny and Jian Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

Patrick Breheny and Jian Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2): 173–187, 2015.

Cristina Butucea, Mohamed Ndaoud, Natalia A. Stepanova, and Alexandre B. Tsybakov. Variable selection with Hamming loss. *The Annals of Statistics*, 46(5):1837 – 1875, 2018. doi: 10.1214/17-AOS1572. URL https://doi.org/10.1214/17-AOS1572.

T. Tony Cai, Anru R. Zhang, and Yuchen Zhou. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory*, 68(9):5975–6002, 2022. doi: 10.1109/TIT.2022.3175455.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Annie P Chiang, John S Beck, Hsan-Jan Yen, Marwan K Tayeh, Todd E Scheetz, Ruth E Swiderski, Darryl Y Nishimura, Terry A Braun, Kwang-Youn A Kim, Jian Huang, et al. Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292, 2006.

Thomas M. Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

Yonina C. Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009. doi: 10.1109/TIT.2009.2030471.

Yonina C. Eldar, Patrick Kuppinger, and Helmut Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6): 3042–3054, 2010. doi: 10.1109/TSP.2010.2044837.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011. doi: 10.1198/jasa.2011.tm09779. URL https://doi.org/10.1198/jasa.2011.tm09779.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

Christophe Giraud. *Introduction to high-dimensional statistics*. CRC Press, 2021.

Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-Dimensional Regression with Unknown Variance. *Statistical Science*, 27(4):500 – 518, 2012. doi: 10.1214/12-STS398. URL https://doi.org/10.1214/12-STS398.

Meiling Hao, Lianqiang Qu, Dehan Kong, Liuquan Sun, and Hongtu Zhu. Optimal minimax variable selection for large-scale matrix linear regression model. *Journal of Machine Learning Research*, 22(147):1–39, 2021. URL http://jmlr.org/papers/v22/19-969.html.

Hussein Hazimeh, Rahul Mazumder, and Peter Radchenko. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *The Annals of Statistics*, 51(1):1 – 32, 2023. doi: 10.1214/21-AOS2155. URL https://doi.org/10.1214/21-AOS2155.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

Jian Huang, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.

Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282 – 2313, 2010. doi: 10.1214/09-AOS781. URL https://doi.org/10.1214/09-AOS781.

Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu. A constructive approach to l_0 penalized regression. *Journal of Machine Learning Research*, 19(10):1–37, 2018. URL http://jmlr.org/papers/v19/17-194.html.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978 – 2004, 2010. doi: 10.1214/09-AOS778. URL https://doi.org/10.1214/09-AOS778.

Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12(103):3371–3412, 2011. URL http://jmlr.org/papers/v12/huang11b.html.

Yasutoshi Ida, Yasuhiro Fujiwara, and Hisashi Kashima. Fast sparse group lasso. In H. Wallach, H. Larochelle, A.Beygelzimer, F.d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d240e3d38a8882ecad8633c8f9c78c9b-Paper.pdf.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems*, 27, 2014.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Zhifan Li, Yanhang Zhang, and Jianxin Yin. Estimating double sparse structures over $\ell_u(\ell_q)$-balls: Minimax rates and phase transition. *IEEE Transactions on Information Theory*, 70(10):7066–7088, 2024. doi: 10.1109/TIT.2024.3451512.

Xiaoxuan Liang, Aaron Cohen, Anibal Sólon Heinsfeld, Franco Pestilli, and Daniel J. McDonald. sparsegl: An r package for estimating sparse group lasso. *Journal of Statistical Software*, 110(6):1–23, 2024. doi: 10.18637/jss.v110.i06. URL https://www.jstatsoft.org/index.php/jss/article/view/v110i06.

Haoyang Liu, Chao Gao, and Richard J. Samworth. Minimax rates in sparse, high-dimensional change point detection. *The Annals of Statistics*, 49(2):1081 – 1112, 2021. doi: 10.1214/20-AOS1994. URL https://doi.org/10.1214/20-AOS1994.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39 (4):2164 – 2204, 2011. doi: 10.1214/11-AOS896. URL https://doi.org/10.1214/11-AOS896.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28:277–289, 2008.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Mohamed Ndaoud. Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach. *arXiv preprint arXiv:2008.12236*, 2020.

Wenliang Pan, Xueqin Wang, Weinan Xiao, and Hongtu Zhu. A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114(526):928–937, 2019. doi: 10.1080/01621459.2018.1462709. URL https://doi.org/10.1080/01621459.2018.1462709. PMID: 31692981.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011. doi: 10.1109/TIT.2011.2165799.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(13):389–427, 2012. URL http://jmlr.org/papers/v13/raskutti12a.html.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865. URL https://doi.org/10.1214/ECP.v18-2865.

Todd E. Scheetz, Kwang-Youn A. Kim, Ruth E. Swiderski, Alisdair R. Philp, Terry A. Braun, Kevin L. Knudtson, Anne M. Dorrance, Gerald F. DiBona, Jian Huang, Thomas L. Casavant, Val C. Sheffield, and Edwin M. Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0602562103. URL https://www.pnas.org/content/103/39/14429.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. doi: https://doi.org/10.1111/j.1467-9868.2005.00532.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x.

Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564 – 2593, 2016. doi: 10.1214/15-AOS1422. URL https://doi.org/10.1214/15-AOS1422.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018. URL http://jmlr.org/papers/v18/14-415.html.

Xiao-Tong Yuan, Bo Liu, Lezi Wang, Qingshan Liu, and Dimitris N. Metaxas. Dual iterative hard thresholding. *Journal of Machine Learning Research*, 21(152):1–50, 2020. URL http://jmlr.org/papers/v21/18-487.html.

Yangjing Zhang, Ning Zhang, Defeng Sun, and Kim-Chuan Toh. An efficient hessian based algorithm for solving large-scale sparse group lasso problems. *Mathematical Programming*, 179:223–263, 2020.

Yanhang Zhang, Junxian Zhu, Jin Zhu, and Xueqin Wang. A splicing approach to best subset of groups selection. *INFORMS Journal on Computing*, 35(1):104–119, 2023. doi: 10.1287/ijoc.2022.1241. URL https://doi.org/10.1287/ijoc.2022.1241.

Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33117–33123, 2020. ISSN 0027-8424. doi: 10.1073/pnas.2014241117. URL https://www.pnas.org/content/117/52/33117.