Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference

T. Tony Cai[1],   Anru Zhang[2],   and   Yuchen Zhou[2]

Abstract

In this paper, we study sparse group Lasso for high-dimensional double sparse linear regression, where the parameter of interest is simultaneously element-wise and group-wise sparse. This problem is an important instance of the simultaneously structured model – an actively studied topic in statistics and machine learning. In the noiseless case, we provide matching upper and lower bounds on sample complexity for the exact recovery of sparse vectors and for stable estimation of approximately sparse vectors, respectively. In the noisy case, we develop upper and matching minimax lower bounds for estimation error. We also consider the debiased sparse group Lasso and investigate its asymptotic property for the purpose of statistical inference. Finally, numerical studies are provided to support the theoretical results.

KEYWORDS: approximate dual certificate, convex optimization, sparsity, sparse group Lasso, simultaneously structured model.

1 Introduction

Consider the *high-dimensional double sparse regression* with simultaneously group-wise and element-wise sparsity structures

$$y = X\beta^* + \varepsilon, \quad \text{or equivalently} \quad y_i = X_i^\top \beta^* + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Here, the covariates $X \in \mathbb{R}^{n \times p}$ and parameter β^* are divided into d known groups, where the ith group contains b_i variables,

$$X = [X_{(1)} \; \cdots \; X_{(d)}], \quad \beta^* = \left((\beta_{(1)}^*)^\top, \cdots (\beta_{(d)}^*)^\top \right)^\top, \quad X_{(j)} \in \mathbb{R}^{n \times b_i}, \beta_{(j)}^* \in \mathbb{R}^{b_i}; \tag{2}$$

β^* is a (s, s_g)-*sparse vector* in the sense that

$$\|\beta^*\|_{0,2} := \sum_{j=1}^d 1_{\{\beta_{(j)}^* \neq 0\}} \leq s_g \quad \text{and} \quad \|\beta^*\|_0 := \sum_{i=1}^p 1_{\{\beta_i^* \neq 0\}} \leq s, \tag{3}$$

The focus of this paper is on the estimation of and inference for β^* based on (y, X). This problem has great importance in a variety of applications. For example in genome-wide association studies (GWAS) [1], the genes can be grouped into pathways and it is believed that only a small portion of the pathways contain causal single nucleotide polymorphisms (SNPs), and the number of causal SNPs is much less than the one of non-causal SNPs in a causal pathway. The sparse group Lasso has been applied to identify causal genes or SNPs associated with a certain trait [1]. Other examples include cancer diagnosis and therapy [2, 3], classification [4], and climate prediction [5] among many others. The problem can also be viewed as a prototype of various problems in statistics and machine learning, such as the sparse multiple response regression [6] and multiple task learning [7, 8, 9].

The sparse group Lasso [10, 11, 12] provides a classic and straightforward estimator for β^*:

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \lambda_g\|\beta\|_{1,2}. \tag{4}$$

Here, $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ and $\|\beta\|_{1,2} = \sum_j \|\beta_{(j)}\|_2$ are ℓ_1 and $\ell_{1,2}$ convex regularizers to account for element-wise and group-wise sparsity structures, respectively. $\lambda, \lambda_g \geq 0$ are tuning parameters. In the noiseless setting that $\varepsilon = 0$, one can apply the constrained $\ell_1 + \ell_{1,2}$ minimization instead to estimate β^*:

$$\begin{aligned} \hat{\beta} \quad &= \quad \arg\min \quad \lambda\|\beta\|_1 + \lambda_g\|\beta\|_{1,2} \\ &\qquad \text{subject to} \quad y = X\beta. \end{aligned} \tag{5}$$

In fact, when λ, λ_g tend to zero while λ/λ_g is fixed as a constant, the sparse group Lasso (4) tends to the $\ell_1 + \ell_{1,2}$ minimization (5).

When β^* is only element-wise sparse, the regular Lasso [13]

$$\hat{\beta}^L = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{6}$$

can be applied and its theoretical properties have been well studied in past decades [14]. When β^* is only group-wise sparse, the group Lasso

$$\hat{\beta}^{GL} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_g\|\beta\|_{1,2} \tag{7}$$

and its variations have been widely investigated [15, 16, 17]. However, to estimate the simultaneously element-wise and group-wise sparse vector β^*, despite many empirical successes of sparse group Lasso in practice, the theoretical properties, including optimal rate of convergence and sample complexity, are still unclear so far to the best of our knowledge.

1.1 Simultaneously Structured Models

More broadly speaking, the simultaneously structured models, i.e., the parameter of interest has multiple structures at the same time, have attracted enormous attention in many fields including statistics, applied mathematics, and machine learning. In addition to the high-dimensional double sparse regression, other simultaneously structured models include sparse principal component analysis [18, 19], tensor singular value decomposition [20, 21], simultaneously sparse and low-rank matrix/tensor recovery [22, 23], sparse matrix/tensor SVD [24], sparse phase retrieval [25, 26], etc. As shown by the seminal work of Oymak, Jalali, Fazel, Yonina, and Hassibi [27, 22], by minimizing multi-objective regularizers with norms associated with these structures (such as ℓ_1 norm for element-wise sparsity, nuclear norm for low-rankness, and total variation norm for piecewise constant structures), one usually cannot do better than applying an algorithm that only exploits one structure. They particularly illustrated that simultaneously sparse and low-rank structured matrix cannot be well estimated by penalizing ℓ_1 and nuclear norm regularizers. Instead, non-convex methods were proposed and shown to achieve better performance.

However based on their results, it remains an open question whether the convex regularization, such as sparse group Lasso or $\ell_1 + \ell_{1,2}$ minimization, can achieve good performance

in estimation of parameter with two types of sparsity structures, such as the aforementioned high-dimensional double sparse regression. Specifically, as illustrated in Section 2.2, a direct application of [22] does not provide a sample complexity lower bound for exact recovery that matches our upper bound.

1.2   Optimality and Related Literature

This paper fills the void of statistical limits of sparse group Lasso and provides an affirmative answer to the aforementioned question: by exploiting both element-wise and group-wise sparsity structures, the $\ell_1 + \ell_{1,2}$ regularization does provide better performance in high-dimensional double sparse regression. Particularly in the noiseless case, it is shown that (s, s_g)-sparse vectors can be exactly recovered and approximately (s, s_g)-sparse vectors can be stably estimated with high probability whenever the sample size satisfies $n \gtrsim s_g \log(d/s_g) + s \log(e s_g b)$, where $b = \max_{1 \leq i \leq d} b_i$. On the other hand, we prove that exact recovery cannot be achieved by $\ell_1 + \ell_{1,2}$ regularization and stable estimation of approximately (s, s_g)-sparse vectors is impossible in general unless $n \gtrsim s_g \log(d/s_g) + s \log(e s_g b/s)$. We then consider the noisy case and develop the matching upper and lower bounds on the convergence rate for the estimation error. Simulation studies are carried out and the results support our theoretical findings. In addition, statistical inference for the individual coordinates of β^* is studied. A confidence interval is constructed based on the debiased sparse group Lasso estimator and its asymptotic property. The results show that by exploring the simultaneously element-wise and group-wise sparsity structures, the debiased sparse group Lasso requires less sample size than the debiased Lasso and debiased group Lasso in the literature [28, 29, 30, 31].

The theoretical analysis of sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization is highly non-trivial. First, the regularizer $\lambda \| \cdot \|_1 + \lambda_g \| \cdot \|_{1,2}$ is not decomposable with respect to the support of β^* so that the classic techniques of decomposable regularizers [32] and null space property [33] may not be suitable here. Despite a substantial body of literature on high-dimensional element-wise sparse vector estimation based on restricted isometry property (RIP) [34, 35, 36, 37, 38] and restricted eigenvalue [14], these techniques cannot provide nearly optimal results for sparse group Lasso here as it is technically difficult to partition general vectors into simultaneously element-

wise and group-wise ones that preserves some ordering structures. Departing from the previous literature, our theoretical analysis relies on a novel construction of approximate dual certificate. See Section 2.3 for further details. Although our results mostly focus on the performance of sparse group Lasso and $\ell_1 + \ell_{1,2}$ estimators, the techniques of approximate dual certificate on multi-norm structures here can also be of independent interest.

The statistical properties of sparse group Lasso and related estimators have been studied previously. For example, [5] developed consistency results for estimators with a general tree-structured norm regularizers, of which the sparse group Lasso is a special case. [39] analyzed the asymptotic behaviors of the adaptive sparse group Lasso estimator. [4, 40] studied the multi-task learning and classification problems based on a variant of sparse group Lasso estimator. [12] studied multivariate linear regression via sparse group Lasso. [41] provided a theoretical framework for developing error bounds of the group Lasso, sparse group Lasso, and group Lasso with tree structured overlapping groups. Specifically, their results imply that the group-wise sparse signal can be exactly recovered with high probability by solving (5) if the sample size satisfies $n \gtrsim s_g (b + \log d)$. Different from previous results, this paper focused on both the required sample size and convergence rate of estimation error of sparse group Lasso. To the best of our knowledge, this is the first paper that provides optimal theoretical guarantees for both the sample complexity and estimation error of sparse group Lasso.

1.3   Organization of the Paper

The rest of the article is organized as follows. After a brief introduction to notation and preliminaries in Section 2.1, the main theoretical results on constrained $\ell_1 + \ell_{1,2}$ minimization in the noiseless setting is presented in Section 2.2 and the key proof ideas are explained in Section 2.3. Results for sparse group Lasso in the noisy setting are discussed in Section 3. In particular, the optimal rate of estimation error and statistical inference are studied in Sections 3.1 and 3.2, respectively. In Section 4, we provide simulation results in both noiseless and noisy cases to justify our theoretical findings. The proofs of technical results are given in Section 6. All technical lemmas and their proofs can be found in Appendix A.

2 $\ell_1 + \ell_{1,2}$ Minimization in Noiseless Case

2.1 Notation and Preliminaries

The following notation will be used throughout the paper. We denote $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$. Let $\text{sgn}(\cdot)$ be the sign function, i.e., $\text{sgn}(x) = 1, 0,$ or -1, if $x > 0, x = 0,$ or $x < 0$, respectively. $H_\alpha(\cdot)$ is the soft-thresholding function such that $H_\alpha(x) = \text{sgn}(x) \cdot \{(|x| - \alpha) \vee 0\}$ for any $x \in \mathbb{R}$. We say $a \lesssim b$ and $a \gtrsim b$ if $a \leq Cb$ and $b \leq Ca$ for some uniform constant $C > 0$, respectively. $a \asymp b$ means $a \lesssim b$ and $a \gtrsim b$ both hold. Let the uppercase $C, C_1, C_0, \ldots$ and lowercase $c, c_1, c_0, \ldots$ denote large and small positive constants respectively, whose actual values vary from time to time. Throughout the paper, we focus on the parameter index set $\{1, \ldots, p\}$ partitioned into d groups. Denote $(1), \ldots, (d) \subseteq \{1, \ldots, p\}$ as the index sets belonging to each group. Additionally, for any group index subset $G \subseteq \{1, \ldots, d\}$, define $(G) = \cup_{j \in G}(j)$, $(G^c) = \cup_{j \notin G}(j)$. For any vector γ and index subset T, $\gamma_T \in \mathbb{R}^{|T|}$ represents the sub-vector of γ with index set T. In particular, $\gamma_{(G)}$ represents the sub-vector of γ in the union of Groups $j \in G$. Define the ℓ_q norm of any vector γ as $\|\gamma\|_q = \left(\sum_i |\gamma_i|^q \right)^{1/q}$. For any vector $\beta \in \mathbb{R}^p$ with group structures, we also define the ℓ_{q_1, q_2} norm for any $0 \leq q_1, q_2 \leq \infty$ as

$$\|\gamma\|_{q_1, q_2} = \left(\sum_{j=1}^d \|\gamma_{(j)}\|_{q_2}^{q_1} \right)^{1/q_1} = \left\{ \sum_{j=1}^d \left(\sum_{i \in (j)} |\gamma_i|^{q_2} \right)^{q_1/q_2} \right\}^{1/q_1}.$$

In particular, $\|\gamma\|_{0,2} = \sum_{j=1}^d 1_{\{\gamma_{(j)} \neq 0\}}$ is the number of non-zero groups of γ, $\|\gamma\|_{\infty,2} = \max_j \|\gamma_{(j)}\|_2$ is the maximum ℓ_2 norm among all groups of γ, and $\|\gamma\|_{1,2} = \sum_{j=1}^d \|\gamma_{(j)}\|_2$ is the group-wise ℓ_1 penalty. With a slight abuse of notation, we simply denote $\|\gamma_T\|_{q_1, q_2} = \|u\|_{q_1, q_2}$ if $u \in \mathbb{R}^p$, u restricted on subset T is γ_T and u restricted on T^c is 0.

The focus of this paper is on simultaneously element-wise and group-wise sparse vectors defined as follows.

Definition 1 (Simultaneous element-wise and group-wise sparsity) *Assume $\beta^* \in \mathbb{R}^p$ is associated with group partition $(1), \ldots, (d)$. For positive integers s, s_g satisfying $s_g \leq d$ and $s_g \leq s \leq \max_{\Omega \subseteq \{1, \ldots, d\}, |\Omega| = s_g} \sum_{i \in \Omega} b_i$, we say β^* is (s, s_g)-sparse if*

$$\|\beta^*\|_{0,2} = \sum_{j=1}^d 1_{\{\beta_{(j)}^* \neq 0\}} \leq s_g, \quad \|\beta^*\|_0 = \sum_i 1_{\{\beta_i^* \neq 0\}} \leq s.$$

6

2.2   Noiseless Case and Sample Complexity

To analyze the performance of sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization, we first introduce the following assumption on the design matrix X.

Assumption 1 (Sub-Gaussian assumption) *Suppose all rows of X are i.i.d. centered sub-Gaussian distributed. Specifically, $\mathbb{E}X_{i\cdot} = 0, \mathrm{Var}(X_{i\cdot}^\top) = \Sigma$, and for any $\alpha \in \mathbb{R}^p$, we have $\mathbb{E}\exp\left(\alpha^\top \Sigma^{-1/2} X_{i\cdot}^\top\right) \le \exp\left(\kappa^2 \|\alpha\|_2^2/2\right)$ for constant $\kappa > 0$. We also assume $\frac{2}{3} \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le \frac{3}{2}$, where $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ are the largest and smallest eigenvalues of Σ, respectively.*

Clear, a random matrix X with i.i.d. standard normal entries satisfies this assumption – this design is referred to as the Gaussian ensemble and has been considered as a benchmark setting in compressed sensing and high-dimensional regression literature [42, 43].

The following theorem shows that the $\ell_1 + \ell_{1,2}$ minimization achieves the exact recovery with high probability when β^* is simultaneously element-wise and group-wise sparse, X is weakly dependent, and Assumption 1 holds. The theorem also provides a more general upper bound on estimation error if β^* is approximately element-wise and group-wise sparse.

Theorem 1 ($\ell_1 + \ell_{1,2}$ minimization in noiseless case) *Suppose one observes $y = X\beta^*$, where X has the group structure (2) and satisfies Assumption 1, β^* is (s, s_g)-sparse, and $b = \max_{1 \le i \le d} b_i$. Let T be the support of β^*. Suppose there exist uniform constants $C, c > 0$ such that*

$$n \ge C\left(s_g \log(d/s_g) + s \log(es_g b)\right), \tag{8}$$

$$\max_{i \in T^c} \left\|\Sigma_{i,T} \Sigma_{T,T}^{-1}\right\|_2 \le c/\sqrt{s}, \tag{9}$$

then the constrained $\ell_1 + \ell_{1,2}$ minimization (5) with $\lambda_g = \sqrt{s/s_g}\lambda$ achieves the exact recovery with probability at least $1 - C\exp(-cn/s)$.

Moreover, if $\beta^* \in \mathbb{R}^p$ is a general vector and $\hat{\beta}$ is the solution to the constrained $\ell_1 + \ell_{1,2}$ minimization (5) with $\lambda_g = \sqrt{s/s_g}\lambda$, then

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \min_{\substack{\|\beta_S^*\|_0 \le s, \|\beta_S^*\|_{0,2} \le s_g, \\ S: \max_{i \in S^c} \|\Sigma_{i,S}\Sigma_{S,S}^{-1}\|_2 \le c/\sqrt{s}}} \left(\frac{1}{\sqrt{s}}\|\beta_{S^c}^*\|_1 + \frac{1}{\sqrt{s_g}}\|\beta_{S^c}^*\|_{1,2}\right). \tag{10}$$

7

with probability at least $1 - C \exp(-cn/s)$.

Remark 1 (Interpretation and comparison) In Theorem 1, the required sample size for achieving exact recovery contains two terms: $s_g \log(d/s_g)$ and $s \log(es_g b)$. Intuitively speaking, $s_g \log(d/s_g)$ corresponds to the complexity of identifying s_g non-zero groups and $s \log(es_g b)$ corresponds to the complexity of estimating s non-zero elements of β in s_g known groups.

When β^* is only element-wise or group-wise sparse, one can apply respectively the classic ℓ_1 or $\ell_{1,2}$ minimization to recover β^*,

$$\hat{\beta}^{\ell_1} = \arg\min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad y = X\beta, \tag{11}$$

$$\hat{\beta}^{\ell_{1,2}} = \arg\min_{\beta} \|\beta\|_{1,2} \quad \text{subject to} \quad y = X\beta. \tag{12}$$

The ℓ_1 minimization and $\ell_{1,2}$ minimization here are respectively the special form of the regular Lasso and group Lasso (if $\lambda, \lambda_g = 0_+$ in (6) and (7)), respectively. Especially if the group size $b_1 \asymp \cdots \asymp b_d \asymp b$, to ensure exact recovery in the noiseless setting with high probability, (11) requires $n \gtrsim Cs \log(ebd/s)$ [44] and group Lasso requires $n \gtrsim s_g(b + \log(ed/s_g))$. The $\ell_1 + \ell_{1,2}$ minimization (5) has provable advantages over both regular and group Lasso when $b \gg \log(d) \gg \log(es_g b)$ and $s_g b / \log(es_g b) \gg s \gg s_g$.

Next, we consider the sample complexity lower bound. Suppose $b_1 = b_2 = \cdots = b_d$ and $d \geq 2s_g$. Recall that one observes $y = X\beta^*$ without noise and aims to estimate the (s, s_g)-sparse vector β^* based on y and X. As indicated by classic results in compressed sensing [45], with sufficient computing power, the ℓ_0 minimization below achieves exact recovery of β^*

$$\hat{\beta}^{\ell_0} = \arg\min \|\beta\|_0 \quad \text{subject to} \quad X\beta = y \tag{13}$$

as along as X is non-degenerate and $n \geq 2s$. This bound is actually sharp: when $n < 2s$, for any set $T \subseteq \{1, \ldots, db\}$ with cardinality $2s$, one can find a vector γ such that $\text{supp}(\gamma) \subseteq T$ and $X\gamma = 0$. By choosing an appropriate T, we can split the support γ to obtain two (s, s_g)-sparse vectors β_1, β_2 satisfying $\beta_1 + \beta_2 = \gamma$. Then, $X\beta_1 = X(-\beta_2)$ but there is no way to distinguish β_1 and β_2 merely based on X and $y = X\beta_1 = X(-\beta_2)$.

However, the ℓ_0 minimization (13) is computational infeasible in practice while a larger sample size is required for applying more practical methods. The following theorem shows that

by performing the convex ℓ_1 regularization, $\ell_{1,2}$ regularization, or any weighted combination of them, one requires at least $\Omega(s_g \log(d/s_g) + s \log(es_g b/s))$ observations to ensure exact recovery of (s, s_g)-sparse vectors.

Theorem 2 (Sample complexity lower bound for exact recovery) *Suppose $b_1 = \cdots = b_d = b$, $d, b \geq 3$. Suppose X is an n-by-(db) matrix. If every $(2s, 2s_g)$-sparse vector $\beta \in \mathbb{R}^{db}$ is a minimizer of the following programming for some $(\lambda, \lambda_g) \in \{(\lambda, \lambda_g) : \lambda, \lambda_g \geq 0, \lambda + \lambda_g > 0\}$:*

$$\min_z \lambda\|z\|_1 + \lambda_g\|z\|_{1,2} \quad subject\ to \quad Xz = y = X\beta.$$

In other words, if the $\ell_1 + \ell_{1,2}$ minimization exactly recover all $(2s, 2s_g)$-sparse vector β, then we must have $n \gtrsim s_g \log(d/s_g) + s \log(es_g b/s)$.

The following sample complexity lower bound shows that for arbitrary methods, to ensure stable estimation of all approximately sparse vectors, one requires at least $\Omega(s_g \log(d/s_g) + s \log(es_g b/s))$ observations.

Theorem 3 (Sample complexity lower bound for stable estimation) *Suppose $b_1 = \cdots = b_d = b$, $b, d \geq 3$. Assume there exists a matrix $X \in \mathbb{R}^{n \times (bd)}$, a map $\Delta : \mathbb{R}^n \to \mathbb{R}^{bd}$ (Δ may depend on X), and a constant $C > 0$ satisfying*

$$\|\beta - \Delta(X\beta)\|_2 \leq C \left(\frac{\|\beta\|_1}{\sqrt{s}} + \frac{\|\beta\|_{1,2}}{\sqrt{s_g}} \right) \tag{14}$$

for all $\beta \in \mathbb{R}^p$ and some s, s_g satisfying $d \geq s_g, s_g b \geq s \geq s_g$. There exists constants C_0 and c_0 that depend only on C such that whenever $s_g \geq C_0$, we must have

$$n \geq c_0(s_g \log(d/s_g) + s \log(es_g b/s)).$$

Remark 2 (Optimality and comparison with previous results) Theorems 2 and 3 show that the sample complexity upper bound in Theorem 1 is rate-optimal under a weak condition: $\log(es_g b) \asymp \log(es_g b) - \log(s)$ or $\log(d) \geq 2s \log(s)/s_g$. Oymak, et al. [22] provided a general analysis for convex regularization of simultaneously structured parameter estimation. Specifically for the high-dimensional double sparse regression, a direct application of their Theorem 3.2 and Corollary 3.1 implies that if $\ell_1 + \ell_{1,2}$ minimization can exactly recover (s, s_g)-sparse vector β^* with a constant probability, one must have $n \gtrsim s$. We can see that Theorem 2 provides a sharper lower bound on sample complexity.

2.3 Proof Sketches

We briefly discuss the proof sketches of the main technical results in this section. The detailed proofs are postponed to Section 6.

The proof of Theorem 1 is based on a novel dual certificate scheme. The dual certificate [46] has been used in the theoretical analysis for various convex optimization methods in high-dimensional problems, such as matrix completion [47, 48], compressed sensing [42], robust PCA [49], tensor completion [50], etc. The high-dimensional double sparse linear regression exhibits different aspects from these previous works due to the simultaneous sparsity structure. In particular, we can show that if the u_{et} defined below is in the row space of X, it can be used as an exact dual certificate for recovery of (s, s_g)-sparse vector β^*:

$$u_{et} = v_{et} + w_{et} \in \mathbb{R}^p, \quad \begin{cases} (v_{et})_{(j)} = \sqrt{s/s_g}\beta^*_{(j)}/\|\beta^*_{(j)}\|_2, & j \in G; \\ \|(v_{et})_{(j)}\|_2 < \sqrt{s/s_g}, & j \in G^c; \end{cases} \quad \begin{cases} (w_{et})_T = \operatorname{sgn}(\beta^*_T) \\ \|(w_{et})_{T^c}\|_\infty < 1. \end{cases}$$

(15)

Here, T and G are the element-wise and group-wise supports of β^*:

$$T = \{i : \beta_i \neq 0\} \subseteq \{1, \ldots, p\}, \quad G = \{j : \beta_{(j)} \neq 0\} \subseteq \{1, \ldots, d\}.$$

Roughly speaking, u_{et} is the sub-gradient of objective function (5) evaluated at $\beta = \beta^*$. If u_{et} is in the row space of X, the sub-gradient will be perpendicular to the feasible set of (5), which implies that β^* is the unique minimizer of $\ell_1 + \ell_{1,2}$ minimization (5).

For more general vector β^* that does not necessarily have a sparse support T or G, we consider the following (s, s_g)-sparse approximation:

$$\beta^{ap} = \arg \min_S \frac{1}{\sqrt{s}}\|\beta^*_{S^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{S^c}\|_{1,2}$$

$$\text{subject to} \quad \|\beta^*_S\|_0 \leq s. \quad \|\beta^*_S\|_{0,2} \leq s_g, \quad \max_{i \in S^c}\|\Sigma_{i,S}\Sigma^{-1}_{S,S}\|_2 \leq c/\sqrt{s}.$$

(16)

Let $T = \{i : \beta^{ap}_i \neq 0\}$ and $G = \{j : (\beta^{ap})_{(j)} \neq 0\}$ be the element-wise and group-wise supports of β^{ap}. Define

$$\widetilde{u}_0 = \widetilde{v}_0 + \widetilde{w}_0 \in \mathbb{R}^p, \quad \begin{cases} (\widetilde{v}_0)_{(j)} = \sqrt{s/s_g}\beta^*_{T,(j)}/\|\beta^*_{T,(j)}\|_2, & j \in G; \\ \|(\widetilde{v}_0)_{(j)}\|_2 < \sqrt{s/s_g}, & j \in G^c; \end{cases} \quad \begin{cases} (\widetilde{w}_0)_T = \operatorname{sgn}(\beta^*_T) \\ \|(\widetilde{w}_0)_{T^c}\|_\infty < 1. \end{cases}$$

(17)

10

Here $\beta^*_{T,(j)} \in \mathbb{R}^{b_j}$ is the subvector β^* restricted on the j-th group with all entries in T^c set to zero. Similarly to the exactly sparse case, if $\widetilde{u}_0$ is in the row space of X and the true β^* is approximately (s, s_g)-sparse, the minimizer of (5) will be close to β^*.

However, it is often difficult to find an exact dual certificate that lies in the row space of X and satisfies stringent conditions in (15) or (17). We instead propose to analyze via the *approximate dual certificate* defined as (18) in the following lemma.

Lemma 1 (Approximate dual certificate for sparse group Lasso) *Suppose T, G are element-wise and group-wise support defined in (16). $\widetilde{u}_0$ is defined in (17). Assume X satisfies $\sigma_{\min}\left(X_T^\top X_T / n\right) \geq 1/2$. If there exists $u \in \mathbb{R}^p$ in the row span of X satisfying*

$$\|u_T - (\widetilde{u}_0)_T\|_2 \cdot \max_{i \in T^c} \left\|X_T^\top X_i / n\right\|_2 \leq 1/8,$$

$$\|H_{1/2}(u_{(G^c)})\|_{\infty,2} \leq \sqrt{s_0}/2, \quad \|u_{(G) \setminus T}\|_\infty \leq 1/2,$$

(18)

Then the conclusion of Theorem 1 (10) holds with probability at least $1 - 2e^{-cn}$. Here, $H_{1/2}(\cdot)$ is the soft-thresholding operator defined at the beginning of Section 2.

If we additionally assume β^* is (s, s_g)-sparse, then β^* is the unique solution to the sparse group $\ell_1 + \ell_{1,2}$ minimization (5) with probability at least $1 - 2e^{-cn}$.

Lemma 1 shows that the conclusion of Theorem 1 holds if there exists an approximate dual certificate u satisfying the condition (18). The following lemma shows that, under the assumptions in Theorem 1, one can find such an approximate dual certificate with high probability.

Lemma 2 *Suppose X has group structure (2) and satisfies Assumption 1. Recall $\sigma_{\min}(X_T^\top X_T / n)$ is the least eigenvalue of $X_T^\top X_T / n$. Then $\sigma_{\min}\left(X_T^\top X_T / n\right) \geq 1/2$ and (18) holds with probability at least $1 - Ce^{-cn/s}$, where T is defined in (16).*

Another key technical tool to the proof of Theorem 1 is the following Lemma, which shows that X satisfies the restricted isometry property for all simultaneously element-wise and group-wise sparse vectors with high probability when there are enough samples.

Lemma 3 *If $n \geq C(s_g \log(d/s_g) + s \log(es_g b))$,*

$$\frac{1}{3}\|\gamma\|_2^2 \leq \frac{1}{n}\|X\gamma\|_2^2 \leq \frac{5}{3}\|\gamma\|_2^2, \quad \forall \gamma \in \{\gamma \in \mathbb{R}^p : \|\gamma\|_0 \leq 2s, \|\gamma\|_{0,2} \leq 2s_g\}$$

(19)

with probability at least $1 - 2e^{-cn}$.

Next we briefly discuss the proof of Theorem 2. Consider the quotient space $\mathbb{R}^{db}/\ker(X) = \{[\gamma] := x + \ker(X), \gamma \in \mathbb{R}^{db}\}$ and define an associated norm as $\|[\gamma]\| = \inf_{v \in \ker(X)}\{\lambda\|\gamma - v\|_1 + \lambda_g\|\gamma - v\|_{1,2}\}$. We show that there exist N different (s, s_g)-sparse vectors $\beta^{(1)}, \ldots, \beta^{(N)}$ such that $\log(N) \asymp s\log(es_g b/s) + s_g\log(d/s_g)$ and $\|[\beta^{(i)}]\| = 1$, $\|[\beta^{(i)}] - [\beta^{(j)}]\| \geq 2/9$ for all $1 \leq i \neq j \leq N$. By a property of the packing number and the fact that $\dim(\mathbb{R}^{db}/\ker(X)) \leq n$, we must have $N \leq 10^n$. Thus $n \gtrsim \log(N) \asymp s\log(es_g b/s) + s_g\log(d/s_g)$.

We prove Theorem 3 by contradiction. Assume that

$$n < c_0\left(s\log(es_g b/s) + s_g\log(d/s_g)\right) \tag{20}$$

for a sufficiently small constant c_0. Let $\|\cdot\| = \|\cdot\|_1 + \sqrt{s/s_g}\|\cdot\|_{1,2}$ and $B = \{x \in \mathbb{R}^{db} : \|x\| \leq 1\}$ be the unit ball associated with $\|\cdot\|$. Define

$$d^n(B, \mathbb{R}^p) = \inf_{\substack{L^n \text{ is a subspace of } \mathbb{R}^p \\ \text{with } \dim(\mathbb{R}^p/L^n) \leq n}} \left\{ \sup_{\beta \in B \cap L^n} \|\beta\|_2 \right\},$$

We have $d^n(B, \mathbb{R}^p) \leq \frac{C}{\sqrt{s}}$ by the assumption of this theorem. We can also show that there exists a uniform constant $c > 0$ such that

$$d^n(B, \mathbb{R}^p) \geq c\min\left\{\frac{1}{\sqrt{s_0}}, \left[\left(\frac{s_g}{s}\log\left(\frac{c\frac{s}{s_g}d\log(es_g b/s)}{n}\right) + \log(es_g b/s)\right)/n\right]^{1/2}\right\}.$$

The previous two inequalities and (20) together imply that

$$n \geq c\left(s_g\log\left(\frac{c\frac{s}{s_g}d\log(es_g b/s)}{n}\right) + s\log(es_g b/s)\right) \geq c_0\left(s\log(es_g b/s) + s_g\log(d/s_g)\right) > n.$$

This contradiction shows that $n \geq c_0\left(s\log(es_g b/s) + s_g\log(d/s_g)\right)$.

3 Sparse Group Lasso in Noisy Case

We now turn to the noisy case.

3.1 Optimal Rate of Estimation Error of Sparse Group Lasso

When observations are noisy, we have the following theoretical guarantee for the sparse group Lasso.

Theorem 4 (Upper bound of estimation error) *Suppose $y = X\beta^* + \varepsilon$, X satisfies Assumption 1, $n \geq C\left(s_g \log(d/s_g) + s \log(es_g b)\right)$ for some uniform constant $C > 0$, $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$, and $b = \max_{1 \leq i \leq d} b_i$. Then the sparse group Lasso estimator (4) with*

$$\lambda = C\sigma\sqrt{(s\log(es_g b) + s_g \log(ed/s_g))n/s} \quad and \quad \lambda_g = \sqrt{s/s_g}\lambda$$

satisfies

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \min_{\substack{S: \|\beta^*_S\|_0 \leq s, \|\beta^*_S\|_{0,2} \leq s_g, \\ \max_{i \in S^c} \|\Sigma_{i,S}\Sigma_{S,S}^{-1}\|_2 \leq c/\sqrt{s}}} \left\{ \sqrt{\frac{\sigma^2(s_g \log(d/s_g) + s\log(es_g b))}{n}} + \frac{\|\beta^*_{S^c}\|_1}{\sqrt{s}} + \frac{\|\beta^*_{S^c}\|_{1,2}}{\sqrt{s_g}} \right\}$$

with probability at least $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}\right)$.

Especially, if β^* is exactly (s, s_g)-sparse and $\max_{i \in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2 \leq c/\sqrt{s}$ holds, then

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{\sigma^2(s_g \log(d/s_g) + s\log(es_g b))}{n} \tag{21}$$

with probability at least $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}\right)$.

In addition, we focus on the following class of simultaneously element-wise and group-wise sparse vectors,

$$\mathcal{F}_{s,s_g} = \{\beta : \|\beta\|_0 \leq s, \|\beta\|_{0,2} \leq s_g\}.$$

The following minimax lower bound of estimation error holds.

Theorem 5 (Lower bound of estimation error) *Suppose X satisfies Assumption 1, $b_1 = \cdots = b_d = b$, and $d, b \geq 3$. Then we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathcal{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \gtrsim \frac{\sigma^2(s_g \log(ed/s_g) + s\log(es_g b/s))}{n}.$$

Remark 3 Theorems 4 and 5 together show that the sparse group Lasso yields the minimax optimal rate of convergence as long as the following condition holds: $\log(es_g b) \asymp \log(es_g b) - \log(s)$ or $\log(d) \gtrsim s\log(s)/s_g$.

Remark 4 We briefly discuss the main proof ideas of Theorem 5 here. First, we randomly generate a series of subsets $\Omega^{(i)} \subseteq \{1, \ldots, p\}$ as feasible supports of (s, s_g)-sparse vectors. Then, we prove by a probabilistic argument that there exist $N \gtrsim (s_g \log(d/s_g) + s \log(es_g b/s))$ subsets $\{\Omega^{(i)}\}_{i=1}^N$ such that $|\Omega^{(i)} \cap \Omega^{(j)}| < 8s_g \lfloor s/s_g \rfloor / 9$ for any $i < j$. Next, we construct a series of candidate (s, s_g)-sparse vectors $\beta^{(i)}$ such that $\beta_k^{(i)} = \tau 1_{\{k \in \Omega^{(i)}\}}$. Intuitively speaking, $\{\beta^{(i)}\}_{i=1}^N$ are non-distinguishable based only on observations (y, X) by such a construction. Theorem 5 then follows by choosing an appropriate τ and the generalized Fano's lemma.

3.2 Statistical Inference via Debiased Sparse Group Lasso

We further consider the statistical inference for β^* under the double sparse linear regression model. First, let $\hat{\beta}$ be the sparse group Lasso estimator given by (4). Inspired by the recent advances in inference for high-dimensional linear regression [28, 51, 29, 31], we propose the following *debiased sparse group Lasso estimator*,

$$\hat{\beta}^u = \hat{\beta} + \frac{1}{n}\hat{M}X^\top \left(Y - X\hat{\beta}\right). \tag{22}$$

Here, $\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^n X_k X_k^\top$ is the sample covariance matrix and $\hat{M} = [\hat{m}_1 \ \cdots \ \hat{m}_p]^\top$ is an approximation of the inverse covariance matrix Σ^{-1}, where $\hat{m}_i$ is the solution to the following convex optimization,

$$
\begin{aligned}
\text{minimize} \quad & m^\top \hat{\Sigma} m \\
\text{subject to} \quad & \|H_\alpha(\hat{\Sigma}m - e_i)\|_{\infty,2} \leq \gamma.
\end{aligned}
\tag{23}
$$

Here, H_α is the soft-thresholding operator with thresholding level α defined at the beginning of Section 2 and e_i is the i-th vector in the canonical basis of $\mathbb{R}^p$. The following theorem establishes an asymptotic result for debiased sparse group Lasso.

Theorem 6 (Asymptotic distribution of debiased sparse group Lasso) *Suppose $\beta^* \in \mathbb{R}^p$ is (s, s_g)-sparse, $X \in \mathbb{R}^{n \times p}$ satisfies Assumption 1, and $\max_{i \in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2 \leq c/\sqrt{s}$. Set $\lambda = C\sigma\sqrt{\frac{(s\log(es_g b) + s_g \log(d/s_g))n}{s}}$ and $\lambda_g = \sqrt{\frac{s}{s_g}}\lambda$ in (4), $\alpha = \frac{\lambda}{n\sigma}, \gamma = \sqrt{\frac{s}{s_g}}\frac{\lambda}{n\sigma}$ in (23). Then with probability at least $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}\right)$, the debiased sparse group Lasso estimator $\hat{\beta}^u$ can be decomposed as $\sqrt{n}\left(\hat{\beta}^u - \beta^*\right) = \Delta + w$, where*

$$\|\Delta\|_\infty \leq \frac{C\left(s\log(es_g b) + s_g \log(ed/s_g)\right)}{\sqrt{n}}\sigma, \quad w|X \sim N\left(0, \sigma^2 \hat{M}\hat{\Sigma}\hat{M}^\top\right). \tag{24}$$

In particular, if $\sqrt{n} \gg s \log(es_g b) + s_g \log(ed/s_g)$, for any $1 \leq i \leq p$,

$$\frac{\sqrt{n}\left(\hat{\beta}_i^u - \beta_i^*\right)}{\sqrt{\hat{m}_i^\top \hat{\Sigma} \hat{m}_i}} \to N\left(0, \sigma^2\right). \tag{25}$$

Remark 5 (25) provides a method to construct confidence intervals for β^*. Specifically if $\hat{\sigma}$ is a consistent estimator of σ, such as the scaled sparse group Lasso to be discussed in Section 5,

$$\left[\hat{\beta}_i^u - \Phi^{-1}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{\hat{m}_i^\top \hat{\Sigma} \hat{m}_i}{n}}, \quad \hat{\beta}_i^u + \Phi^{-1}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{\hat{m}_i^\top \hat{\Sigma} \hat{m}_i}{n}}\right]$$

would be an asymptotic $(1-\alpha)$-confidence interval for β_i^*. We can see that the debiased sparse group Lasso estimator has the provably advantage on sample complexity $(n \gg (s \log(es_g b) + s_g \log(ed/s_g))^2)$ over the ones via debiased Lasso $(n \gg s \log p$, see [28, 29, 31]) or debiased group Lasso $(n \gg (s_g b + s_g \log p)^2$, see [30]) for constructing asymptotic confidence intervals of β^*.

4   Simulation Studies

In this section, we investigate the numerical performance of the sparse group Lasso. The results support our theoretical findings in Sections 2 and 3.

We first study the noiseless case and focus on the sample complexity for the exact recovery. Suppose all group sizes are equal $(b_1 = \cdots = b_d = b)$ and the number of observations n varies from 5 to 200. We consider four simulation designs with (1) $d = 60, b = 20, s_g = 1$; (2) $d = 100, b = 30, s_g = 2$; (3) $d = b = 20, s_g = 1$; and (4) $d = b = 40, s_g = 1$. For each setting, we randomly draw $X \in \mathbb{R}^{n \times db}$ with i.i.d. standard normal entries, construct the fixed vector $\beta^* \in \mathbb{R}^{db}$ satisfying

$$\beta_{(j)}^* = \begin{cases} (1, 2, 3, 4, 5, 0, \ldots, 0) \in \mathbb{R}^b & j = 1, \ldots, s_g; \\ 0 & j = s_g + 1, \ldots, d, \end{cases}$$

and generate $y = X\beta^* = \sum_{j=1}^{s_g} X_{(j)}\beta_{(j)}^*$. We implement the $\ell_1 + \ell_{1,2}$ minimization (5), ℓ_1 minimization (11), and $\ell_{1,2}$ minimization (12). An exact recovery of β^* is considered to be successful if $\|\hat{\beta} - \beta^*\|_2 \leq 10^{-4}$. The successful recovery rate based on 100 replicates is shown in Figure 1. We can see that the sparse group Lasso achieves the best results among all three estimators, which is in line with our theoretical results.

(a) $d = 60, b = 20, s_g = 1$

(b) $d = 100, b = 30, s_g = 2$

(c) $d = 20, b = 20, s_g = 1$
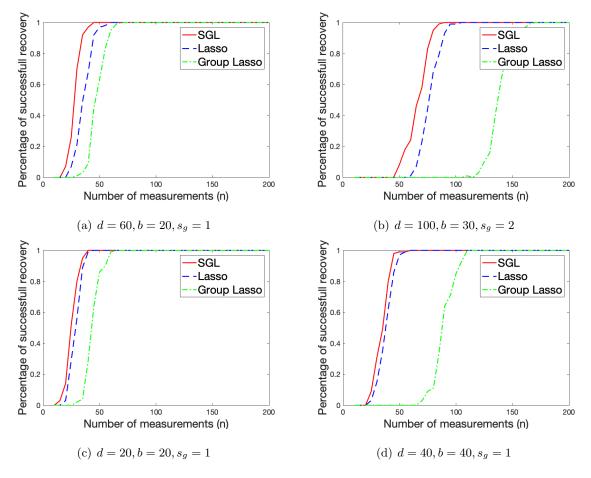
(d) $d = 40, b = 40, s_g = 1$

Figure 1: Exact recovery rate in the noiseless case

Then we consider the noisy case and focus on average estimation errors of different methods. We generate

$$y = X\beta^* + \varepsilon = \sum_{j=1}^{s_g} X_{(j)}\beta^*_{(j)} + \varepsilon,$$

where X, β^* are drawn in the same way as the previous setting and $\varepsilon \overset{iid}{\sim} N(0, 0.1^2)$. We consider four designs: (1) $d = 60, b = 20, s_g = 1$; (2) $d = 100, b = 30, s_g = 2$; (3) $d = b = 20, s_g = 1$; and (4) $d = b = 40, s_g = 2$. For each case, the number of observations n is chosen from an equally spaced sequence from 5 to 200 and the simulation is replicated for 500 times. We compare the following four methods: (1) `SGL_CV`: sparse group Lasso with $\lambda_g = \sqrt{s/s_g}\lambda$ and λ selected via cross validation; (2) `SGL_package`: sparse group Lasso via SGL package[1] in R with the option of automatic tuning parameter selection; (3) `Lasso`: regular Lasso with tuning parameter selected

---

[1] `https://cran.r-project.org/web/packages/SGL/index.html`

via cross validation; (4) `group Lasso`: group Lasso with tuning parameter selected via cross validation. Figure 2 plots the average squared ℓ_2 risks for each design, from which we can clearly see the advantage of `SGL_CV` over the other methods.
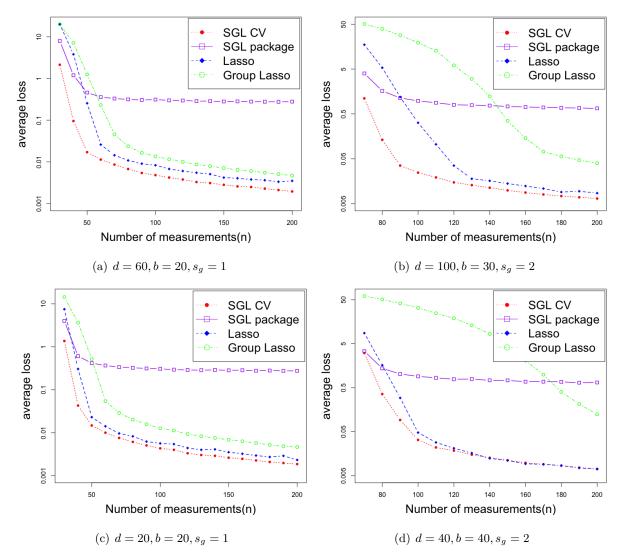

(a) $d = 60, b = 20, s_g = 1$

(b) $d = 100, b = 30, s_g = 2$

(c) $d = 20, b = 20, s_g = 1$

(d) $d = 40, b = 40, s_g = 2$

Figure 2: Average estimation error in the noisy case

5 Discussions

In this paper, we study the high-dimensional double sparse regression and investigate the theoretical properties of the sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization. Particularly, we develop the matching upper and lower bounds on the sample complexity for $\ell_1 + \ell_{1,2}$ minimization in

the noiseless case. We also prove that the sparse group Lasso achieves minimax optimal rate of convergence in a range of settings in the noisy case. Our results give an affirmative answer to the open question for high-dimensional statistical inference for simultaneously structured model: by introducing both ℓ_1 and $\ell_{1,2}$ penalties, one can achieve better performance on estimation and statistical inference for simultaneously element-wise and group-wise sparse vectors.

In addition to β^*, the estimation and inference for noise level σ is another importance task in high-dimensional double sparse regression. Motivated by the recent development of scaled Lasso [52], one may consider the following scaled sparse group Lasso estimator:

$$\{\hat{\beta}^s, \hat{\sigma}\} = \underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \left\{ \frac{\|y - X\beta\|_2^2}{\sigma} + n\sigma + \widetilde{\lambda}\|\beta\|_1 + \widetilde{\lambda}_g\|\beta\|_2 \right\},$$

where $\widetilde{\lambda}$ and $\widetilde{\lambda}_g$ are tuning parameters that do not rely on σ. The consistency of $\hat{\sigma}$ can be established based on similar ideas of scaled Lasso in the literature [52, 29] and the approximate dual certificate in this work.

Moreover, our technical results can be useful in a variety of other problems with simultaneous sparsity structures. For example, [53, 54] considered the estimation of piece-wise constant sparse signals, i.e., both the signal vector and the difference between successive entries of the signal vector are sparse. [55, 56] discussed the estimation of structured parameters where both the number of non-zero elements and the number of distinct values of the parameter vectors are small. [57] considered the estimation of matrices with simultaneous sparsity structures within each block and among different blocks. It is interesting to further study the statistical limits, including the sample complexity and minimax optimal rate of convergence for these problems. In particular, based on the specific sparsity structures of each problem, we can introduce corresponding multi-objective regularizers and the convex regularization methods. The corresponding approximate dual certificates can be proposed, constructed, and analyzed to provide strong theoretical guarantees.

6   Proofs

We collect the proofs of technical results in this section.

6.1 Proof of Lemma 1

Let T satisfy (16). For convenience, we denote $s_0 = s/s_g$ and decompose u as

$$u = v + w, \quad v_i = \begin{cases} u_i - \sqrt{s_0}\beta_i^*/\|\beta_{T,(j)}^*\|_2, & i \in T, i \in (j); \\ u_i, & i \in (G)\backslash T; \\ u_i - H_{1/2}(u_i), & i \in (G^c). \end{cases}$$

$$w_{(j)} = \begin{cases} \sqrt{s_0}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2, & j \in G; \\ H_{1/2}(u_{(j)}), & j \notin G. \end{cases} \tag{26}$$

Note that $|H_{1/2}(x) - x| \leq 1/2$ for any $x \in \mathbb{R}$. Based on the property of (18), $\|u_{(G)\backslash T}\|_\infty \leq 1/2$, then

$$\max_{i \in T^c}|v_i| \leq 1/2, \quad \|v_T - \text{sgn}(\beta_T^*)\|_2 = \|u_T - (\widetilde{u}_0)_T\|_2 \leq \frac{1}{8\max_{i \in T^c}\|X_T^\top X_i/n\|_2}; \tag{27}$$

$$w_{(j)} = \sqrt{s_0}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2, \text{ if } j \in G; \quad \|w_{(j)}\|_2 \leq \sqrt{s_0}/2, \text{ if } j \notin G. \tag{28}$$

Suppose $\hat{\beta}$ is the minimizer to (5), $h = \hat{\beta} - \beta^*$, then based on the sub-differential of $\|\beta\|_1$ and $\|\beta\|_{1,2}$, we have

$$\begin{aligned}
\mathcal{P}(\hat{\beta}) = & \|\hat{\beta}\|_1 + \sqrt{s_0}\|\hat{\beta}\|_{1,2} = \|\beta^* + h\|_1 + \sqrt{s_0}\|\beta^* + h\|_{1,2} \\
\geq & \|\beta_T^*\|_1 + \text{sgn}(\beta_T^*)^\top h_T + \|h_{T^c}\|_1 + \sqrt{s_0}\left(\|\beta_T^*\|_{1,2} + \sum_{j \in G}\frac{\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2} + \sum_{j \notin G}\|h_{(j)}\|_2\right) \\
& - \|\beta_{T^c}^*\|_1 - \sqrt{s_0}\|\beta_{T^c}^*\|_{1,2} \\
\geq & \mathcal{P}(\beta^*) + \|h_{T^c}\|_1 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2} + \text{sgn}(\beta_T^*)^\top h_T + \sum_{j \in G}\frac{\sqrt{s_0}\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2} \\
& - 2\|\beta_{T^c}^*\|_1 - 2\sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}.
\end{aligned} \tag{29}$$

The last inequality comes from $\|\beta^*\|_1 = \|\beta_T^*\|_1 + \|\beta_{T^c}^*\|_1$ and $\|\beta^*\|_{1,2} \leq \|\beta_T^*\|_{1,2} + \|\beta_{T^c}^*\|_{1,2}$.

In particular, given $u^\top h = v^\top h + w^\top h = 0$, we have

$$\operatorname{sgn}(\beta_T^*)^\top h_T + \sum_{j\in G}\frac{\sqrt{s_0}\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2} = \operatorname{sgn}(\beta_T^*)^\top h_T - v^\top h + \sum_{j\in G}\frac{\sqrt{s_0}\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2} - w^\top h$$

$$= -\left(v_T - \operatorname{sgn}(\beta_T^*)\right)^\top h_T - v_{T^c}^\top h_{T^c} - \sum_{j\in G}\left(w_{(j)} - \sqrt{s_0}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2\right)^\top h_{(j)} - (w_{(G^c)})^\top h_{(G^c)}$$

$$\geq -\|v_T - \operatorname{sgn}(\beta_T^*)\|_2\|h_T\|_2 - \|v_{T^c}\|_\infty \cdot \|h_{T^c}\|_1$$

$$\quad - \max_{j\in G}\left\|w_{(j)} - \sqrt{s_0}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2\right\|_2 \cdot \|h_{(G)}\|_{1,2} - \|w_{(G^c)}\|_{\infty,2}\|h_{(G^c)}\|_{1,2}$$

$$\overset{(27)(28)}{\geq} -\|v_T - \operatorname{sgn}(\beta_T^*)\|_2 \cdot \|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2.$$

$$(30)$$

Next note that $h = h_T + h_{T^c}$, we must have $X_T h_T = -X_{T^c}h_{T^c}$, then

$$\|h_T\|_2 = \|(X_T^\top X_T/n)^{-1}X_T^\top X_T h_T/n\|_2 \leq \sigma_{\min}^{-1}(X_T^\top X_T/n)\|X_T X_{T^c}h_{T^c}/n\|_2$$

$$\leq 2 \cdot \max_{i\in T^c}\|X_T^\top X_i/n\|_2 \cdot \|h_{T^c}\|_1.$$

$$(31)$$

Combining (27), (30), and (31), one obtains

$$\operatorname{sgn}(\beta_T^*)^\top h_T + \sum_{j\in G}\frac{\sqrt{s_0}\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2} \geq -3/4 \cdot \|h_{T^c}\|_1 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2.$$

Plug this inequality to (29), we finally have

$$\mathcal{P}(\hat\beta) \geq \mathcal{P}(\beta^*) + \|h_{T^c}\|_1/4 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 - 2\|\beta_{T^c}^*\|_1 - 2\sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}.$$

Since $\hat\beta$ is the minimizer to (5), we must have $\mathcal{P}(\hat\beta) \leq \mathcal{P}(\beta^*)$, then

$$\|h_{T^c}\|_1/4 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 \leq 2\|\beta_{T^c}^*\|_1 + 2\sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}. \qquad (32)$$

If β^* is (s, s_g)-sparse, immediately we have $h_{T^c} = 0$. Then $0 = X_T^\top X h = (X_T^\top X_T)h_T$. By $\sigma_{\min}(X_T^\top X_T/n) \geq 1/2$, we know $X_T^\top X_T/n$ is non-singular, then $h_T = 0$.

Now, we consider the general case. Without loss of generality, suppose $G = \{1, \ldots, g\}$, where $g \leq s_g$. Denote T_1 as the indices of the s largest entries of $h_{(G)\setminus T}$, T_2 as the indices of the s largest entries of $h_{(G)\setminus[T\cup T_1]}$, and so on. For $s_g + 1 \leq i \leq d$, denote $S_{i,1}$ as the indices of the $\lfloor s/s_g \rfloor$ largest entries of $h_{(i)}$, $S_{i,2}$ as the indices of the $\lfloor s/s_g \rfloor$ largest entries of $h_{(i)\setminus S_{i,1}}$, and so on. Let $\widetilde{S}_1, \ldots, \widetilde{S}_{\sum_{i=g+1}^d \lceil b_i/\lfloor s/s_g \rfloor \rceil}$ be an arrangement of $S_{i,j}(1 \leq j \leq \lceil b_i/\lfloor s/s_g \rfloor \rceil, g+1 \leq i \leq d)$

such that $\|h_{\widetilde{S}_1}\|_2^2 \geq \cdots \geq \|h_{\widetilde{S}_{\sum_{i=g+1}^d \lceil b_i/\lfloor s/s_g \rfloor \rceil}}\|_2^2$. Let $R_1 = \cup_{i=1}^{s_g} \widetilde{S}_i$, $R_2 = \cup_{i=s_g+1}^{2s_g} \widetilde{S}_i$, and so on. Then $(T_1, T_2, \ldots, R_1, R_2, \ldots)$ is a partition of T^c, and $|T_i|, |R_j| \leq s$, $|g(T_i)|, |g(R_j)| \leq s_g$, where $g(S) = \{i_1, \ldots, i_k\}$ if $S \subseteq \cup_{j=1}^k (i_j)$ and $S \cap (i_j)$ are not empty for all $1 \leq j \leq k$. Let $\widetilde{T} = T \cup T_1 \cup R_1$. If (19) holds, then

$$\frac{1}{3}\|h_{\widetilde{T}}\|_2^2 \leq \frac{1}{n}\|X_{\widetilde{T}} h_{\widetilde{T}}\|_2^2 = \frac{1}{n}\langle X_{\widetilde{T}} h_{\widetilde{T}}, Xh \rangle - \frac{1}{n}\langle X_{\widetilde{T}} h_{\widetilde{T}}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle. \tag{33}$$

Since $Xh = 0$, we have

$$\langle X_{\widetilde{T}} h_{\widetilde{T}}, Xh \rangle = 0. \tag{34}$$

Now, we consider $\left|\langle X_{\widetilde{T}} h_{\widetilde{T}}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right|$. By triangle inequality,

$$\left|\langle X_{\widetilde{T}} h_{\widetilde{T}}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right| \leq \left|\langle X_T h_T, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right| + \left|\langle X_{T_1} h_{T_1}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right| + \left|\langle X_{R_1} h_{R_1}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right|.$$

The triangle inequality shows that

$$\left|\langle X_T h_T, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right| \leq \sum_{i \geq 2} |\langle X_T h_T, X_{T_i} h_{T_i} \rangle| + \sum_{j \geq 2} \left|\langle X_T h_T, X_{R_j} h_{R_j} \rangle\right|.$$

Combine the parallelogram identity and (19) together, we have

$$|\langle X_T h_T, X_{T_i} h_{T_i} \rangle| \leq \frac{2}{3}n\|h_T\|_2\|h_{T_i}\|_2, \quad \left|\langle X_T h_T, X_{R_j} h_{R_j} \rangle\right| \leq \frac{2}{3}n\|h_T\|_2\|h_{R_j}\|_2.$$

Thus,

$$\left|\langle X_T h_T, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle\right| \leq \frac{2}{3}n\|h_T\|_2\left(\sum_{i \geq 2} \|h_{T_i}\|_2 + \sum_{j \geq 2} \|h_{R_j}\|_2\right). \tag{35}$$

By (3.10) in [35], we have

$$\sum_{i \geq 2} \|h_{T_i}\|_2 \leq s^{-1/2}\|h_{(G) \setminus T}\|_1, \tag{36}$$

and

$$\sum_{j \geq 2} \|h_{R_j}\|_2 = \sum_{j \geq 2} \left(\sum_{i=(j-1)s_g+1}^{js_g} \|h_{\widetilde{S}_i}\|_2^2\right)^{1/2} \leq \sum_{j \geq 2} \sqrt{s_g}\|h_{\widetilde{S}_{(j-1)s_g}}\|_2 \leq \sum_{j \geq 2} \sqrt{s_g} \sum_{i=(j-2)s_g+1}^{(j-1)s_g} \|h_{\widetilde{S}_i}\|_2/s_g$$

$$= s_g^{-1/2}\sum_k \|h_{\widetilde{S}_k}\|_2 = s_g^{-1/2}\sum_{i=g+1}^d \sum_j \|h_{S_{i,j}}\|_2.$$

For all $g + 1 \leq i \leq d$, apply (3.10) in [35] again,

$$\sum_{j \geq 2} \|h_{S_{i,j}}\|_2 \leq (\lfloor s/s_g \rfloor)^{-1/2}\|h_{(i)}\|_1 \leq \sqrt{2}(s/s_g)^{-1/2}\|h_{(i)}\|_1.$$

21

Moreover, by the definition of $S_{i,1}$,

$$\sum_{i=g+1}^{d} \|h_{S_{i,1}}\|_2 \leq \sum_{i=g+1}^{d} \|h_{(i)}\|_2 = \|h_{(G^c)}\|_{1,2}.$$

Therefore,

$$\sum_{j\geq 2} \|h_{R_j}\|_2 \leq s_g^{-1/2} \left(\sum_{i=g+1}^{d} \sqrt{2}(s/s_g)^{-1/2}\|h_{(i)}\|_1 \right) + s_g^{-1/2}\|h_{(G^c)}\|_{1,2} \tag{37}$$

$$= \sqrt{2}s^{-1/2}\|h_{(G^c)}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}.$$

Combine (35), (36) and (37) together, if (19) holds, we have

$$\left|\langle X_T h_T, X_{\widetilde{T}^c} h_{\widetilde{T}^c}\rangle\right| \leq \frac{2}{3}n\|h_T\|_2(s^{-1/2}\|h_{(G)\backslash T}\|_1 + \sqrt{2}s^{-1/2}\|h_{(G^c)}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2})$$

$$\leq \frac{2}{3}n\|h_T\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}).$$

Similarly, if (19) holds, then $\left|\langle X_{T_1} h_{T_1}, X_{\widetilde{T}^c} h_{\widetilde{T}^c}\rangle\right| \leq \frac{2}{3}n\|h_{T_1}\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2})$ and $\left|\langle X_{R_1} h_{R_1}, X_{\widetilde{T}^c} h_{\widetilde{T}^c}\rangle\right| \leq \frac{2}{3}n\|h_{R_1}\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2})$. Thus, with probability at least $1 - 2e^{-cn}$,

$$\left|\langle X_{\widetilde{T}} h_{\widetilde{T}}, X_{\widetilde{T}^c} h_{\widetilde{T}^c}\rangle\right| \leq \frac{2}{3}n\left(\|h_T\|_2 + \|h_{T_1}\|_2 + \|h_{R_1}\|_2\right)(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2})$$

$$\leq \frac{2\sqrt{3}}{3}n\|h_{\widetilde{T}}\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}). \tag{38}$$

The last inequality holds due to Cauchy-Schwarz inequality. Combine (33), (34), (38) and Lemma 3 together, we know that with probability at least $1 - 2e^{-cn}$,

$$\frac{1}{3}\|h_{\widetilde{T}}\|_2^2 \leq \frac{2\sqrt{3}}{3}\|h_{\widetilde{T}}\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}),$$

i.e., with probability at least $1 - 2e^{-cn}$,

$$\|h_{\widetilde{T}}\|_2 \leq 2\sqrt{3}(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}).$$

Finally, by (32), (36), (37) and the previous inequality, with probability at least $1 - 2e^{-cn}$,

$$\|h\|_2 \leq \|h_{\widetilde{T}}\|_2 + \sum_{i\geq 2} \|h_{T_i}\|_2 + \sum_{j\geq 2} \|h_{R_j}\|_2$$

$$\leq 2\sqrt{3}(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}) + \sqrt{2}s^{-1/2}\|h_{T^c}\|_2 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}$$

$$\leq C\left(\frac{1}{\sqrt{s}}\|\beta_{T^c}^*\|_1 + \frac{1}{\sqrt{s_g}}\|\beta_{T^c}^*\|_{1,2}\right).$$

In summary, we have finished the proof of this lemma. $\square$

6.2 Proof of Lemma 2

Let T satisfy (16). Given $\|\beta_T^*\|_{0,2} \leq s_g$, without loss of generally we assume that

$$\beta_{T,(s_g+1)}^*, \cdots, \beta_{T,(d)}^* = 0.$$

We also denote $T_{(j)}$ as the support of $\beta_{T,(j)}^*$. First by Lemma 6 Part 3 with

$$v \in \mathbb{R}^p, v_k = \begin{cases} 1, & k = i; \\ 0, & k \neq i; \end{cases} \quad U \in \mathbb{R}^{p \times |T|} = \mathbb{R}^{(\sum_{i=1}^d b_i) \times |T|}, U_{[T,:]} = I; U_{[T^c,:]} = 0,$$

and notice that $x \log(eu/x) \geq \log(eu)$ for all $1 \leq x \leq u$, we have

$$\begin{aligned}
\mathbb{P}\left(\max_{i \in T^c} \left\|X_T^\top X_i/n\right\|_2 \geq 1/2\right) &\leq \sum_{i \in T^c} \mathbb{P}\left(\left\|X_T^\top X_i/n\right\|_2 \geq 1/2\right) \\
&\leq \sum_{i \in T^c} \mathbb{P}\left(\left\|X_T^\top X_i/n - \mathbb{E}X_T^\top X_i/n\right\|_2 + \|\mathbb{E}X_T^\top X_i/n\|_2 \geq 1/2\right) \\
&\leq \sum_{i \in T^c} \mathbb{P}\left(\left\|X_T^\top X_i/n - \mathbb{E}X_T^\top X_i/n\right\|_2 \geq 1/2 - \|\Sigma_{T,T}\|\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\right) \\
&\leq \sum_{i \in T^c} \mathbb{P}\left(\left\|X_T^\top X_i/n - \mathbb{E}X_T^\top X_i/n\right\|_2 \geq 1/4\right) \\
&\leq db \cdot C \exp\left(Cs - n\right) \leq C \exp\left(\log(d) + \log(b) + Cs - n\right) \\
&\leq C \exp\left(s_g \log(ed/s_g) + s \log(es_g b/s) + Cs - n\right) \leq C \exp(-cn).
\end{aligned} \tag{39}$$

provided that $n \geq C \left(s \log(es_g b/s) + s_g \log(d/s_g)\right)$ for some large constant $C > 0$. By Lemma 7 Part 1, we also know

$$\begin{aligned}
\mathbb{P}\left(\sigma_{\min}(X_T^\top X_T/n) \leq 1/2\right) &\leq \mathbb{P}\left(\|X_T^\top X_T/n - \Sigma_{T,T}\| \geq 1/6\right) \\
&\leq \mathbb{P}\left(\|X_T^\top X_T \Sigma_{T,T}^{-1}/n - I_{|T|}\|\|\Sigma_{T,T}\| \geq 1/6\right) \\
&\leq \mathbb{P}\left(\|X_T^\top X_T \Sigma_{T,T}^{-1}/n - I_{|T|}\| \geq 1/9\right) \\
&\leq C \exp\left(Cs - cn\right) \leq C \exp(-cn).
\end{aligned}$$

Next, we apply the well-regarded golfing scheme [48, 42] to find an approximate dual certificate u that satisfies (18). Let

$$u_0 \in \mathbb{R}^p, \quad (u_0)_{(j)} = \begin{cases} \sqrt{s/s_g}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2 + \text{sgn}(\beta_{T,(j)}^*), & j \in G; \\ 0, & j \in G^c. \end{cases} \tag{40}$$

23

Immediately we have $(u_0)_T = (\tilde{u}_0)_T$. We divide n rows of X into non-overlapping batches, say $X_{[I_1,:]}, X_{[I_2,:]}, \ldots$, with $|I_l| = n_l$. Here, $n_1, n_2, \ldots$ will be specified a little while later. Consider the following sequences

$$
\begin{aligned}
\alpha_0 &= u_0, \\
\gamma_l &= X_{[I_l,:]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T, \quad \alpha_l = \alpha_{l-1} - \gamma_l, \quad l = 1, 2, \ldots, l_{\max}.
\end{aligned}
\tag{41}
$$

Finally the approximate dual certificate is defined as

$$
u = \sum_{l=1}^{l_{\max}} \gamma_l = \sum_{l=1}^{l_{\max}} X_{[I_l,:]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T.
\tag{42}
$$

From the inductive definition we can see

$$
(\alpha_l)_T = (I - X_{[I_l,T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l)(\alpha_{l-1})_T, \quad (\gamma_l)_{T^c} = X_{[I_l,T^c]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T, \quad l = 1, 2, \ldots.
$$

Next, we apply the random matrix results (Lemmas 7 and 6) and obtain the following tail probabilities.

- if $n_l \geq C s t_l$ for large constant $C > 0$ and $t_l \geq C$, by Part 1 of Lemma 7,

$$
\begin{aligned}
&\mathbb{P}\left(\|X_{[I_l,T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l - I_{|T|}\| \geq C\sqrt{st_l/n_l} \right) \\
&\leq C \exp\left(Cs - n_l \min\left\{ \frac{st_l}{n_l}, \left(\frac{st_l}{n_l}\right)^{1/2} \right\} \right) \leq C \exp\left(-cst_l\right);
\end{aligned}
\tag{43}
$$

- Suppose $q_{l-1} = (\alpha_{l-1})_T \in \mathbb{R}^{|T|}$ is independent of $X_{[I_l,:]}$. If $n_l \geq \frac{C(s_0 \log(es_g b/s) + \log d)}{\min\{s_0 \delta_l^2, \sqrt{s_0}\delta_l\}}$ for $\delta_l \geq C \max_{i \in T^c} \|\Sigma_{i,T} \Sigma_{T,T}^{-1}\|_2 \geq C(\max_{i \in T^c} \|\Sigma_{i,T} \Sigma_{T,T}^{-1}\|_2) \|\Sigma_{T,T}^{-1} q_{l-1}\|_2 / \|q_{l-1}\|_2$, by Lemma 7 Part 2,

$$
\begin{aligned}
&\mathbb{P}\left(\max_{j \in G^c} \left\| H_{\|q_{l-1}\|_2 \delta_l} \left(X_{[I_l,(j)]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot q_{l-1} \right) \right\|_2 \geq \sqrt{s_0} \|q_{l-1}\|_2 \delta_l \right) \\
&\leq \sum_{j \in G^c} \mathbb{P}\left(\left\| H_{\|q_{l-1}\|_2 \delta_l} \left(X_{[I_l,(j)]}^\top X_{[I_l,T]} (\Sigma_{T,T}^{-1} q_{l-1})/n_l \right) \right\|_2 \geq \sqrt{s_0} \|q_{l-1}\|_2 \delta_l \right) \\
&\leq d \cdot \binom{b}{\lceil s_0 \rceil} \exp\left(Cs_0 - cn_l \min\left\{ \frac{s_0 \|q_{l-1}\|_2^2 \delta_l^2}{\kappa^4 \|\Sigma_{T,T}^{-1} q_{l-1}\|_2^2}, \frac{\sqrt{s_0} \|q_{l-1}\|_2 \delta_l}{\kappa^2 \|\Sigma_{T,T}^{-1} q_{l-1}\|_2} \right\} \right) \\
&\quad + d \cdot \binom{b}{\lfloor s_0 \rfloor} \exp\left(Cs_0 - cn_l \min\left\{ \frac{s_0 \|q_{l-1}\|_2^2 \delta_l^2}{\kappa^4 \|\Sigma_{T,T}^{-1} q_{l-1}\|_2^2}, \frac{\sqrt{s_0} \|q_{l-1}\|_2 \delta_l}{\kappa^2 \|\Sigma_{T,T}^{-1} q_{l-1}\|_2} \right\} \right) \\
&\leq 2d \cdot \left(\frac{eb}{\lfloor s_0 \rfloor} \right)^{\lceil s_0 \rceil} \exp\left(Cs_0 - cn_l \min\{s_0 \delta_l^2, \sqrt{s_0}\delta_l\} \right) \\
&\leq C \exp\left(\log(d) + Cs_0 \log(2es_g b/s) + Cs_0 - cn_l \min\{s_0 \delta_l^2, \sqrt{s_0}\delta_l\} \right) \\
&\leq C \exp(-cn_l \min\{s_0 \delta_l^2, \sqrt{s_0}\delta_l\});
\end{aligned}
\tag{44}
$$

The third inequality comes from $\|\Sigma_{T,T}^{-1}\| \leq \frac{3}{2}$.

- Suppose $q_{l-1} = (\alpha_{l-1})_T \in \mathbb{R}^{|T|}$ is fixed. If $n_l \min\{\theta_l^2, \theta_l\} \geq C \log(es_g b), \theta_l \geq 2\max_{i \in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2$, by Lemma 6 Part 2,

$$
\mathbb{P}\left(\|X_{[I_l,(G)\setminus T]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l \cdot q_{l-1}\|_\infty \geq \theta_l \|q_{l-1}\|_2\right)
$$

$$
\leq \sum_{i \in (G)\setminus T} \mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l \cdot (\Sigma_{T,T}^{-1}q_{l-1})| \geq \theta_l \|q_{l-1}\|_2\right)
$$

$$
\leq \sum_{i \in (G)\setminus T} \mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l \cdot (\Sigma_{T,T}^{-1}q_{l-1}) - \Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}| \geq \theta_l \|q_{l-1}\|_2 - |\Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}|\right)
$$

$$
\leq \sum_{i \in (G)\setminus T} \mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l \cdot (\Sigma_{T,T}^{-1}q_{l-1}) - \Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}| \geq \theta_l \|q_{l-1}\|_2 - \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\|q_{l-1}\|_2\right)
$$

$$
\leq \sum_{i \in (G)\setminus T} \mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l \cdot (\Sigma_{T,T}^{-1}q_{l-1}) - \Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}| \geq \frac{1}{2}\theta_l \|q_{l-1}\|_2\right)
$$

$$
\leq \sum_{i \in (G)\setminus T} \mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l \cdot (\Sigma_{T,T}^{-1}q_{l-1}) - \Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}| \geq \frac{1}{3}\theta_l \|\Sigma_{T,T}^{-1}q_{l-1}\|_2\right)
$$

$$
\leq s_g b \cdot C \exp\left(-cn_l \min\{\theta_l^2, \theta_l\}\right) = C \exp(\log(s_g b) - cn_l \min\{\theta_l^2, \theta_l\})
$$

$$
\leq C \exp\left(-cn_l \min\{\theta_l^2, \theta_l\}\right).
$$

$$(45)$$

Then we specify $\{n_l, t_l, \delta_l, \theta_l\}_{l \geq 1}$ as follows,

- $n_1 = n_2 \geq C(s\log(es_g b) + s_g \log(d/s_g)),\ t_1 = t_2 = cn_1/(s\log(es)) \geq C,\ \delta_1 = \delta_2 = 1/(16\sqrt{s}),\ \theta_1 = \theta_2 = 1/(16\sqrt{s})$;

- $n_3 = \cdots = n_{l_{\max}} \asymp \frac{n_1}{l_{\max}-2} \geq C(s\log(es_g b) + s_g \log(d/s_g))/\log(es),\ t_3 = \cdots = t_{l_{\max}} = cn_3/s \geq C,\ \delta_3 = \cdots = \delta_{l_{\max}} = \log(es)/(16\sqrt{s}) \geq \max\{(\log(es)/s)^{1/2}/16, \log(es)\sqrt{s_0}/(16s)\},\ \theta_3 = \cdots = \theta_{l_{\max}} = (\log(es)/s)^{1/2}/16,$ with $l_{\max} = \lceil C\log(es) \rceil + 2$.

We can see the following events happen

$$
\|X_{[I_l,T]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l - I_{|T|}\| \leq C\sqrt{st_l/n_l} \leq \sqrt{1/\log(es)}, \quad l = 1,2;
$$

$$
\|X_{[I_l,T]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l - I_{|T|}\| \leq C\sqrt{st_l/n_l} \leq 1/2, \quad l = 3, \ldots, l_{\max};
$$

$$(46)$$

$$\max_{j \in G^c} \left\| H_{\|q_{l-1}\|_2/(16\sqrt{s})} \left(X_{[I_l,(j)]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot q_{l-1} \right) \right\|_2 \leq \sqrt{s_0} \|q_{l-1}\|_2/(16\sqrt{s}), \quad l = 1, 2;$$

$$\max_{j \in G^c} \left\| H_{\|q_{l-1}\|_2 \cdot \log(es)/(16\sqrt{s})} \left(X_{[I_l,(j)]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot q_{l-1} \right) \right\|_2 \tag{47}$$

$$\leq \sqrt{s_0} \|q_{l-1}\|_2 \log(es)/(16\sqrt{s}), \quad l = 3, \ldots, l_{\max};$$

$$\left\| X_{[I_l,(G)\backslash T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot q_{l-1} \right\|_\infty \leq \|q_{l-1}\|_2/(16\sqrt{s}), \quad l = 1, 2$$

$$\left\| X_{[I_l,(G)\backslash T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot q_{l-1} \right\|_\infty \leq \|q_{l-1}\|_2 \cdot (\log(es)/s)^{1/2}/16, \quad l = 3, \ldots, l_{\max}. \tag{48}$$

with probability at least $1 - C \log(es) \exp(-c\frac{n}{\log(es)}) - C \log(es) \exp\left(-c\frac{n}{s_g}\right) - C \log(es) \exp\left(-c\frac{n}{s}\right)$.

By triangle inequality, u_0 satisfies

$$\|u_0\|_2 \leq \sqrt{s/s_g} \left(\sum_{j \in G} \left\| \frac{\beta_{T,(j)}^*}{\|\beta_{T,(j)}^*\|_2} \right\|_2^2 \right)^{1/2} + \|\text{sgn}(\beta_T^*)\|_2 \leq 2\sqrt{s}. \tag{49}$$

When $\max_{i \in T^c} \|X_T^\top X_i/n\|_2 \leq \frac{1}{2}$ and (46)-(49) hold, we have

$$\|q_0\|_2 \leq 2\sqrt{s},$$

$$\|q_1\|_2 = \left\| (I_{|T|} - X_{I_1,T}^\top X_{I_1,T} \Sigma_{T,T}^{-1}/n_1) q_0 \right\| \leq \|I_{|T|} - X_{I_1,T}^\top X_{I_1,T} \Sigma_{T,T}^{-1}/n_1\| \cdot \|q_0\|_2$$

$$\leq 2\sqrt{s/\log(es)}; \tag{50}$$

similarly, $\|q_2\|_2 \leq \|q_1\|_2/\sqrt{\log(es)} \leq 2\sqrt{s}/(\log(es))$;

$$\|q_l\|_2 \leq \|q_{l-1}\|_2/2 \leq \cdots \leq \|q_2\|/2^{l-2} \leq 2^{3-l}\sqrt{s}/(\log(es)), \quad l \geq 3.$$

For large constant $C > 0$, $\|q_{l_{\max}}\|_2 \leq 2^{3-C\log(es)}\sqrt{s}/\log(es) \leq 1/8$. Notice that

$$u_T = (\sum_{l=1}^{l_{\max}} \gamma_l)_T = (\sum_{l=1}^{l_{\max}} (\alpha_{l-1} - \alpha_l))_T = (\alpha_0 - \alpha_{l_{\max}})_T = (\widetilde{u}_0)_T - (q_{l_{\max}})_T,$$

we know that

$$\|u_T - (\widetilde{u}_0)_T\|_2 \cdot \max_{i \in T^c} \|X_T^\top X_i/n\|_2 = \|q_{l_{\max}}\|_2 \cdot \max_{i \in T^c} \|X_T^\top X_i/n\|_2 \leq \frac{1}{8} \cdot \frac{1}{2} < \frac{1}{8}.$$

In addition,

$$\|u_{(G)\backslash T}\|_\infty \leq \sum_{l=1}^{l_{\max}} \left\| X_{[I_l,(G)\backslash T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T \right\|_\infty$$

$$\leq \|q_0\|_2/(16\sqrt{s}) + \|q_1\|_2/(16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \|q_{l-1}\|_2 \cdot (\log(es)/s)^{1/2}/16$$

$$\leq 1/8 + 1/8 + \sum_{l=3}^{\infty} 2^{4-l}/16 \leq 1/2.$$

Since

$$\|q_0\|_2/(16\sqrt{s}) + \|q_1\|_2/(16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \|q_{l-1}\|_2 \cdot \log(es)/(16\sqrt{s})$$

$$\leq \frac{1}{8} + \frac{1}{8} + \sum_{l=3}^{l_{\max}} 2^{4-l}\sqrt{s}/(\log(es)) \cdot \log(es)/(16\sqrt{s}) \leq \frac{1}{2},$$

$$\left\|H_{1/2}(u_{(G^c)})\right\|_{\infty,2} \leq \left\|H_{\|q_0\|_2/(16\sqrt{s})+\|q_1\|_2/(16\sqrt{s})+\sum_{l=3}^{l_{\max}}\|q_{l-1}\|_2\cdot\log(es)/(16\sqrt{s})}(u_{(G^c)})\right\|_{\infty,2}$$

$$\leq \sum_{l=1}^{2} \left\|H_{\|q_{l-1}\|_2/(16\sqrt{s})}\left(X_{[I_l,(G^c)]}^{\top}X_{[I_l,T^c]}q_{l-1}\right)\right\|_{\infty,2}$$

$$+ \sum_{l=3}^{l_{\max}} \left\|H_{\|q_{l-1}\|_2\cdot\log(es)/(16\sqrt{s})}\left(X_{[I_l,(G^c)]}^{\top}X_{[I_l,T^c]}q_{l-1}\right)\right\|_{\infty,2}$$

$$\leq \sum_{l=1}^{2} \sqrt{s_0}\|q_{l-1}\|_2/(16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \sqrt{s_0}\|q_{l-1}\|_2 \cdot \log(es)/(16\sqrt{s})$$

$$\leq \sqrt{s_0}/2.$$

Thus, the construction of u satisfies all required condition in Lemma 1 with probability at least $1 - C\exp\left(-c\frac{n}{s}\right)$. This has finished the proof of this lemma. $\square$

6.3 Proof of Lemma 3

Let $g(S)$ be the group support of set S, that is, $g(S) = \{i_1, \ldots, i_k\}$ if $S \subset \cup_{j=1}^k (i_j)$ and $S \cap (i_j)$ are not empty for all $1 \leq j \leq k$. Lemma 7 Part 1 and the union bound show that

$$\mathbb{P}\left(\exists \gamma \in \mathbb{R}^p, \|\gamma\|_0 \leq 2s, \|\gamma\|_{0,2} \leq 2s_g, \frac{1}{n}\|X\gamma\|_2^2 \notin \left[\frac{1}{3}\|\gamma\|_2^2, \frac{5}{3}\|\gamma\|_2^2\right]\right)$$

$$=\mathbb{P}\left(\exists x \in \mathbb{R}^{2s \wedge p}, S \subseteq 1, \cdots, p, |S| = 2s \wedge p, |g(S)| \leq 2s_g, \frac{1}{n}\|X_S x\|_2^2 \notin \left[\frac{1}{3}\|x\|_2^2, \frac{5}{3}\|x\|_2^2\right]\right)$$

$$\leq \sum_{S \subseteq \{1,\ldots,p\}, |S|=2s \wedge p, |g(S)| \leq 2s_g} \mathbb{P}\left(\forall x \in \mathbb{R}^{2s \wedge p}, \frac{1}{n}\|X_S x\|_2^2 \notin \left[\frac{1}{3}\|x\|_2^2, \frac{5}{3}\|x\|_2^2\right]\right)$$

$$\leq \sum_{S \subseteq \{1,\ldots,p\}, |S|=2s \wedge p, |g(S)| \leq 2s_g} \mathbb{P}\left(\|\frac{1}{n}X_S^\top X_S - \Sigma_{S,S}\| \geq \frac{1}{6}\right)$$

$$\leq \sum_{S \subseteq \{1,\ldots,p\}, |S|=2s \wedge p, |g(S)| \leq 2s_g} \mathbb{P}\left(\|\frac{1}{n}X_S^\top X_S \Sigma_{S,S}^{-1} - I_{|S|}\| \geq \frac{1}{9}\right)$$

$$\leq \left[\binom{d}{2s_g} \vee 1\right]\binom{2s_g b}{2s} \cdot 2\exp\left(Cs - cn\right)$$

$$\leq \left(\frac{ed}{2s_g}\right)^{2s_g}\left(\frac{e \cdot 2s_g b}{2s}\right)^{2s} \cdot 2\exp\left(Cs - cn\right)$$

$$\leq 2\exp\left(2s\log(es_g b/s) + 2s_g\log(ed/s_g) + Cs - cn\right)$$

$$\leq 2e^{-cn}.$$

$\square$

6.4 Proof of Theorem 2

If $d \geq 3s_g$ and $b \geq 3s/s_g$, by (76), we can find $\Omega^{(1)}, \ldots, \Omega^{(N)} \subset \{1, \ldots, db\}$ such that $|\Omega^{(i)}| = s_g\lfloor s/s_g\rfloor$, $|\Omega_{(k)}^{(i)}| = \lfloor s/s_g\rfloor 1_{\{\Omega_{(k)}^{(i)} \text{ is not empty}\}}$ for all $1 \leq i \leq N$, $1 \leq k \leq d$, and

$$\left|\Omega^{(i)} \cap \Omega^{(j)}\right| \leq 8s_g\lfloor s/s_g\rfloor/9, \quad 1 \leq i \neq j \leq N, \tag{51}$$

$$\left|\left\{k \,\middle|\, \left|\Omega_{(k)}^{(i)} \cap \Omega_{(k)}^{(j)}\right| \geq 2\lfloor s/s_g\rfloor/3\right\}\right| \leq 2s_g/3, \quad 1 \leq i \neq j \leq N, \tag{52}$$

where $N = \left\lfloor\left(\frac{d}{2\sqrt{2}s_g}\right)^{s_g/3}\left(\frac{b}{2\sqrt{2}\lfloor s/s_g\rfloor}\right)^{s/9}\right\rfloor$. For any $1 \leq j \leq db, 1 \leq i \leq N$, define

$$\beta_j^{(i)} = \begin{cases} \frac{1}{\lambda s_g\lfloor s/s_g\rfloor + \lambda_g s_g\sqrt{\lfloor s/s_g\rfloor}}, & j \in \Omega^{(i)} \\ 0 & j \notin \Omega^{(i)}, \end{cases}$$

then $\|\beta^{(i)}\|_0 \le s, \|\beta^{(i)}\|_{0,2} \le s_g$. We consider the quotient space

$$\mathbb{R}^{db}/\ker(X) = \left\{ [x] := x + \ker(X), x \in \mathbb{R}^{db} \right\}.$$

Then the dimension of $\mathbb{R}^{db}/\ker(X)$ is $\operatorname{rank}(X) \le n$. Define the norm $\|[x]\| = \inf_{v \in \ker(X)}\{\lambda\|x - v\|_1 + \lambda_g\|x - v\|_{1,2}\}$. For any vector $x \in \mathbb{R}^{db}$ satisfying $\|x\|_0 \le 2s, \|x\|_{0,2} \le 2s_g$, note that $x - v$ with $v \in \ker(X)$ satisfies $X(x - v) = Xx$, by our assumption, we have $\|[x]\| = \lambda\|x\|_1 + \lambda_g\|x\|_{1,2}$. Thus $\|[\beta^{(1)}]\| = \cdots = \|[\beta^{(N)}]\| = 1$. Moreover, by (51) and (52),

$$\begin{aligned}
\|\beta^{(i)} - \beta^{(j)}\|_1 &= \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}} \left(|\Omega^{(i)}| + |\Omega^{(j)}| - 2|\Omega^{(i)} \cap \Omega^{(j)}| \right) \\
&\ge \frac{2s_g \lfloor s/s_g \rfloor}{9(\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor})},
\end{aligned}$$

and

$$\begin{aligned}
\|\beta^{(i)} - \beta^{(j)}\|_{1,2} &= \sum_{k=1}^{d} \|\beta_{(k)}^{(i)} - \beta_{(k)}^{(j)}\|_2 \\
&\ge \sum_{k \in S_{i,j}} \|\beta_{(k)}^{(i)} - \beta_{(k)}^{(j)}\|_2 \\
&\ge \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}} \sqrt{\frac{2\lfloor s/s_g \rfloor}{3}} \cdot |S_{i,j}| \\
&\ge \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}} \sqrt{\frac{2\lfloor s/s_g \rfloor}{3}} \cdot \frac{s_g}{3},
\end{aligned}$$

where $S_{i,j} = \left\{ k | \Omega_{(k)}^{(i)}, \Omega_{(k)}^{(j)} \text{ are not empty sets}, \left| \Omega_{(k)}^{(i)} \cap \Omega_{(k)}^{(j)} \right| < 2\lfloor s/s_g \rfloor/3 \right\}$.
Since $\beta^{(i)} - \beta^{(j)}$ is $(2s, 2s_g)$-sparse,

$$\left\| [\beta^{(i)}] - [\beta^{(j)}] \right\| = \left\| [\beta^{(i)} - \beta^{(j)}] \right\| = \lambda \left\| \beta^{(i)} - \beta^{(j)} \right\|_1 + \lambda_g \left\| \beta^{(i)} - \beta^{(j)} \right\|_{1,2} \ge 2/9.$$

By [44, Proposition C.3], we have $N \le 10^{\operatorname{rank}(X)} \le 10^n$. Therefore we have

$$\left\lfloor \left(\frac{d}{2\sqrt{2}s_g} \right)^{s_g/3} \left(\frac{b}{2\sqrt{2}\lfloor s/s_g \rfloor} \right)^{s/9} \right\rfloor \le 10^n,$$

which means that $n \ge c(s_g \log(d/s_g) + s \log(es_g b/s))$.

If $d < 3s_g$ or $b < 3s/s_g$, let $s_g' = \lceil s_g/3 \rceil \vee 1 \ge s_g/5, s' = \lceil s/15 \rceil \vee s_g'$, then $d \ge 3s_g'$ and $b \ge 3s'/s_g'$, we have $n \ge c(s_g' \log(d/s_g') + s' \log(es_g' b/s')) \ge c(s_g \log(d/s_g) + s \log(es_g b/s))$.

□

29

6.5 Proof of Theorem 3

We would like prove Theorem 3 by contradiction. Let

$$c = \min\left\{\frac{1}{8}, c', \sqrt{\frac{c'}{256}}\right\}, \quad c_0 = \min\left\{\frac{c}{2e}, \frac{c^2}{2C^2}, 16c^2\right\}, \quad C_0 = \max\left\{\frac{C^2}{c^2}, \frac{1}{32c^2}\right\},$$

where c' is a uniform constant such that $n \geq c'(s\log(es_g b/s) + s_g \log(d/s_g))$ if the conditions in Theorem 2 are satisfied. Assume for contradiction that

$$n < c_0(s\log(es_g b/s) + s_g \log(d/s_g)). \tag{53}$$

Let $s_0 = s/s_g$, define the norm $\|\cdot\| = \|\cdot\|_1 + \sqrt{s_0}\|\cdot\|_{1,2}$. Let $B = \{x \in \mathbb{R}^p \|\|x\| \leq 1\}$,

$$d^n(B, \mathbb{R}^p) = \inf_{\substack{L^n \text{ is a subspace of } \mathbb{R}^p \\ \text{with } \dim(\mathbb{R}^p/L^n) \leq n}} \left\{ \sup_{\beta \in B \cap L^n} \|\beta\|_2 \right\}.$$

By [44, Theorem 10.4], we have

$$d^n(B, \mathbb{R}^p) \leq \sup_{\beta \in B} \|\beta - \Delta(X\beta)\|_2 \leq \frac{C}{\sqrt{s}} \sup_{\beta \in B} (\|\beta\|_1 + \sqrt{s_0}\|\beta\|_{1,2}) = \frac{C}{\sqrt{s}}. \tag{54}$$

If

$$d^n(B, \mathbb{R}^p) \geq c\min\left\{ \frac{1}{\sqrt{s_0}}, \left[\left(\frac{s_g}{s} \log\left(\frac{c\frac{s}{s_g} d\log(es_g b/s)}{n}\right) + \log(es_g b/s)\right) / n \right]^{1/2} \right\}, \tag{55}$$

since

$$\frac{C}{\sqrt{s}} \leq \frac{c\sqrt{C_0}}{\sqrt{s}} \leq \frac{c\sqrt{s_g}}{\sqrt{s}} = \frac{c}{\sqrt{s_0}},$$

(54) and (55) together imply that

$$n \geq \frac{c^2}{C^2}\left(s_g \log\left(\frac{c\frac{s}{s_g} d\log(es_g b/s)}{n}\right) + s\log(es_g b/s)\right). \tag{56}$$

By (53),

$$\begin{aligned}
\frac{c\frac{s}{s_g} d\log(es_g b/s)}{n} &> \frac{c\frac{s}{s_g} d\log(es_g b/s)}{c_0(s\log(es_g b/s) + s_g \log(d/s_g))} \\
&\geq 2e\frac{\frac{s}{s_g} d\log(es_g b/s)}{s\log(es_g b/s) + s_g \log(d/s_g)} \\
&\geq \min\left\{ e\frac{\frac{s}{s_g} d\log(es_g b/s)}{s\log(es_g b/s)}, \frac{e\frac{s}{s_g} d\log(es_g b/s)}{s_g \log(ed/s_g)} \right\} \\
&\geq \min\left\{ \frac{ed}{s_g}, \frac{\frac{ed}{s_g}}{\log(\frac{ed}{s_g})} \right\} \geq \left(\frac{ed}{s_g}\right)^{1/2}.
\end{aligned} \tag{57}$$

30

In the last inequality, we used $x^{1/2} \geq \log(x)/2$ for all $x \geq 1$.

Combine (56) and (57) together, we have

$$n \geq \frac{c^2}{2C^2} \left(s \log(es_g b/s) + s_g \log(d/s_g) \right) \geq c_0 \left(s \log(es_g b/s) + s_g \log(d/s_g) \right) > n,$$

contradiction!

Thus, we only need to prove (55) based on (53). We still use the proof of contradiction. If

$$d^n(B, \mathbb{R}^p) < c \min \left\{ \frac{1}{\sqrt{s_0}}, \left[\left(\frac{s_g}{s} \log \left(\frac{c \frac{s}{s_g} d \log(es_g b/s)}{n} \right) + \log(es_g b/s) \right) / n \right]^{1/2} \right\} := \mu,$$

then there exists a subspace L^n of $\mathbb{R}^p$ with $\dim(\mathbb{R}^p/L^n) \leq n$ such that for all $v \in L^n \backslash \{0\}$,

$$\|v\|_2 < \mu \left(\|v\|_1 + \sqrt{s_0} \|v\|_{1,2} \right).$$

Let $B \in \mathbb{R}^{n \times p}$ satisfying $\ker(B) = L^n$. Let $s' = \lfloor \frac{1}{32\mu^2} \rfloor, s_g' = \lfloor s'/s_0 \rfloor$, by (53) and (57),

$$\frac{1}{8} s_0^{-1/2} \geq c s_0^{-1/2} \geq \mu \geq c \min \left\{ \sqrt{\frac{C_0}{s}}, \left(\frac{\frac{s_g}{2s} \log(d/s_g) + \log(es_g b/s)}{c_0 (s_g \log(d/s_g) + s \log(es_g b/s))} \right)^{1/2} \right\} \geq \frac{1}{4\sqrt{2}} s^{-1/2},$$

which means that

$$1 \leq s' \leq s, \quad 1 \leq s_g' \leq s_g.$$

Moreover, we have $\frac{1}{64\mu^2} < s' \leq \frac{1}{32\mu^2}$. For any $(2s', 2s_g')$-sparse β with support set T and group support set G, and $v \in \ker(A)$, by Cauchy-Schwarz inequality,

$$\|v_T\|_1 + \sqrt{s_0} \|v_{(G)}\|_{1,2} \leq \sqrt{2s'} \|v_T\|_2 + \sqrt{s_0} \sqrt{2s_g'} \|v_T\|_2 \leq 2\sqrt{2s'} \|v_T\|_2$$

$$< 2\sqrt{2} \frac{1}{4\sqrt{2}\mu} \mu \left(\|v\|_1 + \sqrt{s_0} \|v\|_{1,2} \right) = \frac{1}{2} \left(\|v\|_1 + \sqrt{s_0} \|v\|_{1,2} \right),$$

i.e.,

$$\|v_T\|_1 + \sqrt{s_0} \|v_{(G)}\|_{1,2} < \|v_{T^c}\|_1 + \sqrt{s_0} \|v_{(G^c)}\|_{1,2}.$$

Based on Cauchy-Schwarz inequality and the sub-differential of $\|\beta\|_1$ and $\|\beta\|_{1,2}$, we have

$$\|\beta + v\|_1 + \sqrt{s_0} \|\beta + v\|_{1,2}$$

$$\geq \|\beta\|_1 + \text{sgn}(\beta)^\top v_T + \|v_{T^c}\|_1 + \sqrt{s_0} \left(\|\beta\|_{1,2} + \sum_{j \in G} \frac{\beta_{(j)}^\top v_{(j)}}{\|\beta_{(j)}\|_2} + \sum_{j \in G^c} \|v_j\|_2 \right)$$

$$\geq \|\beta\|_1 - \|v_T\|_1 + \|v_{T^c}\|_1 + \sqrt{s_0} \left(\|\beta\|_{1,2} - \|v_{(G)}\|_2 + \|v_{(G^c)}\|_2 \right)$$

$$> \|\beta\|_1 + \sqrt{s_0} \|\beta\|_{1,2}.$$

31

By Theorem 2,

$$n \geq c'(s'\log(es_g'b/s') + s_g'\log(d/s_g')) \geq c's'\left\{\log\left(\frac{es_g b}{2s}\right) + \frac{1}{2s_0}\log(s_0 d/s')\right\} \geq c's'\log\left(\frac{es_g b}{2s}\right).$$

Thus

$$
\begin{aligned}
n \geq & c's'\left(\log\left(\frac{es_g b}{2s}\right) + \frac{s_g}{s}\log\left(\frac{c'\frac{s}{s_g}d\log(es_g b/s)}{n}\right)\right) \\
> & \frac{c'}{64\mu^2}\left(\frac{1}{4}\log(es_g b/s) + \frac{s_g}{s}\log\left(\frac{c\frac{s}{s_g}d\log(es_g b/s)}{n}\right)\right) \\
\geq & n
\end{aligned}
$$

provided that $c = \min\left\{\frac{1}{8}, c', \sqrt{\frac{c'}{256}}\right\}$, contradiction! This means that (55) holds if (53) is true.

Therefore, we have finished the proof of Theorem 3. $\square$

6.6 Proof of Theorem 4

Let $\lambda = C\sigma\sqrt{\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}n}$, $\lambda_g = \sqrt{s/s_g}\lambda$. By (82) in Lemma 5 and (97), one has

$$
\begin{aligned}
&\mathbb{P}\left(\left\|H_{\frac{1}{10}\lambda}(X^\top\varepsilon)\right\|_{\infty,2} \geq \frac{1}{10}\lambda_g\right) \\
\leq & \mathbb{P}\left(\exists 1 \leq j \leq d, \left\|H_{\frac{1}{10}\lambda}(X_{(j)}^\top\varepsilon)\right\|_2 \geq \frac{1}{10}\lambda_g, \|\varepsilon\|_2 \geq 5\sqrt{n\sigma^2}\right) + \mathbb{P}\left(\|\varepsilon\|_2 \geq 5\sqrt{n\sigma^2}\right) \\
\leq & \mathbb{P}\left(\exists 1 \leq j \leq d, \left\|H_{\frac{1}{10}\lambda}(X_{(j)}^\top\varepsilon)\right\|_2 \geq \frac{1}{10}\lambda_g\Big|\|\varepsilon\|_2 \geq 5\sqrt{n\sigma^2}\right) + \mathbb{P}\left(\|\varepsilon\|_2 \geq 5\sqrt{n\sigma^2}\right) \\
\leq & d\exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n} \\
= & \exp\left(\log(s_g) + \log(d/s_g) - C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n} \\
\leq & \exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n}.
\end{aligned}
$$

(58)

By the definition of $\hat{\beta}$ and KKT condition, we have

$$X^\top(y - X\hat{\beta}) + \lambda z_1 + \lambda_g z_2 = 0,$$

where

$$
\begin{cases} (z_1)_i = \text{sgn}(\hat{\beta}_i), & \hat{\beta}_i \neq 0; \\ |(z_1)_i| \leq 1, & \hat{\beta}_i = 0; \end{cases} \qquad \begin{cases} (z_2)_{(j)} = \frac{\hat{\beta}_{(j)}}{\|\hat{\beta}_{(j)}\|_2}, & \hat{\beta}_{(j)} \neq 0; \\ \|(z_2)_{(j)}\|_2 \leq 1, & \hat{\beta}_{(j)} = 0. \end{cases}
$$

Therefore,

$$\|H_\lambda(X^\top(X\hat\beta - y))\|_{\infty,2} \le \lambda_g.$$

(58), Lemma 8 Part 1 and the previous inequality together imply that

$$\mathbb{P}\left(\left\|H_{(1+\frac{1}{10})\lambda}(X^\top Xh)\right\|_{\infty,2} \le (1+\frac{1}{10})\lambda_g\right) \ge 1 - \exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right) - e^{-n},$$

(59)

where $h = \hat\beta - \beta^*$. By the definition of $\hat\beta$, we have

$$\|y - X\hat\beta\|_2^2 + \lambda\|\hat\beta\|_1 + \lambda_g\|\hat\beta\|_{1,2} \le \|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1 + \lambda_g\|\beta^*\|_{1,2}.$$

(29) and the previous inequality show that

$$\begin{aligned}
&\|Xh\|_2^2 + \lambda\|h_{T^c}\|_1 + \lambda_g\|h_{(G^c)}\|_{1,2} \\
&\le 2\langle Xh, \varepsilon\rangle - \lambda\cdot\mathrm{sgn}(\beta_T^*)^\top h_T - \lambda_g\sum_{j\in G}\frac{\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2} + 2\lambda\|\beta_{T^c}^*\|_1 + 2\lambda_g\|\beta_{T^c}^*\|_{1,2}.
\end{aligned}$$

(60)

First, we consider $\langle Xh, \varepsilon\rangle$. Denote $P = X_T(X_T^\top X_T)^{-1}X_T^\top$, since $Xh = X_T h_T + X_{T^c}h_{T^c}$ and $(I_n - P)X_T = 0$,

$$\begin{aligned}
|\langle Xh, \varepsilon\rangle| &\le |\langle PXh, \varepsilon\rangle| + |\langle(I_n - P)Xh, \varepsilon\rangle| \\
&= \left|\langle X_T^\top Xh, (X_T^\top X_T)^{-1}X_T^\top\varepsilon\rangle\right| + |\langle(I_n - P)X_{T^c}h_{T^c}, \varepsilon\rangle| \\
&= \left|\langle X_T^\top Xh, (X_T^\top X_T)^{-1}X_T^\top\varepsilon\rangle\right| + |\langle X_{T^c}h_{T^c}, (I_n - P)\varepsilon\rangle|.
\end{aligned}$$

(61)

Therefore, to give an upper bound of $|\langle Xh, \varepsilon\rangle|$, we only need to bound $\left|\langle X_T^\top Xh, (X_T^\top X_T)^{-1}X_T^\top\varepsilon\rangle\right|$ and $|\langle X_{T^c}h_{T^c}, (I_n - P)\varepsilon\rangle|$, respectively. By Part 1 of Lemma 7 and also notice that $\frac{2}{3} \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le \frac{3}{2}$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\left(\frac{1}{n}X_T^\top X_T\right)^{-1}\right\| \ge 2\right) &\le \mathbb{P}\left(\|\frac{1}{n}X_T^\top X_T - \Sigma_{T,T}\| \ge \frac{1}{6}\right) \\
&\le \mathbb{P}\left(\|\frac{1}{n}X_T^\top X_T\Sigma_{T,T}^{-1} - I_s\| \ge \frac{1}{9}\right) \\
&\le 2\exp\left(Cs - cn\right) \le 2\exp\left(-cn\right).
\end{aligned}$$

(62)

(62), Lemma 9 and Cauchy-Schwarz inequality together imply that with probability at least $1 - \exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}\right) - 2\exp(-cn)$,

$$\|(X_T^\top X_T)^{-1}X_T^\top\varepsilon\|_1 \le 2\frac{\sqrt{s}}{n}\|X_T^\top\varepsilon\|_2 \le 2\frac{s}{n}\|X_T^\top\varepsilon\|_\infty \le C\frac{s}{n}\sqrt{n\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}\sigma^2} \le C\frac{s}{n}\lambda,$$

33

$$\|(X_T^\top X_T)^{-1} X_T^\top \varepsilon\|_{1,2} \le \sqrt{s_g} \|(X_T^\top X_T)^{-1} X_T^\top \varepsilon\|_2 \le 2\frac{\sqrt{s_g}}{n}\|X_T^\top \varepsilon\|_2 \le C\frac{\sqrt{s \cdot s_g}}{n}\lambda.$$

Combine Lemma 8 Part 2, (59) and the previous two inequalities together, with probability at least $1 - 2\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}\right) - 3e^{-cn}$,

$$\begin{aligned}
\left|\langle X_T^\top X h, (X_T^\top X_T)^{-1} X_T^\top \varepsilon\rangle\right| \le& \frac{11}{10}\lambda\|(X_T^\top X_T)^{-1} X_T^\top \varepsilon\|_1 + \frac{11}{10}\lambda_g\|(X_T^\top X_T)^{-1} X_T^\top \varepsilon\|_{1,2}\\
\le& C\frac{s}{n}\lambda^2.
\end{aligned} \tag{63}$$

Similarly to the proof of (58), also notice that $\|(I_n - P)\varepsilon\|_2 \le \|\varepsilon\|_2$ and $X_{(G^c)}$ is independent of $I_n - P$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|H_{\frac{1}{10}\lambda}\left(X_{(G^c)}^\top (I_n - P)\varepsilon\right)\right\|_{\infty,2} \ge \frac{1}{10}\lambda_g\right)\\
\le& \mathbb{P}\left(\exists j \in G^c, \left\|H_{\frac{1}{10}\lambda}\left(X_{(j)}^\top (I_n - P)\varepsilon\right)\right\|_2 \ge \frac{1}{10}\lambda_g \middle| \|(I_n - P)\varepsilon\|_2 \ge 5\sqrt{n\sigma^2}\right)\\
&+ \mathbb{P}\left(\|\varepsilon\|_2 \ge 5\sqrt{n\sigma^2}\right)\\
\le& \exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n}.
\end{aligned}$$

By Lemma 8 Part 2 and (58), with probability at least $1 - \exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) - e^{-n}$,

$$\left|\langle X_{(G^c)} h_{(G^c)}, (I_n - P)\varepsilon\rangle\right| = \left|\langle h_{(G^c)}, X_{(G^c)}^\top (I_n - P)\varepsilon\rangle\right| \le \frac{1}{10}\lambda\|h_{(G^c)}\|_1 + \frac{1}{10}\lambda_g\|h_{(G^c)}\|_{1,2}.$$

Notice that $X_{T^c\setminus(G^c)}$ and $I_n - P$ are independent and $|T^c\setminus(G^c)| \le |G| \le s_g b$, by Lemma 9, with probability at least $1 - \exp(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}) - e^{-n}$,

$$\begin{aligned}
\left|\langle X_{T^c\setminus(G^c)} h_{T^c\setminus(G^c)}, (I_n - P)\varepsilon\rangle\right| \le& \|h_{T^c\setminus(G^c)}\|_1 \|X_{T^c\setminus(G^c)}^\top (I_n - P)\varepsilon\|_\infty\\
\le& C\sqrt{n\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}\sigma^2}\|h_{T^c\setminus(G^c)}\|_1\\
\le& \frac{1}{10}\lambda\|h_{T^c\setminus(G^c)}\|_1.
\end{aligned}$$

Combine the previous two inequalities together, we have

$$\begin{aligned}
\left|\langle X_{T^c} h_{T^c}, (I_n - P)\varepsilon\rangle\right| \le& \left|\langle X_{(G^c)} h_{(G^c)}, (I_n - P)\varepsilon\rangle\right| + \left|\langle X_{T^c\setminus(G^c)} h_{T^c\setminus(G^c)}, (I_n - P)\varepsilon\rangle\right|\\
\le& \frac{1}{10}\lambda\|h_{T^c}\|_1 + \frac{1}{10}\lambda_g\|h_{(G^c)}\|_{1,2}
\end{aligned} \tag{64}$$

with probability $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}\right) - Ce^{-cn}$. Combine (61), (63) and (64) together, we know that with probability at least $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s}\right) - Ce^{-cn}$,

$$|\langle X h, \varepsilon\rangle| \le C\frac{s}{n}\lambda^2 + \frac{1}{10}\lambda\|h_{T^c}\|_1 + \frac{1}{10}\lambda_g\|h_{(G^c)}\|_{1,2}. \tag{65}$$

Moreover, by the proof of Theorem 1, with probability at least $1 - C\exp(-cn/s)$, there exists an approximate dual certificate $u \in \mathbb{R}^p$ in the row span of X satisfying (18), and $\|v_T - \mathrm{sgn}(\beta_T^*)\|_2 \leq \frac{1}{8}$, where v is defined in (26). Similarly to (30), we have

$$\mathrm{sgn}(\beta_T^*)^\top h_T + \sum_{j \in G} \frac{\sqrt{s_0}\beta_{T,(j)}^{*\top} h_{(j)}}{\|\beta_{T,(j)}^*\|_2}$$

$$\geq -\|v_T - \mathrm{sgn}(\beta_T^*)\|_2 \cdot \|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 + \langle h, u \rangle$$

$$\geq -\frac{1}{8} \cdot \|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 + \langle h, u \rangle.$$

By Lemma 10, with probability at least $1 - Ce^{-cn/s}$, $u = X^\top w$ with $\|w\|_2 \leq C\sqrt{s/n}$. Therefore, with probability at least $1 - Ce^{-cn/s}$,

$$|\langle h, u \rangle| = |\langle Xh, w \rangle| \leq \|Xh\|_2 \|w\|_2 \leq C\sqrt{s/n}\|Xh\|_2.$$

The two previous inequalities together imply that

$$\mathrm{sgn}(\beta_T^*)^\top h_T + \sum_{j \in G} \frac{\sqrt{s_0}\beta_{T,(j)}^{*\top} h_{(j)}}{\|\beta_{T,(j)}^*\|_2} \geq -\frac{1}{8} \cdot \|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 - C\sqrt{s/n}\|Xh\|_2 \quad (66)$$

with probability at least $1 - Ce^{-cn/s}$.

Combine (60), (65) and (66) together, with probability at least $1 - Ce^{-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}} - Ce^{-cn/s}$,

$$\|Xh\|_2^2 + \frac{3}{10}\lambda\|h_{T^c}\|_1 + \frac{3}{10}\lambda_g\|h_{(G^c)}\|_{1,2}$$
$$\leq C\frac{s}{n}\lambda^2 + \frac{1}{8}\lambda\|h_T\|_2 + C\sqrt{s/n}\lambda\|Xh\|_2 + 2\lambda\|\beta_{T^c}^*\|_1 + 2\lambda_g\|\beta_{T^c}^*\|_{1,2}. \quad (67)$$

By (39), (59) and (62), with probability at least $1 - \exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right) - Ce^{-cn}$,

$$\|h_T\|_2 \leq \|(X_T^\top X_T)^{-1}\|\|X_T^\top X_T h_T\|_2$$

$$\leq \frac{2}{n}\|X_T^\top Xh - X_T^\top X_{T^c} h_{T^c}\|_2$$

$$\leq \frac{2}{n}\left(\|X_T^\top Xh\|_2 + \|X_T^\top X_{T^c} h_{T^c}\|_2\right)$$

$$\leq \frac{2}{n}\left(\|H_{\frac{11}{10}\lambda}(X_T^\top Xh)\|_2 + \frac{11}{10}\sqrt{s}\lambda + n\sum_{i \in T^c}\|X_T^\top X_i/n\|_2 |h_i|\right)$$

$$\leq \frac{2}{n}\left(\sqrt{s_g}\|H_{\frac{11}{10}\lambda}(X_T^\top Xh)\|_{\infty,2} + \frac{11}{10}\sqrt{s}\lambda + n\max_{i \in T^c}\|X_T^\top X_i/n\|_2\|h_{T^c}\|_1\right)$$

$$\leq \frac{2}{n}\left(\sqrt{s_g}\frac{11}{10}\lambda_g + \frac{11}{10}\sqrt{s}\lambda + \frac{n}{2}\|h_{T^c}\|_1\right)$$

$$\leq 5\frac{\sqrt{s}}{n}\lambda + \|h_{T^c}\|_1. \quad (68)$$

The fourth inequality comes from $\|x\|_2 \le \|H_\alpha(x)\|_2 + \sqrt{s}\alpha$ for $x \in \mathbb{R}^s$; the fifth inequality holds since $\|X_T^\top X h\|_{0,2} \le s_g$.

(67) and (68) together imply that

$$\|Xh\|_2^2 + \frac{7}{40}\lambda\|h_{T^c}\|_1 + \frac{3}{10}\lambda_g\|h_{(G^c)}\|_{1,2} \le C\frac{s}{n}\lambda^2 + C\sqrt{s/n}\lambda\|Xh\|_2 + 2\lambda\|\beta_{T^c}^*\|_1 + 2\lambda_g\|\beta_{T^c}^*\|_{1,2}$$

with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$. Also notice that

$$C\sqrt{s/n}\lambda\|Xh\|_2 \le \|Xh\|_2^2 + C\frac{s}{n}\lambda^2,$$

with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$,

$$\|h_{T^c}\|_1 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2} \le C\left(\frac{s}{n}\lambda + \|\beta_{T^c}^*\|_1 + \sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}\right). \tag{69}$$

From the proof of Lemma 1, we know that (33) and (38) hold with probability at least $1 - 2e^{-cn}$. By Lemma 8 Part 2 and (59), with probability at least $1-\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) - e^{-n}$,

$$\begin{aligned}\left|\langle X_{\widetilde{T}}h_{\widetilde{T}}, Xh\rangle\right| = \left|\langle h_{\widetilde{T}}, X_{\widetilde{T}}^\top Xh\rangle\right| &\le \frac{11}{10}\left(\lambda\|h_{\widetilde{T}}\|_1 + \lambda_g\|h_{\widetilde{T}}\|_{1,2}\right) \\ &\le \frac{11}{10}\left(\lambda\cdot\sqrt{3s}\|h_{\widetilde{T}}\|_2 + \lambda_g\sqrt{2s_g}\|h_{\widetilde{T}}\|_2\right) \le 4\lambda\sqrt{s}\|h_{\widetilde{T}}\|_2.\end{aligned} \tag{70}$$

The second inequality is due to $\|h_{\widetilde{T}}\|_0 \le 3s$, $\|h_{\widetilde{T}}\|_{0,2} \le 2s_g$ and Cauchy-Schwarz inequality.

Combine (33), (38), (69) and (70) together, with probability at least $1-C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$, we have

$$\begin{aligned}\frac{1}{3}\|h_{\widetilde{T}}\|_2^2 &\le \frac{1}{n}4\lambda\sqrt{s}\|h_{\widetilde{T}}\|_2 + \frac{2\sqrt{3}}{3}\|h_{\widetilde{T}}\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}) \\ &\le \frac{1}{n}4\lambda\sqrt{s}\|h_{\widetilde{T}}\|_2 + \frac{2\sqrt{3}}{3}\|h_{\widetilde{T}}\|_2\cdot\frac{C}{\sqrt{s}}\left(\frac{s}{n}\lambda + \|\beta_{T^c}^*\|_1 + \sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}\right) \\ &\le C\left(\frac{\sqrt{s}}{n}\lambda + \frac{1}{\sqrt{s}}\|\beta_{T^c}^*\|_1 + \frac{1}{\sqrt{s_g}}\|\beta_{T^c}^*\|_{1,2}\right)\|h_{\widetilde{T}}\|_2.\end{aligned}$$

Therefore, with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$,

$$\|h_{\widetilde{T}}\|_2 \le C\left(\frac{\sqrt{s}}{n}\lambda + \frac{1}{\sqrt{s}}\|\beta_{T^c}^*\|_1 + \frac{1}{\sqrt{s_g}}\|\beta_{T^c}^*\|_{1,2}\right). \tag{71}$$

By (36), (37), (69) and the previous inequality, also notice that $e^{-cn/s} \le e^{-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}}$,

36

with probability at least $1 - C \exp(-C \frac{s \log(es_g b) + s_g \log(d/s_g)}{s})$,

$$\|h\|_2 \leq \|h_{\widetilde{T}}\|_2 + \sum_{i \geq 2} \|h_{T_i}\|_2 + \sum_{j \geq 2} \|h_{R_j}\|_2 \leq \|h_{\widetilde{T}}\|_2 + \sqrt{2} s^{-1/2} \|h_{T^c}\|_2 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2}$$
$$\leq C \left(\frac{\sqrt{s}}{n} \lambda + \frac{1}{\sqrt{s}} \|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}} \|\beta^*_{T^c}\|_{1,2} \right), \tag{72}$$

i.e., with probability at least $1 - C \exp(-C \frac{s \log(es_g b) + s_g \log(d/s_g)}{s})$,

$$\|h\|_2 \leq C \left(\sqrt{\frac{\sigma^2 (s_g \log(d/s_g) + s \log(es_g b))}{n}} + \frac{1}{\sqrt{s}} \|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}} \|\beta^*_{T^c}\|_{1,2} \right).$$

Moreover, if β^* is (s, s_g)-sparse, then $\|\beta^*_{T^c}\|_1 = \|\beta^*_{T^c}\|_{1,2} = 0$. Therefore, with probability at least $1 - C \exp(-C \frac{s \log(es_g b) + s_g \log(d/s_g)}{s})$,

$$\|h\|_2^2 \leq \frac{C \sigma^2 (s_g \log(d/s_g) + s \log(es_g b))}{n}.$$

$\square$

6.7 Proof of Theorem 5

First, we consider the case that $d \geq 3s_g$ and $b \geq 3s/s_g$. Let $\omega^{(1)}, \ldots, \omega^{(N)}$ be uniformly randomly vectors from

$$A = \{\omega \in \{0,1\}^{db} | \sum_j 1_{\{\omega_{(j)} \neq 0\}} = s_g, \|\omega_{(j)}\|_0 = \lfloor s/s_g \rfloor \text{ if } \omega_{(j)} \neq 0\}.$$

Denote $\Omega^{(i)} = \{j | \omega_j^{(i)} \neq 0\}$, $\Omega_{(k)}^{(i)} = \{j | j \in (k), \omega_j^{(i)} \neq 0\}$ and $\beta^{(i)} = \tau \omega^{(i)}$, for all $1 \leq i \leq N, 1 \leq k \leq d$, where τ is a parameter that will be specified later. Obviously, $\|\beta^{(i)}\|_0 \leq s$, therefore $\|\beta^{(i)} - \beta^{(j)}\|_2^2 \leq 2s\tau^2$.

Moreover, if $|\Omega^{(i)} \cap \Omega^{(j)}| \geq 8s_g \lfloor s/s_g \rfloor / 9$, then we must have

$$\left| \left\{ k | \omega_{(k)}^{(i)}, \omega_{(k)}^{(j)} \neq 0, \left| \Omega_{(k)}^{(i)} \cap \Omega_{(k)}^{(j)} \right| \geq 2 \lfloor s/s_g \rfloor / 3 \right\} \right| \geq 2s_g/3,$$

otherwise $|\Omega^{(i)} \cap \Omega^{(j)}| \leq \frac{2s_g}{3} \lfloor s/s_g \rfloor + \frac{s_g}{3} 2 \lfloor s/s_g \rfloor / 3 \leq 8s_g \lfloor s/s_g \rfloor / 9$, which is a contradiction.

Therefore,

$$\mathbb{P}\left(\|\beta^{(i)} - \beta^{(j)}\|_2^2 \leq 2s_g\lfloor s/s_g\rfloor\tau^2/9\right)$$

$$=\mathbb{P}\left(|\Omega^{(i)} \cap \Omega^{(j)}| \geq 8s_g\lfloor s/s_g\rfloor/9\right)$$

$$\leq\mathbb{P}\left(\left|\left\{k|\omega_{(k)}^{(i)}, \omega_{(k)}^{(j)} \neq 0, \left|\Omega_{(k)}^{(i)} \cap \Omega_{(k)}^{(j)}\right| \geq 2\lfloor s/s_g\rfloor/3\right\}\right| \geq 2s_g/3\right)$$

$$\leq\frac{\sum_{l=\lceil 2s_g/3\rceil}^{s_g} \binom{s_g}{l} \left[\sum_{t=\lceil 2\lfloor s/s_g\rfloor/3\rceil}^{\lfloor s/s_g\rfloor} \binom{\lfloor s/s_g\rfloor}{t}\binom{b-\lfloor s/s_g\rfloor}{\lfloor s/s_g\rfloor-t}\right]^l \binom{b}{\lfloor s/s_g\rfloor}^{s_g-l}\binom{d-l}{s_g-l}}{\binom{d}{s_g}\binom{b}{\lfloor s/s_g\rfloor}^{s_g}} \tag{73}$$

$$= \sum_{l=\lceil 2s_g/3\rceil}^{s_g} \binom{s_g}{l}\frac{\binom{d-l}{s_g-l}}{\binom{d}{s_g}} \cdot \left[\sum_{t=\lceil 2\lfloor s/s_g\rfloor/3\rceil}^{\lfloor s/s_g\rfloor} \binom{\lfloor s/s_g\rfloor}{t}\frac{\binom{b-\lfloor s/s_g\rfloor}{\lfloor s/s_g\rfloor-t}}{\binom{b}{\lfloor s/s_g\rfloor}}\right]^l.$$

Note that

$$\frac{\binom{d-l}{s_g-l}}{\binom{d}{s_g}} = \frac{\frac{(d-l)\cdots(d-s_g+1)}{(s_g-l)!}}{\frac{d(d-1)\cdots(d-s_g+1)}{s_g!}} = \frac{s_g(s_g-1)\cdots(s_g-l+1)}{d(d-1)\cdots(d-l+1)} \leq \left(\frac{s_g}{d}\right)^l,$$

The inequality holds since $\frac{s_g-i}{d-i} \leq \frac{s_g}{d}$ for all $1 \leq i \leq s_g$.

Similarly, for $1 \leq t \leq \lfloor s/s_g\rfloor$,

$$\frac{\binom{b-\lfloor s/s_g\rfloor}{\lfloor s/s_g\rfloor-t}}{\binom{b}{\lfloor s/s_g\rfloor}} \leq \frac{\binom{b-t}{\lfloor s/s_g\rfloor-t}}{\binom{b}{\lfloor s/s_g\rfloor}} \leq \left(\frac{\lfloor s/s_g\rfloor}{b}\right)^t.$$

Combine (73) and the previous two inequalities together, we have

$$\mathbb{P}\left(\|\beta^{(i)} - \beta^{(j)}\|_2^2 \leq 2s_g\lfloor s/s_g\rfloor\tau^2/9\right)$$

$$\leq \sum_{l=\lceil 2s_g/3\rceil}^{s_g} \binom{s_g}{l}\left(\frac{s_g}{d}\right)^l \cdot \left[\sum_{t=\lceil 2\lfloor s/s_g\rfloor/3\rceil}^{\lfloor s/s_g\rfloor} \binom{\lfloor s/s_g\rfloor}{t}\left(\frac{\lfloor s/s_g\rfloor}{b}\right)^t\right]^l$$

$$\leq \sum_{l=\lceil 2s_g/3\rceil}^{s_g} \binom{s_g}{l}\left(\frac{s_g}{d}\right)^l \cdot \left[\sum_{t=\lceil 2\lfloor s/s_g\rfloor/3\rceil}^{\lfloor s/s_g\rfloor} \binom{\lfloor s/s_g\rfloor}{t}\left(\frac{\lfloor s/s_g\rfloor}{b}\right)^{2\lfloor s/s_g\rfloor/3}\right]^l$$

$$\leq \sum_{l=\lceil 2s_g/3\rceil}^{s_g} \binom{s_g}{l}\left(\frac{s_g}{d}\right)^l \cdot \left[2^{\lfloor s/s_g\rfloor}\left(\frac{\lfloor s/s_g\rfloor}{b}\right)^{2\lfloor s/s_g\rfloor/3}\right]^l \tag{74}$$

$$\leq \sum_{l=\lceil 2s_g/3\rceil}^{s_g} \binom{s_g}{l}\left(\frac{s_g}{d}\right)^{2s_g/3} \cdot \left[\left(\frac{2\sqrt{2}\lfloor s/s_g\rfloor}{b}\right)^{2\lfloor s/s_g\rfloor/3}\right]^{2s_g/3}$$

$$\leq \left(\frac{2\sqrt{2}s_g}{d}\right)^{2s_g/3} \cdot \left(\frac{2\sqrt{2}\lfloor s/s_g\rfloor}{b}\right)^{2s/9}.$$

38

Set $N = \left\lfloor \left(\frac{d}{2\sqrt{2}s_g}\right)^{s_g/3} \left(\frac{b}{2\sqrt{2}\lfloor s/s_g \rfloor}\right)^{s/9} \right\rfloor$, then

$$\mathbb{P}\left(\forall 1 \le i \ne j \le N, \|\beta^{(i)} - \beta^{(j)}\|_2^2 > 2s_g \lfloor s/s_g \rfloor \tau^2/9\right)$$

$$\ge 1 - \frac{N(N-1)}{2}\left(\frac{2\sqrt{2}s_g}{d}\right)^{2s_g/3} \cdot \left(\frac{2\sqrt{2}\lfloor s/s_g \rfloor}{b}\right)^{2s/9}$$

$$> 0.$$

i.e., the probability that $\beta^{(1)}, \ldots, \beta^{(N)}; \Omega^{(1)}, \cdots, \Omega^{(N)}$ satisfy

$$\frac{s}{9}\tau^2 < 2s_g \lfloor s/s_g \rfloor \tau^2/9 < \min_{i \ne j} \|\beta^{(i)} - \beta^{(j)}\|_2^2 \le 2s\tau^2, \tag{75}$$

$$|\Omega^{(i)} \cap \Omega^{(j)}| < 8s_g \lfloor s/s_g \rfloor/9, \quad \forall 1 \le i < j \le N \tag{76}$$

is positive. For convenience, we fix $\beta^{(1)}, \ldots, \beta^{(N)}$ to be the vectors satisfying (75).

Denote $y^{(i)} = X\beta^{(i)} + \varepsilon$ for all $1 \le i \le N$. We consider the Kullback-Leibler divergence between different distribution pairs:

$$D_{KL}\left((y^{(i)}, X), (y^{(j)}, X)\right) = \mathbb{E}_{(y^{(j)}, X)}\left[\log\left(\frac{p(y^{(i)}, X)}{p(y^{(j)}, X)}\right)\right],$$

where $p(y^{(i)}, X)$ is the probability density of $(y^{(i)}, X)$. Conditioning on X, we have

$$\mathbb{E}_{(y^{(j)}, X)}\left[\log\left(\frac{p(y^{(i)}, X)}{p(y^{(j)}, X)}\right)|X\right] = \frac{\|X(\beta^{(i)} - \beta^{(j)})\|_2^2}{2\sigma^2}.$$

Thus for $1 \le i \ne j \le N$,

$$\begin{aligned}
D_{KL}\left((y^{(i)}, X), (y^{(j)}, X)\right) &= \mathbb{E}_X \frac{\|X(\beta^{(i)} - \beta^{(j)})\|_2^2}{2\sigma^2} = \frac{n(\beta^{(i)} - \beta^{(j)})^\top \Sigma(\beta^{(i)} - \beta^{(j)})}{2\sigma^2} \\
&\le \frac{3n\|\beta^{(i)} - \beta^{(j)}\|_2^2}{4\sigma^2} \le \frac{3ns\tau^2}{2\sigma^2}.
\end{aligned} \tag{77}$$

In the first inequality, we used $\sigma_{\max}(\Sigma) \le \frac{3}{2}$.

By generalized Fano's Lemma,

$$\inf_{\hat\beta} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat\beta - \beta\|_2 \ge \frac{\sqrt{s\tau^2/9}}{2}\left(1 - \frac{\frac{3ns\tau^2}{2\sigma^2} + \log 2}{\log N}\right).$$

Since $\log N \asymp s_g \log(\frac{d}{s_g}) + s \log\left(\frac{es_g b}{s}\right)$, by setting $\tau = c\sqrt{\frac{\sigma^2\left(s_g \log(\frac{d}{s_g}) + s \log\left(\frac{es_g b}{s}\right)\right)}{ns}}$, we have

$$\inf_{\hat\beta} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat\beta - \beta\|_2^2 \ge \left(\inf_{\hat\beta} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat\beta - \beta\|_2\right)^2 \ge c\frac{\sigma^2\left(s_g \log(d/s_g) + s \log(es_g b/s)\right)}{n}.$$

If $d < 3s_g$ or $b < 3s/s_g$, let $s'_g = [s_g/3] \vee 1 \geq s_g/5, s' = [s/15] \vee s'_g$, then $d \geq 3s'_g$ and $b \geq 3s'/s'_g$, we have

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \geq c\frac{\sigma^2 \left(s'_g \log(d/s'_g) + s' \log(es'_g b/s')\right)}{n} \geq c\frac{\sigma^2 \left(s_g \log(d/s_g) + s \log(es_g b/s)\right)}{n}.$$

$\square$

6.8   Proof of Theorem 6

The proof of Theorem 6 relies on the following key lemma, which shows that Σ^{-1} is in the feasible set of the optimization problem (23) with high probability by choosing appropriate α and γ.

Lemma 4 *By setting* $\alpha = C\sqrt{\frac{s\log(es_g b) + s_g \log(d/s_g)}{sn}}, \gamma = \sqrt{\frac{s}{s_g}}\alpha$ *in* (23), *we have*

$$\mathbb{P}\left(\max_{1 \leq i \leq p} \|H_\alpha(e_i - \frac{1}{n}X^\top X \Sigma^{-1} e_i)\|_{\infty,2} \leq \gamma\right) \geq 1 - 4\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right).$$

Note that $Y = X\beta^* + \varepsilon$, we have

$$\sqrt{n}(\hat{\beta}^u - \beta^*) = \sqrt{n}\left(\hat{\beta} - \beta^* + \frac{1}{n}\hat{M}X^\top\left(Y - X\hat{\beta}\right)\right) = \sqrt{n}\left(I - \frac{1}{n}\hat{M}X^\top X\right)(\hat{\beta} - \beta^*) + \frac{1}{\sqrt{n}}\hat{M}X^\top\varepsilon.$$

Since $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, we know that

$$\frac{1}{\sqrt{n}}\hat{M}X^\top\varepsilon|X \sim N\left(0, \hat{M}\hat{\Sigma}\hat{M}^\top\right).$$

Denote $h = \hat{\beta} - \beta^*$. Since β^* is (s, s_g)-sparse, by (69), (72) and Cauchy-Schwarz inequality, with probability at least $1 - C\exp(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s})$,

$$\|h\|_1 \leq \|h_T\|_1 + \|h_{T^c}\|_1 \leq \sqrt{s}\|h_T\|_2 + \|h_{T^c}\|_1 \leq \sqrt{s}\|h\|_2 + \|h_{T^c}\|_1 \leq C\frac{s}{n}\lambda.$$

$$\|h\|_{1,2} \leq \|h_{(G)}\|_{1,2} + \|h_{(G^c)}\|_{1,2} \leq \sqrt{s_g}\|h_{(G)}\|_2 + \|h_{(G^c)}\|_{1,2} \leq \sqrt{s_g}\|h\|_2 + \|h_{(G^c)}\|_{1,2} \leq C\frac{\sqrt{s \cdot s_g}}{n}\lambda.$$

In addition, Lemma 4 shows that Σ^{-1} is in the feasible set of (23) with probability at least $1 - C\exp(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s})$. By the definition of $\hat{M}$,

$$\max_i \|H_\alpha(e_i - \hat{\Sigma}\hat{M}^\top e_i)\|_{\infty,2} = \max_i \|H_\alpha(e_i - \hat{\Sigma}\hat{m}_i)\|_{\infty,2} \leq \gamma. \tag{78}$$

Combining these facts, by Lemma 8 Part 2, we must have

$$\left\| (I - \frac{1}{n}\hat{M}XX^\top)(\hat{\beta} - \beta^*) \right\|_\infty = \max_i \left| \langle e_i - \hat{\Sigma}\hat{M}^\top e_i, h \rangle \right|$$

$$\leq \alpha \|h\|_1 + \gamma \|h\|_{1,2}$$

$$\leq C \frac{s}{n}\alpha\lambda + C\frac{\sqrt{s \cdot s_g}}{n}\gamma\lambda$$

$$= \frac{C(s\log(es_g b) + s_g \log(d/s_g))}{n}\sigma$$

with probability at least $1 - C\exp(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s})$. This has finished the proof of (24).

Next, we consider $\hat{m}_i^\top \hat{\Sigma}\hat{m}_i$. By (78) and Lemma 8 Part 2, we have

$$1 - \langle e_i, \hat{\Sigma}\hat{m}_i \rangle = \langle e_i, e_i - \hat{\Sigma}\hat{m}_i \rangle \leq \alpha \|e_i\|_1 + \gamma\|e_i\|_{1,2} = \alpha + \gamma.$$

Therefore, for any $c \geq 0$,

$$\hat{m}_i^\top \hat{\Sigma}\hat{m}_i \geq \hat{m}_i^\top \hat{\Sigma}\hat{m}_i + c(1 - \alpha - \gamma) - c\langle e_i, \hat{\Sigma}\hat{m}_i \rangle \geq \min_m \left\{ m^\top\hat{\Sigma}m + c(1 - \alpha - \gamma) - c\langle e_i, \hat{\Sigma}m \rangle \right\}.$$

Since $m = ce_i/2$ achieves the minimum of the right hand side, we have

$$\hat{m}_i^\top \hat{\Sigma}\hat{m}_i \geq c(1 - \alpha - \gamma) - \frac{c^2}{4}\hat{\Sigma}_{i,i}.$$

If $\hat{\Sigma}_{ii} > 0$ for all $1 \leq i \leq p$, by setting $c = 2(1 - \alpha - \gamma)/\hat{\Sigma}_{i,i}$, we have

$$\hat{m}_i^\top \hat{\Sigma}\hat{m}_i \geq \frac{(1 - \alpha - \gamma)^2}{\hat{\Sigma}_{i,i}}, \quad \forall 1 \leq i \leq p. \tag{79}$$

Moreover, by Lemma 6 Part 2 with $u = v = e_i$, we have

$$\mathbb{P}\left(\left| \hat{\Sigma}_{i,i} - \Sigma_{i,i} \right| \geq \frac{1}{6} \right) \leq 2\exp\left(-cn \right).$$

By the union bound,

$$\mathbb{P}\left(\exists 1 \leq i \leq p, \left| \hat{\Sigma}_{i,i} - \Sigma_{i,i} \right| \geq \frac{1}{6} \right)$$

$$\leq \sum_{i=1}^p \mathbb{P}\left(\left| \hat{\Sigma}_{i,i} - \Sigma_{i,i} \right| \geq \frac{1}{6} \right)$$

$$\leq db \cdot 2\exp\left(-cn \right)$$

$$\leq 2\exp\left(-cn \right).$$

Therefore, with probability at least $1 - 2\exp(-cn)$,

$$\frac{1}{2} \leq \hat{\Sigma}_{i,i} \leq \frac{5}{3}, \quad \forall 1 \leq i \leq p.$$

(79) and the previous inequality together imply that with probability at least $1 - 2\exp(-cn)$,

$$\hat{m}_i^\top \hat{\Sigma} \hat{m}_i \geq \frac{1}{2}, \quad \forall 1 \leq i \leq p.$$

(24) and the previous inequality together imply (25). $\quad\square$

References

[1] M. Silver, P. Chen, R. Li, C.-Y. Cheng, T.-Y. Wong, E.-S. Tai, Y.-Y. Teo, and G. Montana, "Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two asian cohorts," *PLoS genetics*, vol. 9, no. 11, p. e1003939, 2013.

[2] M. Vidyasagar, "Machine learning methods in the computational biology of cancer," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 470, no. 2167, p. 20140081, 2014.

[3] A. Allahyar and J. De Ridder, "Feral: network-based classifier with application to breast cancer outcome prediction," *Bioinformatics*, vol. 31, no. 12, pp. i311–i319, 2015.

[4] N. Rao, R. Nowak, C. Cox, and T. Rogers, "Classification with the sparse group lasso," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 448–463, 2015.

[5] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, "Sparse group lasso: Consistency and climate applications," in *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 47–58, SIAM, 2012.

[6] W. Wang, Y. Liang, and E. Xing, "Block regularized lasso for multivariate multi-response linear regression," in *Artificial Intelligence and Statistics*, pp. 608–617, 2013.

[7] K. Lounici, M. Pontil, A. Tsybakov, and S. Van De Geer, "Taking advantage of sparsity in multi-task learning," in *COLT 2009-The 22nd Conference on Learning Theory*, 2009.

[8] A. C. Lozano and G. Swirszcz, "Multi-level lasso for sparse multi-task regression," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 595–602, Omnipress, 2012.

[9] H. H. Zhou, Y. Zhang, V. K. Ithapu, S. C. Johnson, V. Singh, *et al.*, "When can multi-site datasets be pooled for regression? hypothesis tests, ℓ_2-consistency and neuroscience applications," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4170–4179, JMLR. org, 2017.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[11] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[12] Y. Li, B. Nan, and J. Zhu, "Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure," *Biometrics*, vol. 71, no. 2, pp. 354–363, 2015.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[14] P. J. Bickel, Y. Ritov, A. B. Tsybakov, *et al.*, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[15] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[16] K. Lounici, M. Pontil, S. Van De Geer, A. B. Tsybakov, *et al.*, "Oracle inequalities and optimal inference under group sparsity," *The Annals of Statistics*, vol. 39, no. 4, pp. 2164–2204, 2011.

[17] F. Bunea, J. Lederer, and Y. She, "The group square-root lasso: Theoretical properties and fast algorithms," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1313–1325, 2013.

[18] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009.

[19] Z. Ma, "Sparse principal component analysis and iterative thresholding," *The Annals of Statistics*, vol. 41, no. 2, pp. 772–801, 2013.

[20] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational limits," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 7311–7338, 2018.

[21] M. Wang and L. Li, "Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality," *arXiv preprint arXiv:1811.05076*, 2018.

[22] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[23] B. Hao, A. Zhang, and G. Cheng, "Sparse and low-rank tensor estimation via cubic sketchings," *arXiv preprint arXiv:1801.09326*, 2018.

[24] A. Zhang and R. Han, "Optimal sparse singular value decomposition for high-dimensional high-order data," *Journal of the American Statistical Association*, 2018.

[25] K. Jaganathan, S. Oymak, and B. Hassibi, "Sparse phase retrieval: Convex algorithms and limitations," in *2013 IEEE International Symposium on Information Theory*, pp. 1022–1026, IEEE, 2013.

[26] Y. Shechtman, A. Beck, and Y. C. Eldar, "Gespar: Efficient phase retrieval of sparse signals," *IEEE transactions on signal processing*, vol. 62, no. 4, pp. 928–938, 2014.

[27] S. Oymak, A. Jalali, M. Fazel, and B. Hassibi, "Noisy estimation of simultaneously structured models: Limitations of convex relaxation," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pp. 6019–6024, IEEE, 2013.

[28] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, 2014.

[29] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[30] R. Mitra, C.-H. Zhang, *et al.*, "The benefit of group sparsity in group inference with de-biased scaled group lasso," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 1829–1873, 2016.

[31] T. T. Cai and Z. Guo, "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity," *The Annals of statistics*, vol. 45, no. 2, pp. 615–646, 2017.

[32] S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, *et al.*, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[33] M. Stojnic, W. Xu, and B. Hassibi, "Compressed sensing-probabilistic analysis of a null-space characterization," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3377–3380, IEEE, 2008.

[34] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[35] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[36] T. T. Cai and A. Zhang, "Compressed sensing and affine rank minimization under restricted isometry," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3279–3290, 2013.

[37] T. T. Cai and A. Zhang, "Sharp rip bound for sparse signal and low-rank matrix recovery," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 74–93, 2013.

[38] T. T. Cai and A. Zhang, "Sparse representation of a polytope and recovery of sparse signals and low-rank matrices," *IEEE transactions on information theory*, vol. 60, no. 1, pp. 122–132, 2014.

[39] B. Poignard, "Asymptotic theory of the adaptive sparse group lasso," *Annals of the Institute of Statistical Mathematics*, pp. 1–32, 2018.

[40] N. Rao, C. Cox, R. Nowak, and T. T. Rogers, "Sparse overlapping sets lasso for multitask learning and its application to fmri analysis," in *Advances in neural information processing systems*, pp. 2202–2210, 2013.

[41] M. E. Ahsen and M. Vidyasagar, "Error bounds for compressed sensing algorithms with group sparsity: A unified approach," *Applied and Computational Harmonic Analysis*, vol. 43, no. 2, pp. 212–232, 2017.

[42] E. J. Candes and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *IEEE transactions on information theory*, vol. 57, no. 11, pp. 7235–7254, 2011.

[43] A. Javanmard, A. Montanari, *et al.*, "Debiasing the lasso: Optimal sample size for gaussian designs," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2593–2622, 2018.

[44] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, vol. 1. Birkhäuser Basel, 2013.

[45] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[46] D. Bertsekas and A. Nedic, "Convex analysis and optimization (conservative)," 2003.

[47] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[48] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[49] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[50] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Foundations of Computational Mathematics*, vol. 16, no. 4, pp. 1031–1068, 2016.

[51] S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, *et al.*, "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, vol. 42, no. 3, pp. 1166–1202, 2014.

[52] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.

[53] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[54] A. Rinaldo *et al.*, "Properties and refinements of the fused lasso," *The Annals of Statistics*, vol. 37, no. 5B, pp. 2922–2952, 2009.

[55] A. Jalali and M. Fazel, "A convex method for learning d-valued models," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 1123–1126, IEEE, 2013.

[56] A. Jalali, A. Javanmard, and M. Fazel, "New computational and statistical aspects of regularized regression with application to rare feature selection and aggregation," *arXiv preprint arXiv:1904.05338*, 2019.

[57] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. Eldar, "Collaborative hierarchical sparse modeling," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2010.

[58] D. Hsu, S. Kakade, T. Zhang, *et al.*, "A tail inequality for quadratic forms of subgaussian random vectors," *Electronic Communications in Probability*, vol. 17, 2012.

[59] M. Rudelson, R. Vershynin, *et al.*, "Hanson-wright inequality and sub-gaussian concentration," *Electronic Communications in Probability*, vol. 18, 2013.

[60] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[61] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

A   Technical Lemmas

We collect all additional technical lemmas and their proofs in this section.

Lemma 5 (Bernstein-type Inequality for Soft-thresholded Sub-Gaussian Vectors)
Suppose the rows of $X \in \mathbb{R}^{n \times p}$ are independent sub-Gaussian vectors satisfying Assumption 1. $w \in \mathbb{R}^n$ is a fixed vector, Ω is a subset of $\{1, \ldots, p\}$ with $|\Omega| = r$. Then

$$\mathbb{P}\left(\left\| \sum_{k=1}^{n} w_k X_{k,\Omega} \right\|_2 \geq \sqrt{\frac{3}{2}} \kappa \|w\|_2 \cdot \left(\sqrt{r} + \sqrt{2t} \right) \right) \leq \exp(-t). \tag{80}$$

For any fixed vector $w \in \mathbb{R}^n$ and fixed index subset $\Omega \subseteq \{1, \ldots, p\}$ with $|\Omega| = r$,

$$\mathbb{P}\left(\left\| H_{(\delta\|w\|_2)}(w^\top X_\Omega) \right\|_2 \geq t\|w\|_2 \right) \leq \binom{r}{\lfloor (t/\delta)^2 \rfloor \wedge r} \cdot \exp\left(-(t/(\kappa\sqrt{3/2}) - (t/\delta) \wedge \sqrt{r})_+^2 /2 \right)$$
$$+ \binom{r}{\lceil (t/\delta)^2 \rceil} \cdot \exp\left(-(t/(\kappa\sqrt{3/2}) - \sqrt{\lceil (t/\delta)^2 \rceil})_+^2 /2 \right). \tag{81}$$

In particular, for any $b \geq r$, if $\bar{\lambda} = C\|w\|_2 \sqrt{\frac{s \log(es_g b) + s_g \log(d/s_g)}{s}}$, $\bar{\lambda}_g = \sqrt{s/s_g}\bar{\lambda}$, we have

$$\mathbb{P}\left(\left\| H_{\bar{\lambda}}(w^\top X_\Omega) \right\|_2 \geq \bar{\lambda}_g \right) \leq \exp\left(-C \frac{s \log(es_g b) + s_g \log(d/s_g)}{s_g} \right). \tag{82}$$

Proof of Lemma 5. We only need to focus on the case where $\|w\|_2 = 1$. Let $W_\Omega = X_\Omega \Sigma_{\Omega,\Omega}^{-1/2}$, immediately we know that $W_{1,\Omega}, \ldots, W_{n,\Omega}$ are isotropic sub-Gaussian distributed. Then for any fixed w, $w^\top W_\Omega$ is also an isotropic sub-Gaussian vector such that for any $\alpha \in \mathbb{R}^r$,

$$\mathbb{E} \exp\left(w^\top W_\Omega \alpha \right) = \mathbb{E} \exp\left(w^\top X_\Omega \Sigma_{\Omega,\Omega}^{-1/2} \alpha \right) = \mathbb{E} \exp\left(w^\top X \Sigma^{-1/2} (\Sigma^{1/2})_{\cdot,\Omega} \Sigma_{\Omega,\Omega}^{-1/2} \alpha \right)$$
$$\leq \exp\left(\kappa^2 \|(\Sigma^{1/2})_{\cdot,\Omega} \Sigma_{\Omega,\Omega}^{-1/2} \alpha\|_2^2 /2 \right) = \exp\left(\kappa^2 \|\alpha\|_2^2 /2 \right).$$

The last equation holds since $(\Sigma^{1/2})_{\Omega,\cdot}(\Sigma^{1/2})_{\cdot,\Omega} = (\Sigma^{1/2}\Sigma^{1/2})_{\Omega,\Omega} = \Sigma_{\Omega,\Omega}$.

By the tail inequality of sub-Gaussian quadratic form ([58, Theorem 2.1]),

$$\mathbb{P}\left(\left\| w^\top W_\Omega \right\|_2^2 \geq \kappa^2 \left(r + 2\sqrt{rt} + 2t \right) \right) \leq \exp(-t).$$

By taking square-root of the previous inequality, we have

$$\mathbb{P}\left(\left\| w^\top W_\Omega \right\|_2 \geq \kappa\|w\|_2 \cdot \left(\sqrt{r} + \sqrt{2t} \right) \right) \leq \exp(-t).$$

Also note that

$$\left\|w^\top X_\Omega\right\|_2 = \left\|w^\top W_\Omega \Sigma_{\Omega,\Omega}^{1/2}\right\|_2 \leq \left\|\Sigma_{\Omega,\Omega}^{1/2}\right\| \left\|w^\top W_\Omega\right\|_2 \leq \|\Sigma\|^{1/2} \left\|w^\top W_\Omega\right\|_2 \leq \sqrt{\frac{3}{2}} \left\|w^\top W_\Omega\right\|_2,$$

we obtain (80).

For the second part of proof, note that

$$\mathbb{P}\left(\left\|H_\delta(w^\top X_\Omega)\right\| \geq t\right)$$

$$\leq \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \text{ such that all entries of } |w^\top X_\Lambda| \geq \delta \text{ and } \|w^\top X_\Lambda\|_2 \geq t\right)$$

$$\leq \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta \leq t, \|w^\top X_\Lambda\|_2 \geq t\right)$$

$$+ \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta > t, \text{ all entries of } |w^\top X_\Lambda| \geq \delta\right)$$

$$\leq \sum_{\substack{\Lambda \subseteq \Omega \\ |\Lambda| = \lfloor (t/\delta)^2 \rfloor \wedge r}} \mathbb{P}\left(\|w^\top X_\Lambda\|_2 \geq t\right) + \sum_{\substack{\Lambda \subseteq \Omega \\ |\Lambda| = \lceil (t/\delta)^2 \rceil}} \mathbb{P}\left(\|w^\top X_\Lambda\|_2 \geq t\right).$$

By the first part of this lemma,

$$\mathbb{P}\left(\|w^\top X_\Lambda\|_2 \geq \sqrt{\frac{3}{2}}\kappa\|w\|_2 t\right) \leq \exp\left(-\left(t - \sqrt{|\Lambda|}\right)_+^2 /2\right).$$

Plug in this to the previous inequality, one has

$$\mathbb{P}\left(\left\|H_\delta(w^\top X_\Omega)\right\| \geq t\right) \leq \binom{r}{\lfloor (t/\delta)^2 \rfloor \wedge r} \cdot \exp\left(-(t/(\kappa\sqrt{3/2}) - (t/\delta) \wedge \sqrt{r})_+^2 /2\right)$$

$$+ \binom{r}{\lceil (t/\delta)^2 \rceil} \cdot \exp\left(-(t/(\kappa\sqrt{3/2}) - \sqrt{\lceil (t/\delta)^2 \rceil})_+^2\right).$$

Specifically, if $\delta = C\sqrt{\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}}$, $t = \sqrt{s/s_g}\delta$,

$$t/(\kappa\sqrt{3/2}) - \sqrt{\lceil (t/\delta)^2 \rceil} \geq C\sqrt{\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}} - \sqrt{2\frac{s}{s_g}} \geq C\sqrt{\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}}.$$

Therefore, (81) shows that

$$\mathbb{P}\left(\left\|H_{\bar\lambda}(w^\top X_\Omega)\right\|_2 \geq \bar\lambda_g\right) \leq r^{\lfloor (t/\delta)^2 \rfloor} \exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right)$$

$$+ r^{\lceil (t/\delta)^2 \rceil} \exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right)$$

$$\leq 2r^{2s/s_g} \exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right)$$

$$\leq \exp\left(\log 2 + \frac{2s\log(eb)}{s_g} - C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right)$$

$$\leq \exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right).$$

□

Lemma 6 (sub-Gaussian quadratic form concentrations) *Suppose $Z \in \mathbb{R}^p$ is a sub-Gaussian vector satisfying Assumption 1.*

1. *For any fixed $u, v \in \mathbb{R}^p, u, v \neq 0$, $u^\top Z Z^\top v$ is sub-exponential such that for every $t > 0$,*

$$\mathbb{P}\left(\left|u^\top Z Z^\top v - \mathbb{E}u^\top Z Z^\top v\right| \geq t\|u\|_2\|v\|_2\right) \leq C\exp(-ct/\kappa^2). \tag{83}$$

2. *In addition, suppose $X = [X_1^\top, \ldots, X_n^\top]^\top \in \mathbb{R}^{n \times p}$ is a random matrix with independent random sub-Gaussian rows satisfying Assumption 1,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n u^\top X_k X_k^\top v - u^\top \Sigma v\right| \geq t\|u\|_2\|v\|_2\right) \leq 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \tag{84}$$

3. *More generally, for any fixed matrix $U \in \mathbb{R}^{p \times r}$, the following concentration inequality in spectral norm holds,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{k=1}^n U^\top X_k X_k^\top v - U^\top \Sigma v\right\|_2 \geq t\|U\|\|v\|_2\right) \leq 2\exp\left(Cr - cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \tag{85}$$

Proof of Lemma 6. Since we can rescale u and v without essentially changing the problem, without loss of generality we assume $\|u\|_2 = \|v\|_2 = 1$. Let $A = uv^\top$, then $u^\top Z Z^\top v = Z^\top uv^\top Z = Z^\top AZ$. By Assumption 1, $\mathbb{E}Z = 0$ and $\|\langle Z, e_i\rangle\|_{\psi_2} \leq C\kappa$. By Hanson-Wright inequality ([59, Theorem 1.1]),

$$\begin{aligned}
\mathbb{P}\left(\left|u^\top Z Z^\top v - \mathbb{E}u^\top Z Z^\top v\right| \geq t\right) &= \mathbb{P}\left(|Z^\top AZ - \mathbb{E}Z^\top AZ| \geq t\right) \\
&\leq 2\exp\left[-c\min\left(\frac{t^2}{\kappa^4\|A\|_{HS}^2}, \frac{t}{\kappa^2\|A\|}\right)\right] \\
&\leq 2\exp\left[-c\min\left(\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right)\right],
\end{aligned}$$

where

$$\|A\|_{HS} = \left(\sum_{i,j}|a_{i,j}|^2\right)^{1/2} = \left(\sum_{i,j}|u_iv_j|^2\right)^{1/2} = \|u\|_2\|v\|_2 = 1,$$

$$\|A\| = \max_{\|x\|_2\leq 1}\|Ax\|_2 = \max_{\|x\|_2\leq 1}\|uv^\top x\|_2 = \|u\|_2\max_{\|x\|_2\leq 1}|v^\top x| = \|u\|_2\|v\|_2 = 1.$$

Therefore, for every $t \geq \kappa^2$,

$$\mathbb{P}\left(\left|u^\top ZZ^\top v - \mathbb{E}u^\top ZZ^\top v\right| \geq t\right) \leq 2\exp\left(-ct/\kappa^2\right).$$

Thus, there exists a constant $c < \log 2$, for every $t \geq 0$,

$$\mathbb{P}\left(\left|u^\top ZZ^\top v - \mathbb{E}u^\top ZZ^\top v\right| \geq t\right) \leq 2\exp\left(-ct/\kappa^2\right).$$

Notice that $\mathbb{E}u^\top X_k X_k^\top v = u^\top \Sigma v$ for all $1 \leq k \leq n$, by Bernstein-type concentration inequality (c.f., [60, Proposition 5.16]),

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n u^\top X_k X_k^\top v - u^\top \Sigma v\right| \geq t\right) \leq 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

This has finished the proof of (84).

Finally, we consider (85), which can be done by an ε-net argument and the result in (84). For any $w \in \mathbb{R}^r, \|w\|_2 = 1$, set $u = Uw$ in (84), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geq \frac{t}{2}\|Uw\|_2\|v\|_2\right) \leq 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

By [60, Lemma 5.3], we can find a $\frac{1}{2}$-net $\mathcal{N}_{\frac{1}{2}}$ of $S^{r-1} = \{x | x \in \mathbb{R}^r, \|x\|_2 = 1\}$ with $|\mathcal{N}_{\frac{1}{2}}| \leq 5^r$. By the union bound,

$$\begin{aligned} &\mathbb{P}\left(\forall w \in \mathcal{N}_{\frac{1}{2}}, \left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geq \frac{t}{2}\|Uw\|_2\|v\|_2\right) \\ &\leq 5^r \cdot 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \end{aligned} \tag{86}$$

For any $g \in \mathbb{R}^r, g \neq 0$, set $x = \frac{g}{\|g\|_2} \in \arg\max_{w \in \mathbb{R}^r, \|w\|_2 = 1} |w^\top g|$, we can find $y \in \mathcal{N}_{\frac{1}{2}}$ such that $\|x - y\|_2 \leq \frac{1}{2}$. By triangle inequality,

$$\|g\|_2 - |y^\top g| = |x^\top g| - |y^\top g| \leq |x^\top g - y^\top g| \leq \|x - y\|_2\|g\|_2 \leq \frac{1}{2}\|g\|_2.$$

Therefore,

$$\sup_{w \in \mathbb{R}^r, \|w\|_2 = 1}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \leq 2\sup_{w \in \mathcal{N}_{\frac{1}{2}}}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right|.$$

51

The (86) and the previous inequality together, also notice that $\|U\| = \sup_{w \in \mathbb{R}^r, \|w\|_2 = 1} \|Uw\|_2$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{w \in \mathbb{R}^r, \|w\|_2 = 1} \left|\frac{1}{n}\sum_{k=1}^{n} w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geq t\|U\|\|v\|_2\right) \\
&\leq \mathbb{P}\left(\sup_{w \in \mathcal{N}_{\frac{1}{2}}} \left|\frac{1}{n}\sum_{k=1}^{n} w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geq \frac{t}{2}\|U\|\|v\|_2\right) \\
&\leq \mathbb{P}\left(\forall w \in \mathcal{N}_{\frac{1}{2}}, \left|\frac{1}{n}\sum_{k=1}^{n} w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geq \frac{t}{2}\|Uw\|_2\|v\|_2\right) \\
&\leq 5^r \cdot 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).
\end{aligned}
\tag{87}
$$

Finally, note that

$$
\left\|\frac{1}{n}\sum_{k=1}^{n} U^\top X_k X_k^\top v - U^\top \Sigma v\right\|_2 = \sup_{w \in \mathbb{R}^r, \|w\|_2 = 1} \left|\frac{1}{n}\sum_{k=1}^{n} w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right|,
$$

we have proved (85). $\square$

We collect the random matrix properties of X in the following lemma. These properties will be extensively used in the main content of the paper.

Lemma 7 *Suppose $X = [X_1^\top, \ldots, X_n^\top]^\top \in \mathbb{R}^{n \times p}$ is a random matrix with independent random sub-Gaussian rows satisfying Assumption 1.*

1. *Suppose $T \subseteq \{1, \ldots, p\}$ is with cardinality s. Then,*

$$
\mathbb{P}\left(\left\|\frac{1}{n}X_T^\top X_T \Sigma_{T,T}^{-1} - I_s\right\| \geq t\right) \leq 2\exp\left(Cs - cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right);
\tag{88}
$$

2. *For any fixed vector $\alpha \in \mathbb{R}^s$, $\delta > 0$, and fixed index subset $\Omega \subseteq T^c$ satisfying $|\Omega| = r$, $t \geq \delta \geq C(\max_{i \in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2)\|\alpha\|_2$,*

$$
\begin{aligned}
&\mathbb{P}\left(\left\|H_\delta(\alpha^\top X_T^\top X_\Omega/n)\right\|_2 \geq t\right) \\
&\leq \binom{r}{\lfloor(t/\delta)^2\rfloor \wedge r} \exp\left(C\lfloor(t/\delta)^2\rfloor \wedge r - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right) \\
&\quad + \binom{r}{\lceil(t/\delta)^2\rceil}_+ \exp\left(C\lceil(t/\delta)^2\rceil - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right).;
\end{aligned}
\tag{89}
$$

Here, $H_\lambda(\cdot)$ is the soft-thresholding estimator at level λ.

Proof of Lemma 7.

1. The first statement is via ε-net. Denote $W_T = X_T \Sigma_{T,T}^{-1/2}$, then the rows of W_T are independent isotropic sub-Gaussian distributed. For any fixed vector $x \in S^{s-1} = \{x : x \in \mathbb{R}^s, \|x\|_2 = 1\}$, by [60, Lemma 5.5], $Z_i = \langle (W_T)_{i\cdot}^\top, x \rangle$ are independent sub-Gaussian random variables with $\mathbb{E} Z_i^2 = 1$ and $\|Z_i\|_{\psi_2} \le C\kappa$. Therefore, by Remark 5.18 and Lemma 5.14 in [60], $\|Z_i^2 - 1\|_{\psi_1} \le 2\|Z_i^2\|_{\psi_1} \le 4\|Z_i\|_{\psi_2}^2 \le C\kappa^2$. Bernstein-type inequality shows that

$$\mathbb{P}\left(\left|\frac{1}{n}\|W_T x\|_2^2 - 1\right| \ge \frac{t}{2}\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i^2 - 1)\right| \ge \frac{t}{2}\right) \le 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

By [60, Lemma 5.2], we can find a $\frac{1}{4}$-net $\mathcal{N}_{\frac{1}{4}}$ of $S^{s-1} = \{x : x \in \mathbb{R}^s, \|x\|_2 = 1\}$ with $|\mathcal{N}_{\frac{1}{4}}| \le 9^s$. The union bound tells us

$$\mathbb{P}\left(\max_{x \in \mathcal{N}_{\frac{1}{4}}} \left|\frac{1}{n}\|W_T x\|_2^2 - 1\right| \ge \frac{t}{2}\right) \le 9^s \cdot 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \tag{90}$$

By [60, Lemma 5.4],

$$\left\|\frac{1}{n}W_T^\top W_T - I_s\right\| \le 2\max_{x \in \mathcal{N}_{\frac{1}{4}}} \left|\langle (\frac{1}{n}W_T^\top W_T - I_s)x, x\rangle\right| = 2\max_{x \in \mathcal{N}_{\frac{1}{4}}} \left|\frac{1}{n}\|W_T x\|_2^2 - 1\right|. \tag{91}$$

Since $\frac{2}{3} \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le \frac{3}{2}$, we have $\|\Sigma_{T,T}^{1/2}\|, \|\Sigma_{T,T}^{-1/2}\| \le \sqrt{\frac{3}{2}}$. Therefore,

$$\begin{aligned}
\left\|\frac{1}{n}X_T^\top X_T \Sigma_{T,T}^{-1} - I_s\right\| &= \left\|\Sigma_{T,T}^{1/2}(\frac{1}{n}W_T^\top W_T - I_s)\Sigma_{T,T}^{-1/2}\right\| \\
&\le \|\Sigma_{T,T}^{1/2}\|\|\frac{1}{n}W_T^\top W_T - I_s\|\|\Sigma_{T,T}^{-1/2}\| \\
&\le \frac{3}{2}\|\frac{1}{n}W_T^\top W_T - I_s\|.
\end{aligned} \tag{92}$$

Combine (90), (91) and (92) together, we have arrived at the conclusion.

2. Now we consider the proof for (89). Note that $\|H_\delta(\alpha^\top X_T^\top X_\Omega)\|_2 \le t$ implies there exists $\Lambda \subset \Omega$ such that all entry of $|\alpha^\top X_T^\top X_\Lambda|$ are greater than δ, and $\left\||\alpha^\top X_T^\top X_\Lambda| - \delta\right\|_2 \ge t$.

Thus,

$$\mathbb{P}\left(\left\|H_\delta(\alpha^\top X_T^\top X_\Omega/n)\right\|_2 \geq t\right)$$

$$\leq \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \text{ such that all entries of } \left|\alpha^\top X_T^\top X_\Lambda/n\right| \geq \delta, \text{and } \|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geq t\right)$$

$$\leq \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta \leq t, \|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geq t\right)$$

$$+ \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta > t, \text{ all entries of } \left|\alpha^\top X_T^\top X_\Lambda/n\right| \geq \delta\right)$$

$$\leq \sum_{\substack{\Lambda \subseteq \Omega \\ |\Lambda| = \lfloor (t/\delta)^2 \rfloor \wedge r}} \mathbb{P}\left(\|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geq t\right) + \sum_{\substack{\Lambda \subseteq \Omega \\ |\Lambda| = \lceil (t/\delta)^2 \rceil}} \mathbb{P}\left(\text{all entries of } \left|\alpha^\top X_T^\top X_\Lambda/n\right| \geq \delta\right)$$

$$\leq \sum_{\substack{\Lambda \subseteq \Omega \\ |\Lambda| = \lfloor (t/\delta)^2 \rfloor \wedge r}} \mathbb{P}\left(\|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geq t\right) + \sum_{\substack{\Lambda \subseteq \Omega \\ |\Lambda| = \lceil (t/\delta)^2 \rceil}} \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t\right).$$

$$(93)$$

Since $t \geq \delta \geq C \max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2 \|\alpha\|_2$,

$$3\sqrt{\lceil (t/\delta)^2 \rceil}(\max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2)\|\alpha\|_2 \leq 3\sqrt{2}(t/\delta)(\max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2)\|\alpha\|_2 \leq t.$$

By Part 3 of Lemma 6, for any $\Lambda \subseteq \Omega$, $t \geq 3\sqrt{|\Lambda|} \max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2 \|\alpha\|_2$, we have

$$\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t\right)$$

$$\leq \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t - \left\|\mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2\right)$$

$$\leq \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t - \|\Sigma_{\Lambda,T}\alpha\|_2\right)$$

$$= \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t - \left(\sum_{i\in\Lambda}(\Sigma_{i,T}\alpha)^2\right)^{1/2}\right)$$

$$\leq \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t - \sqrt{|\Lambda|}\max_{i\in T^c}|\Sigma_{i,T}\alpha|\right)$$

$$\leq \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t - \sqrt{|\Lambda|}\max_{i\in T^c}\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\|\Sigma_{T,T}\|\|\alpha\|_2\right)$$

$$\leq \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geq t/2\right)$$

$$\leq 2\exp\left(C|\Lambda| - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right).$$

54

Combine (93) and the previous inequality, one obtains

$$\mathbb{P}\left(\left\|H_\delta(\alpha^\top X_T^\top X_\Omega/n)\right\|_2 \geq t\right)$$

$$\leq \binom{r}{\lfloor(t/\delta)^2\rfloor \wedge r} \exp\left(C\lfloor(t/\delta)^2\rfloor \wedge r - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right)$$

$$+ \binom{r}{\lceil(t/\delta)^2\rceil}_+ \exp\left(C\lceil(t/\delta)^2\rceil - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right).$$

□

Lemma 8 (Properties of Soft-thresholding) *1. Suppose $a, b > 0$, $x, y \in \mathbb{R}$, $H.(\cdot)$ is the soft-thresholding operator satisfying $H_a(x) = \text{sgn}(x)\cdot(|x|-a)_+$. Then the following triangular inequality holds,*

$$|H_{a+b}(x+y)| \leq |H_a(x)| + |H_b(y)|. \tag{94}$$

2. Suppose $a, b > 0$, $x, y \in \mathbb{R}^p$, if $\|H_a(x)\|_{\infty,2} \leq b$, then

$$|\langle x, y\rangle| \leq a\|y\|_1 + b\|y\|_{1,2}. \tag{95}$$

Proof of Lemma 8.

1.

$$|H_{a+b}(x+y)| = (|x+y| - a - b)_+ \leq (|x| - a + |y| - b)_+ \leq (|x| - a)_+ + (|y| - b)_+$$
$$= |H_a(x)| + |H_b(y)|.$$

2.

$$|\langle x, y\rangle| \leq |\langle H_a(x), y\rangle| + |\langle x - H_a(x), y\rangle| = |\sum_{j=1}^d \langle [H_a(x)]_{(j)}, y_{(j)}\rangle| + |\langle x - H_a(x), y\rangle|$$

$$\leq \sum_{j=1}^d \|[H_a(x)]_{(j)}\|_2\|y_{(j)}\|_2 + \|x - H_a(x)\|_\infty\|y\|_1 \leq \|H_a(x)\|_{\infty,2}\|y\|_{1,2} + \|x - a\|y\|_1$$

$$\leq b\|y\|_{1,2} + a\|y\|_1.$$

□

55

Lemma 9 *Suppose $X = [X_1^\top, \ldots, X_n^\top]^\top \in \mathbb{R}^{n \times p}$ is a random matrix with independent random sub-Gaussian rows satisfying Assumption 1, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Suppose $T \subseteq \{1, \ldots, p\}$ is with cardinality s, $P \in \mathbb{R}^{n \times n}$ is a projection matrix and independent of X_T. Then, for any $t \geq \log(es)$,*

$$\mathbb{P}\left(\|X_T^\top P\varepsilon\|_\infty \geq C\kappa\sqrt{n}t\sigma^2 \right) \leq e^{-n} + e^{-Ct}.$$

Proof of Lemma 9. For fixed vector $w \in \mathbb{R}^n$, since Assumption 2 is satisfied, for $i \in T$, $X_{1i}, \ldots, X_{ni}$ are independent sub-Gaussian distributed such that

$$\mathbb{E}\exp\left(tX_{ji}\right) = \mathbb{E}\exp\left(te_i^\top \Sigma^{1/2}\Sigma^{-1/2}X_{j\cdot}^\top \right) \leq \exp\left(\frac{\kappa^2\|\Sigma^{1/2}e_i\|_2^2 t^2}{2} \right) \leq \exp\left(\frac{\kappa^2\Sigma_{i,i}t^2}{2} \right) \leq \exp\left(\frac{3\kappa^2 t^2}{4} \right).$$

By Hoeffding-type inequality,

$$\mathbb{P}\left(|X_{\cdot i}^\top w| \geq t\|w\|_2 \right) \leq 2\exp\left(-c\frac{t^2}{\kappa^2} \right). \tag{96}$$

Moreover, by [61, Lemma 1], for any $x \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i^2 \geq (n + 2\sqrt{nx} + 2x)\sigma^2 \right) \leq e^{-x}.$$

Set $x = n$ in the last inequality, we have

$$\mathbb{P}\left(\|\varepsilon\|_2 \geq \sqrt{5n}\sigma^2 \right) \leq e^{-n}. \tag{97}$$

Combine (96) and (97) together and notice that $\|P\varepsilon\|_2 \leq \|\varepsilon\|_2$, we have

$$\mathbb{P}\left(\|X_T^\top P\varepsilon\|_\infty \geq C\kappa\sqrt{n}t\sigma^2 \right) \leq \sum_{i \in T} \mathbb{P}\left(|X_{\cdot i}^\top P\varepsilon| \geq C\kappa\sqrt{n}t\sigma^2 \right)$$

$$\leq \mathbb{P}\left(\|P\varepsilon\|_2 \geq \sqrt{5n}\sigma^2 \right) + \sum_{i \in T} \mathbb{P}\left(|X_{\cdot i}^\top P\varepsilon| \geq C\kappa\sqrt{n}t\sigma^2, \|P\varepsilon\|_2 \leq \sqrt{5n}\sigma^2 \right)$$

$$\leq \mathbb{P}\left(\|\varepsilon\|_2 \geq \sqrt{5n}\sigma^2 \right) + \sum_{i \in T} \mathbb{P}\left(|X_{\cdot i}^\top P\varepsilon| \geq C\kappa\sqrt{n}t\sigma^2 \Big| \|P\varepsilon\|_2 \leq \sqrt{5n}\sigma^2 \right)$$

$$\leq e^{-n} + s \cdot 2\exp\left(-Ct \right) \leq e^{-n} + e^{-Ct}.$$

$\square$

Lemma 10 *With probability at least $1 - Ce^{-cn/s}$, the approximate dual certificate defined in (42) can be written as $u = X^\top w$, where $\|w\|_2 \leq C\sqrt{s/n}$.*

56

Proof of Lemma 10. By (42), we have $u = X^\top w$, where $w = (w_1^\top, \ldots, w_{l_{\max}}^\top)^\top$ and $w_l = \frac{1}{n_l} X_{I_l,T} \Sigma_{T,T}^{-1} q_{l-1}$. Thus $\|w\|_2^2 = \sum_{l=1}^{l_{\max}} \|w_l\|_2^2$. Also note that

$$
\begin{aligned}
\frac{1}{n_l} \|X_{I_l,T} \Sigma_{T,T}^{-1} q_{l-1}\|_2^2 =& \langle \frac{1}{n_l} X_{I_l,T}^\top X_{I_l,T} \Sigma_{T,T}^{-1} q_{l-1}, \Sigma_{T,T}^{-1} q_{l-1} \rangle \\
=& \langle (\frac{1}{n_l} X_{I_l,T}^\top X_{I_l,T} \Sigma_{T,T}^{-1} - I_{|T|}) q_{l-1}, \Sigma_{T,T}^{-1} q_{l-1} \rangle + \|\Sigma_{T,T}^{-1/2} q_{l-1}\|_2^2 \\
=& \langle -q_l, \Sigma_{T,T}^{-1} q_{l-1} \rangle + \|\Sigma_{T,T}^{-1/2} q_{l-1}\|_2^2 \\
\leq& \|q_l\|_2 \|\Sigma_{T,T}^{-1} q_{l-1}\|_2 + \|\Sigma_{T,T}^{-1/2} q_{l-1}\|_2^2 \\
\leq& \frac{3}{2} \|q_l\|_2 \|q_{l-1}\|_2 + \frac{3}{2} \|q_{l-1}\|_2^2 \leq 3 \|q_{l-1}\|_2^2.
\end{aligned}
$$

By (50), with probability at least $1 - C \exp(-cn/s)$,

$$
\begin{aligned}
\|w\|_2^2 \leq& \sum_{l=1}^{l_{\max}} \frac{3}{n_l} \|q_{l-1}\|_2^2 \\
\leq& \frac{C}{n} (2\sqrt{s})^2 + \frac{C}{n} \left(2\sqrt{s/\log(es)}\right)^2 + \frac{C\log(es)}{n} \sum_{l=3}^{l_{\max}} \left(2^{4-l}\sqrt{s}/\log(es)\right)^2 \\
\leq& C\frac{s}{n}.
\end{aligned}
$$

$\square$

Proof of Lemma 4. For any $1 \leq i \leq p$, $1 \leq j \leq d$, $, \Lambda \subseteq (j), |\Lambda| = k$, by Lemma 6 with

$$
v = \Sigma^{-1} e_i, \quad U \in \mathbb{R}^{p \times k}, U_{[\Lambda,:]} = I, U_{[\Lambda^c,:]} = 0,
$$

we have

$$
\mathbb{P}\left(\left\| (e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i \right\|_2 \geq t \right) \leq 2 \exp\left(Ck - cn \min\left\{ \frac{t^2}{\kappa^4}, \frac{t}{\kappa^2} \right\} \right). \tag{98}
$$

By the same method in Lemma 5 Part 2,

$$\mathbb{P}\left(\left\|H_\alpha\left((e_i)_{(j)} - \frac{1}{n}X_{(j)}^\top X\Sigma^{-1}e_i\right)\right\|_2 \geq \gamma\right)$$

$$\leq \mathbb{P}\left(\exists\Lambda \subseteq (j),\ \text{all entries of }|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i| \geq \alpha \text{ and } \|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2 \geq \gamma\right)$$

$$\leq \mathbb{P}\left(\exists\Lambda \subseteq (j),\ \sqrt{|\Lambda|}\alpha \leq \gamma,\ \|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2 \geq \gamma\right)$$

$$+ \mathbb{P}\left(\exists\Lambda \subseteq (j),\ \sqrt{|\Lambda|}\alpha > \gamma,\ \text{all entries of }|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i| \geq \alpha\right)$$

$$\leq \sum_{\substack{\Lambda\subseteq(j) \\ |\Lambda|=\lfloor s/s_g\rfloor}} \mathbb{P}\left(\|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2 \geq \gamma\right)$$

$$+ \sum_{\substack{\Lambda\subseteq(j) \\ |\Lambda|=\lceil s/s_g\rceil}} \mathbb{P}\left(\text{all entries of }|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i| \geq \alpha\right)$$

$$\leq \sum_{\substack{\Lambda\subseteq(j) \\ |\Lambda|=\lfloor s/s_g\rfloor}} \mathbb{P}\left(\|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2 \geq \gamma\right) + \sum_{\substack{\Lambda\subseteq(j) \\ |\Lambda|=\lceil s/s_g\rceil}} \mathbb{P}\left(\|(e_i)_\Lambda - \frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2 \geq \gamma\right).$$

Combine (98) and the previous inequality together, we have

$$\mathbb{P}\left(\left\|H_\alpha\left((e_i)_{(j)} - \frac{1}{n}X_{(j)}^\top X\Sigma^{-1}e_i\right)\right\|_2 \geq \gamma\right)$$

$$\leq \binom{b_j}{\lfloor s/s_g\rfloor} \cdot 2\exp\left(C\lfloor s/s_g\rfloor - cn\cdot C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g n}\right)$$

$$+ \binom{b_j}{\lceil s/s_g\rceil} \cdot 2\exp\left(C\lceil s/s_g\rceil - cn\cdot C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g n}\right) \tag{99}$$

$$\leq 4\left(\frac{2es_g b}{s}\right)^{2s/s_g}\exp\left(Cs/s_g - C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right)$$

$$\leq 4\exp\left(\frac{2s}{s_g}\log\left(\frac{2es_g b}{s}\right) + Cs/s_g - C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right).$$

By (99) and the union bound, we have

$$\mathbb{P}\left(\max_{1\leq i\leq p}\|H_\alpha(e_i - \frac{1}{n}X^\top X\Sigma^{-1}e_i)\|_{\infty,2} \leq \gamma\right)$$

$$\leq \sum_{i=1}^p\sum_{j=1}^d \mathbb{P}\left(\left\|H_\alpha\left((e_i)_{(j)} - \frac{1}{n}X_{(j)}^\top X\Sigma^{-1}e_i\right)\right\|_2 \geq \gamma\right)$$

$$\leq d^2 b\cdot 4\exp\left(\frac{2s}{s_g}\log\left(\frac{2es_g b}{s}\right) + Cs/s_g - C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right)$$

$$\leq 4\exp\left(2\log(s_g) + 2\log(d/s_g) + \frac{3s}{s_g}\log(2eb) + Cs/s_g - C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right)$$

$$\leq 4\exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right).$$

$\square$