Journal of Machine Learning Research 25 (2024) 1-70 Submitted 9/22; Revised 2/24; Published 4/24

The Non-Overlapping Statistical Approximation to Overlapping Group Lasso

Mingyu Qi MQ3SQ@VIRGINIA.EDU
Department of Statistics
University of Virginia
Charlottesville, VA 22904, USA

Tianxi Li TIANXILI@UMN.EDU
School of Statistics
University of Minnesota, Twin Cities
Minneapolis, MN 55455, USA

Editor: Francis Bach

Abstract

The group lasso penalty is widely used to introduce structured sparsity in statistical learning, characterized by its ability to eliminate predefined groups of parameters automatically. However, when the groups overlap, solving the group lasso problem can be time-consuming in high-dimensional settings due to groups' non-separability. This computational challenge has limited the applicability of the overlapping group lasso penalty in cutting-edge areas, such as gene pathway selection and graphical model estimation. This paper introduces a non-overlapping and separable penalty designed to efficiently approximate the overlapping group lasso penalty. The approximation substantially enhances the computational efficiency in optimization, especially for large-scale and high-dimensional problems. We show that the proposed penalty is the tightest separable relaxation of the overlapping group lasso norm within the family of ℓ_{q_1}/ℓ_{q_2} norms. Moreover, the estimators derived from our proposed norm are statistically equivalent to those derived from the overlapping group lasso penalty in terms of estimation error, support recovery, and minimax rate under the squared loss. The effectiveness of our method is demonstrated through extensive simulation examples and a predictive task of cancer tumors.

Keywords: overlapping group lasso, separable approximation, computational efficiency, statistical error bound, support recovery, high-dimensional regression

1. Introduction

Grouping patterns of variables are commonly observed in real-world applications. For example, in regression modeling, explanatory variables might belong to different groups with the expectation that the variables within the same group are highly correlated. In this context, variable selection or model regularization should also consider the grouping patterns, and one may prefer to either include the entire group of variables in the selection or completely exclude the group. Group lasso (Yuan and Lin, 2006) is one popular method designed for this group selection task via adding ℓ_1/ℓ_2 regularization, and is part of a broader class for group selection (Bach, 2008; Levina et al., 2008; Meier et al., 2008; Ravikumar et al., 2009; Zhao et al., 2009b; Danaher et al., 2014; Loh, 2014; Basu et al., 2015; Xiang et al., 2015; Campbell and Allen, 2017; Tank et al., 2017; Yan and Bien, 2017; Austin et al., 2020; Yang and Peng, 2020).

While the original group lasso penalty (Yuan and Lin, 2006) focuses on regularizing disjoint parameter groups, overlapping groups appear frequently in many applications such as tumor metastasis analysis (Jacob et al., 2009; Zhao et al., 2009b; Yuan et al., 2011; Chen et al., 2012) and structured model selection problems (Mohan et al., 2014; Cheng et al., 2017; Yu and Bien, 2017; Tarzanagh and Michailidis, 2018). For example, in tumor metastasis analysis, scientists usually aim to select a small number of tumor-related genes. Biological theory suggests that rather than functioning in isolation, genes act in groups to perform biological functions. Therefore, the gene selection is more meaningful when co-functioning groups of genes are selected together (Ma and Kosorok, 2010). In particular, gene pathways, which represent overlapping groups of genes, render mechanistic insights into the co-functioning patterns. Applying group lasso with these overlapping groups is thus a natural way to incorporate the prior group information into tumor metastasis analysis. For another example, graphical models have been widely used to represent conditional dependency structures among variables. Cheng et al. (2017) developed a mixed graphical model for high-dimensional data with both continuous and discrete variables. In their model, groups of parameters corresponding to each edge emerged naturally, with these groups overlapping as edges share common nodes. Selecting the graph structures under this class of models requires the elimination of groups of parameters, which is achieved by the overlapping group lasso penalty.

The optimization involving the group lasso penalty with non-overlapping groups is efficient (Friedman et al., 2010; Qin et al., 2013; Yang and Zou, 2015). However, the overlapping group lasso problems present more complex challenges despite their convex nature. This complexity arises because the non-separability between groups intrinsically increases the problem's dimensionality compared with the non-overlapping situation (Yan and Bien, 2017). Proposed methods for such optimization problems include the second-order cone program method, SLasso (Jenatton et al., 2011a), the ADMM-based methods (Boyd et al., 2011; Deng et al., 2013), and their smoothed improvement, FoGLasso (Yuan et al., 2011). Nevertheless, these exact solvers involve expensive calculations when the overlapping becomes severe, which may limit the applicability of the overlapping group lasso penalty in many large-scale applications such as genome-wide association studies (Yang et al., 2010; Lee and Xing, 2012, 2014) or graphical model fitting problems (Cheng et al., 2017). For instance, Cheng et al. (2017) showed that although overlapping group lasso is a natural choice for their problem, it is infeasible even for estimating moderate-size graphs. Instead, they used a fast lasso approach (Tibshirani, 1996) to solve the graph estimation problem without theoretical support. As we introduce later, our proposed solution includes the method of Cheng et al. (2017) as a special case, but our method is more general and comes with theoretical guarantees.

In this paper, we propose a non-overlapping approximation alternative to the overlapping group lasso penalty. The approximation is formulated as a weighted non-overlapping group lasso penalty that respects the original overlapping group patterns, thereby simplifying the optimization significantly. The proposed penalty is shown to be the tightest separable relaxation of the original overlapping group lasso penalty within a broad family of penalties. Our analysis reveals that the estimator derived from our method is statistically equivalent to the original overlapping group lasso estimator in terms of estimator error and support recovery. The practical effectiveness of our proposed method is demonstrated through simulation examples and its application to a predictive task involving a breast cancer gene dataset. As a high-level summary, our major contribution to the paper is the design of a novel approximation penalty to the overlapping group lasso penalty, which enjoys substantially better computational efficiency in optimization while maintaining equivalent statistical properties to the original penalty.

The remainder of this paper is organized as follows: Section 2 introduces the overlapping group lasso problem and the proposed approximation method. We also establish the optimality of the proposed penalty from the optimization perspective. Section 3 details the statistical properties of the penalized estimator derived from the proposed penalty. Comparisons between our estimator and the original overlapping group lasso estimator are made to demonstrate their statistical equivalence in terms of estimation errors and variable selection performance. Empirical evaluations using both simulated and real breast cancer gene expression data are presented in Sections 5 and 6, respectively. Finally, Section 7 concludes the paper with additional discussions.

2. Methodology

Notation and Preliminaries. Throughout this paper, given a positive integer z, we define $[z] = \{1, 2, \ldots, z\}$. For a vector $x \in \mathbb{R}^p$, we define $\|x\|_z = (|x_1|^z + |x_2|^z + \ldots + |x_p|^z)^{\frac{1}{z}}$. Given a set T, $|T|$ represents the cardinality. When referring to a matrix A, A_T denotes the sub-matrix consisting of columns indexed by T, and $A_{T,T}$ denotes the sub-matrix induced by both rows and columns indexed by T. The operator norm is defined as: $\|A\|_{a,b} = \sup_{\|u\|_a \leq 1} \|Au\|_b$. When A is a symmetric matrix, $\gamma_{\min}(A)$ and $\gamma_{\max}(A)$ denote its smallest and largest eigenvalues, respectively. Given two sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \leqslant Cb_n$ for a sufficiently large n and a universal constant $C > 0$. We write $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n/b_n \to 0$. Furthermore, $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We will introduce other notations within the text as needed. Table 9 in Appendix A lists all the notations used in the paper.

2.1 Overlapping Group Lasso

In a statistical learning problem, consider parameters represented by a vector $\beta \in \mathbb{R}^p$, with β_j representing the j-th element of β. Let $G = \{G_1, \cdots, G_m\}$ be m predefined groups for the p parameters, where each group G_g is a subset of $[p]$ and $\cup_{g\in[m]} G_g = [p]$. For each group G_g, $d_g^G = |G_g|$ denotes the group size, with $d_{\max}^G = \max_{g\in[m]} d_g^G$. For any set $T \subset [p]$, β_T denotes the subvector of β indexed by T. Let $w = \{w_1, \cdots, w_m\}$ be the user-defined positive weights associated with the groups. The group lasso penalty (Yuan and Lin, 2006) is defined as:

$$\phi^G(\beta) = \sum_{g\in[m]} w_g \left\|\beta_{G_g}\right\|_2. \tag{1}$$

We will omit G in all notations when the group structure is clear.

In statistical estimation problems that involve group selection, the group lasso norm is combined with a convex empirical loss function L_n, and the estimator is determined by solving the following M-estimation problem:

$$\text{minimize}_{\beta\in\mathbb{R}^p} \big\{L_n(\beta) + \lambda_n \phi(\beta)\big\}. \tag{2}$$

When the groups are disjoint, the group lasso penalty selects and eliminates variables by groups. When the groups overlap, the above estimation enforces an “all-out” pattern by simultaneously setting all variables within certain groups to zero, thus the zero-out variables are form a union of a subset of the groups (Jenatton et al., 2011a). Such a pattern is desirable in various applications, such as graphical models, multi-task learning, and gene analysis (Jacob et al., 2009; Zhao et al., 2009b; Mohan et al., 2014; Cheng et al., 2017; Tarzanagh and Michailidis, 2018). Another generalization of the group lasso for overlapping groups is the latent overlapping group lasso (Jacob et al., 2009; Mairal and Yu, 2013), which follows an “all-in” pattern by maintaining the nonzero patterns as a union of groups. As noted by Yan and Bien (2017), the decision to adopt an "all-in" or "all-out" strategy depends on the problem and the corresponding scientific interpretation. The comparison between these two strategies is not our objective. However, both methods suffer from computational difficulties. Our emphasis in this paper is on introducing an approximation method for the overlapping group lasso penalty (1), leaving the computational improvements of the latent overlapping group lasso for future work.

Problem (2) is a non-smooth convex optimization problem (Jenatton et al., 2011a; Chen et al., 2012), and the proximal gradient method (Beck and Teboulle, 2009; Nesterov, 2013) is one of the most general yet efficient strategies to solve it. Intuitively, proximal gradient descent iteratively minimizes the objective by applying the proximal operator of $\lambda_n \phi(\beta)$ at each step. The proximal operator associated with the group lasso penalty in (1) is:

$$\text{prox}_{\lambda_n}(\mu) = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \Big\{ \frac{1}{2} \|\mu - \beta\|^2 + \lambda_n \phi(\beta) \Big\}, \tag{3}$$

whose dual problem has been shown by Jenatton et al. (2011b) to be:

$$\underset{\{\xi^g \in \mathbb{R}^p\}_{g \in [m]}}{\text{minimize}} \Big(\frac{1}{2} \|\mu - \sum_{g=1}^{m} \xi^g \|_2^2 \Big), \quad \text{s.t. } \|\xi^g\|_2 \leq \lambda_n w_g, \text{ and } \xi^g_j = 0 \text{ if } j \notin G_g. \tag{4}$$

The proximal operator (3) and its dual can be computed using a block coordinate descent (BCD) algorithm (Jenatton et al., 2011b). We list the procedure in Algorithm 1 for readers' information. The convergence of this algorithm is guaranteed by Bertsekas (1997, Proposition 2.7.1).

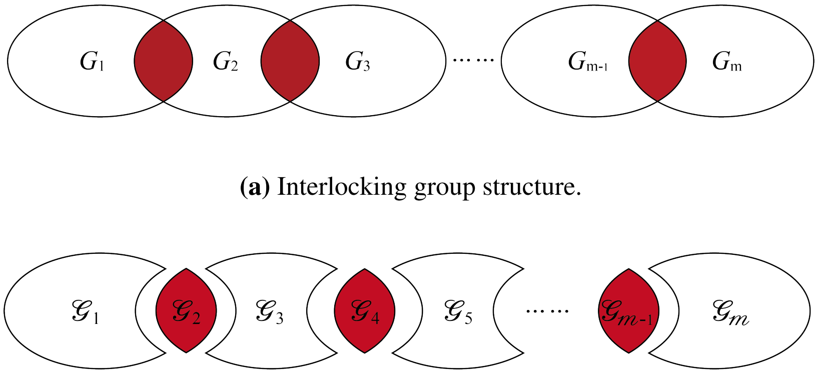
Algorithm 1 BCD algorithm for the proximal operator of the overlapping group lasso

```
Input: G, $\{w_g\}_{g=1}^m > 0$, $u$, $\lambda_n > 0$.
Output: $\beta^*$.
Initialization: $\{\xi^g\}_{g=1}^m = \mathbf{0} \in \mathbb{R}^p$.
while stopping criterion not reached do
    for all $g \in \{1, \cdots, m\}$ do
        Calculate $r^g = \mu - \sum_{h \neq g} \xi^h$.
        if $||r^g||_2 \leqslant \lambda_n w_g$ then $\xi^g_j = \begin{cases} 0 & \text{if } j \notin G_g \\ r^g_j & \text{if } j \in G_g \end{cases}$
        else $\xi^g_j = \begin{cases} 0 & \text{if } j \notin G_g \\ \frac{\lambda w_g r^g_j}{||r^g||_2} & \text{if } j \in G_g \end{cases}$
        end if
    end for
end while
$\beta^* = u - \sum_{g \in [m]} \xi^g$.
```

Although additional techniques that employ smoothing techniques have been developed to improve optimization (Yuan et al., 2011; Chen et al., 2012), (3) and (4) still offer crucial insights into the computational bottlenecks caused by overlapping groups. Notably, the duality between (3) and (4) reveals that the intrinsic dimension of the overlapping group lasso problem is equivalent to that of a $\sum_{g \in [m]} d_g$-dimensional separable problem. When the groups contain a nontrivial proportion of overlapping variables, computing the overlapping group lasso problems becomes substantially more difficult, eventually prohibitive in large-scale problems. This issue significantly limits the applicability of the overlapping group lasso penalty. Next, we introduce our non-overlapping approximation to rectify this challenge.

2.2 The Non-overlapping Approximation of the Overlapping Group Lasso

The fundamental challenge in solving overlapping group lasso problems stems from the non-separability of the penalty. Thus, to enhance computational efficiency, our approach hinges on introducing separable operators. As a starting point, we will illustrate this concept with a toy example of an interlocking group structure. In this structure, the groups are arranged sequentially, each overlapping with its adjacent neighbors (Figure 1a). For simplicity, we consider a scenario where the weight for all groups is uniformly set to $w_g \equiv 1$.



(a) Interlocking group structure.

(b) Partitioned group structure.

Figure 1: Illustration of proposed group partition in an interlocking group structure. Red regions are the overlapping variables in the original group structure.

We now partition the original overlapping groups in Figure 1b into smaller groups as in Figure 1b. This partition treats intersections as individual groups. We define these new groups as $\mathscr{G} = \{\mathscr{G}_1, \cdots, \mathscr{G}_{\mathscr{m}}\}$, where, for this specific instance, $\mathscr{m} = 2m - 1$. Taking G_1 as an example, it comprises $G_1 = \mathscr{G}_1 \cup \mathscr{G}_2$. By the triangle inequality, we have:

$$\|\beta_{G_1}\|_2 \leq \|\beta_{\mathscr{G}_1}\|_2 + \|\beta_{\mathscr{G}_2}\|_2.$$

Extending this principle to each group, the norm of the overlapping group lasso based on G can be bounded by a reweighted non-overlapping group norm based on $\mathscr{G}$:

$$\sum_{g \in [m]} \|\beta_{G_g}\|_2 \leq \sum_{\mathscr{g} \in [\mathscr{m}]} w_{\mathscr{g}} \|\beta_{\mathscr{G}_{\mathscr{g}}}\|_2, \tag{5}$$

where $w_{\mathscr{g}}$ equals 1 for odd $\mathscr{g}$ and 2 for even $\mathscr{g}$. Consequently, controlling the sum on the right-hand side of (5) effectively controls the overlapping group lasso norm on the left-hand side. The key advantage of this approach is the separability of the right-hand side norm, which substantially enhances the efficiency of optimization.

While the previous example focuses on interlocking group structures, the underlying idea is applicable to any general overlapping pattern, as we introduce in the following two steps.

Step 1: Overlapping-induced partition construction. The predefined group structure G can be represented by a $m \times p$ binary matrix $\mathbf{G}$, where $\mathbf{G}_{gj} = 1$ if and only if the j-th variable is a member of the g-th group, and $\mathbf{G}_{gj} = 0$ otherwise. Our method starts from constructing a new non-overlapping group structure $\mathscr{G}$ from G, detailed in Algorithm 2. To distinguish clearly between the original group structure G and the derived non-overlapping structure $\mathscr{G}$, standard letters, such as $\{g, d, m, w, G\}$, are used to denote quantities related to the original group structure, while calligraphic letters, like $\{\mathscr{g}, \mathscr{d}, \mathscr{m}, \mathscr{w}, \mathscr{G}\}$, are used to represent quantities about $\mathscr{G}$. For instance, $\mathscr{m}$ denotes the number of groups in $\mathscr{G}$, and $\mathscr{g} \in [\mathscr{m}]$ serves as the index for groups within $\mathscr{G}$.

Algorithm 2 Algorithm to construct the overlapping-induced partition $\mathscr{G}$

```
Input: Binary matrix $\mathbf{G}$.
Output: New group structure $\mathscr{G}$.
Initialize the column index set as $C = \{1, \ldots, p\}$.
Initialize $k = 1$.
while $C$ is not empty do
    Choose the first column index $j$ in $C$, and set $I$ to be the set of all column indices in $G$ identical to $G_{,j}$: $I = \{j' \in C, G_{,j'} = G_{,j}\}$.
    Set $\mathscr{G}_k = I$, and remove $I$ from $C$: $C \leftarrow C \setminus I$.
    $k = k + 1$.
end while
Return $\mathscr{G} \leftarrow \{\mathscr{G}_1, \mathscr{G}_2, \cdots, \mathscr{G}_{\mathscr{m}}\}$.
```

Step 2: Overlapping-based group weights calculation. Note that each group within $\mathscr{G}$ is a subset of at least one of the original groups in G. Conversely, each group in G can be reconstructed as the union of several groups in $\mathscr{G}$. Consequently, we define two mappings:

$$F(\mathscr{g}) = \{g : g \in [m], \mathscr{G}_{\mathscr{g}} \subset G_g\} \quad \text{and} \quad F^{-1}(g) = \{\mathscr{g} : \mathscr{g} \in [\mathscr{m}], \mathscr{G}_{\mathscr{g}} \subset G_g\}.$$

Given the positive weights w of G, we set the weights $\mathscr{w}$ of $\mathscr{G}$ as:

$$\mathscr{w}_{\mathscr{g}} = \sum_{g \in F(\mathscr{g})} w_g, \quad \mathscr{g} \in [\mathscr{m}]. \tag{6}$$

With the new partition $\mathscr{G}$ and the new weights $\mathscr{w}$ from the previous two steps, we define the following norm as the proposed alternative to the original overlapping group lasso norm:

$$\psi^{\mathscr{G}}(\beta) = \sum_{\mathscr{g}=1}^{\mathscr{m}} \mathscr{w}_{\mathscr{g}} \left\|\beta_{\mathscr{G}_{\mathscr{g}}}\right\|_2. \tag{7}$$

In general, by the triangle inequality, the proposed norm is always an upper bound of the original group lasso norm:

$$\phi^G(\beta) = \sum_{g=1}^{m} w_g \left\|\beta_{G_g}\right\|_2 \leqslant \sum_{\mathscr{g}=1}^{\mathscr{m}} \mathscr{w}_{\mathscr{g}} \left\|\beta_{\mathscr{G}_{\mathscr{g}}}\right\|_2 = \psi^{\mathscr{G}}(\beta). \tag{8}$$

Our proposed penalty is essentially a weighted non-overlapping group lasso on $\mathscr{G}$. For illustration, Figure 2 shows the unit ball of these two norms based on $G_1 = \{\beta_1, \beta_2\}$ and $G_2 = \{\beta_1, \beta_2, \beta_3\}$ in a three dimensional problem. All singular points of the ϕ^G-ball (where exactly zero happens in (2)) are also singular points of the $\psi^{\mathscr{G}}$-ball.

Readers may observe that the inequality in (8) could apply to other separable norms. For instance, consider partitioning all p variables into individual groups and using a weighted lasso norm as another upper bound for ϕ^G, represented by:

$$\sum_{j=1}^{p} \Big(\sum_{\{g|\beta_j \in G_g\}} w_g \Big) |\beta_j|. \tag{9}$$

This approach to employing a weighted lasso norm was previously explored by Cheng et al. (2017). We will now explain what makes our proposed norm in (7) special.

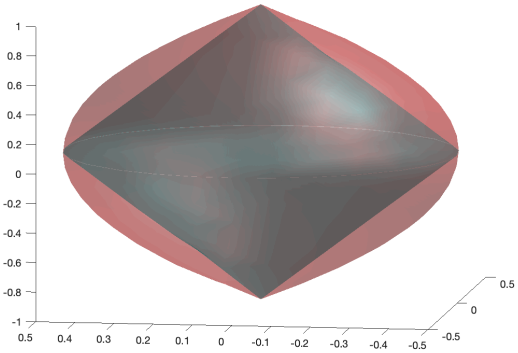


Figure 2: Illustration of two norms in $\mathbb{R}^3$: The outer region depicts the unit ball of the overlapping group lasso norm defined by $\{\beta : \phi^G(\beta) \leqslant 1\}$; The inner region represents the unit ball of our proposed separable norm $\{\beta : \psi^{\mathscr{G}}(\beta) \leqslant 1\}$.

Intuitively, as illustrated by our construction process for $\mathscr{G}$ (Figure 2), our method introduces additional singular points in the norm only when necessary to achieve separability. Unlike the lasso upper bound, this approach avoids adding redundancy. As such, our approximation is expected to maintain a certain level of tightness. We now formally substantiate this intuition. Given any group structure G and weights w, following Cai et al. (2022), we define the ℓ_{q_1}/ℓ_{q_2} norm of β for any $0 \leqslant q_1, q_2 \leqslant \infty$ as:

$$||\beta_{\{G,w\}}||_{q_1,q_2} = \Big(\sum_{g \in [m]} w_g ||\beta_{G_g}||_{q_2}^{q_1} \Big)^{\frac{1}{q_1}}. \tag{10}$$

This general class of norms potentially includes most commonly used penalties, such as the weighted lasso penalty. The subsequent theorem demonstrates that the proposed $\psi^{\mathscr{G}}(\beta)$ is the tightest separable relaxation of the original overlapping group lasso norm among all separable ℓ_{q_1}/ℓ_{q_2} norms.

Theorem 1. *Let $\mathbb{G}$ represent the set of all possible partitions of $[p]$. Given the original groups G and group weights w, there does not exist $0 \leqslant q_1, q_2 \leqslant \infty, \tilde{G} \in \mathbb{G}, \tilde{w} \in (0, \infty)^p$ such that:*

$$\begin{cases} \phi^G(\beta) \leqslant ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} \leqslant \psi^{\mathscr{G}}(\beta) & \textit{for all } \beta \in \mathbb{R}^p \\ ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} < \psi^{\mathscr{G}}(\beta) & \textit{for some } \beta \in \mathbb{R}^p \end{cases}. \tag{11}$$

3. Statistical Properties

Incorporating the proposed norm $\psi^{\mathscr{G}}$ into an M-estimation procedure leads to the following optimization problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \{L_n(\beta) + \lambda_n \psi^{\mathscr{G}}\}, \tag{12}$$

which is different but related to (2). In this section, we explore the statistical properties of the regularized estimator based on $\psi^{\mathscr{G}}$ and the estimator based on ϕ^G, demonstrating that $\psi^{\mathscr{G}}$ could serve as an effective alternative to ϕ^G. Following previous group lasso studies (Huang and Zhang, 2010; Lounici et al., 2011; Chen et al., 2012; Negahban et al., 2012; Dedieu, 2019), our analysis will focus on high-dimensional linear models. Specifically, the linear model is defined as:

$$Y = X\beta^* + \varepsilon, \tag{13}$$

where $Y \in \mathbb{R}^{n\times 1}$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the covariate matrix, and $\varepsilon \in \mathbb{R}^{n\times 1}$ is a random noise vector. The overlapping group lasso estimator under the linear regression model is defined by a solution of (2) under the squared loss:

$$\hat{\beta}^G \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \ \Big\{\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n \phi^G(\beta)\Big\}. \tag{14}$$

Correspondingly, we define the regularized estimator by our approximation norm as:

$$\hat{\beta}^{\mathscr{G}} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \ \Big\{\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n \psi^{\mathscr{G}}(\beta)\Big\}. \tag{15}$$

The solution uniqueness of (14) and (15) has been studied by Jenatton et al. (2011a), and we include their results in Appendix B for completeness. However, our study only requires the estimator to be one solution to the problem, as in Jenatton et al. (2011a); Negahban et al. (2012); Wainwright (2019). Therefore, we will not specifically focus on the uniqueness in our discussion.

As a remark, our objective is **not** to present (15) as an approximate optimization problem of (14). Instead, we aim to establish the statistical equivalence of the two classes of estimators defined by (14) and (15) in terms of their statistical properties under sparse regression models when appropriate values of λ_n are chosen (which may differ for each estimator). Our theoretical analysis focuses on three aspects. In Section 3.1, we establish that under reasonable assumptions, the ℓ_2 estimation error bound for (15) is no larger than that for (14). In Section 3.2, we present the minimax error rate for the overlapping sparse group regression problem, showing that both (14) and (15) achieve minimax optimality under specific requirements of the group structures. Lastly, in Section 3.3, we demonstrate that both estimators consistently recover the support of the sparse β^* with high probability under similar sample size requirements.

3.1 Estimation Error Bounds

We start by introducing additional quantities. Define the overlapping degree h_j^G as the number of groups in G that contain β_j, with $h_{\max}^G = \max h_j$. Given a group index set $I \subseteq [m]$, we use G_I to denote the union $\bigcup_{g \in I} G_g$. Given G and I, following Wainwright (2019), we define two parameter spaces:

$$M(I) = \{\beta \in \mathbb{R}^p \mid \beta_j = 0 \text{ for all } j \in (G_I)^c\},$$
$$M^{\perp}(I) = \{\beta \in \mathbb{R}^p \mid \beta_j = 0 \text{ for all } j \in G_I\},$$

and we further use $\beta_{M(I)}$ to denote the projection of β onto $M(I)$.

Given a set $T \subseteq [p]$, we define a set of groups $\mathsf{G}_T = \{g \in [m] \mid G_g \cap T \neq \emptyset\}$. Note that $(G_{\mathsf{G}_T})^c$ is referred to as the hull of T in Jenatton et al. (2011a). Let $\mathit{supp}(\beta) = \{j \in [p] \mid \beta_j \neq 0\}$ denote the support set. We define the group support set $S^G(\beta) = \mathsf{G}_{\mathit{supp}(\beta)}$, and the augmented group support $\overline{S^G(\beta)} = \{g \in [m] \mid G_g \cap G_{S(\beta)} \neq \emptyset\}$. Furthermore, we define $s = |\mathit{supp}(\beta)|$, $s_g = |S(\beta)|$, and $\overline{s_g} = |\overline{S(\beta)}|$. We omit the subscript G in notations when G is clearly given in context. Now, we introduce additional assumptions under the regression model (13).

Assumption 1 (Sub-Gaussian noise for the response variable)**.** *The coordinates of ε are $i.i.d.$ zero-mean sub-Gaussian with parameter σ. Specifically, there exists $\sigma > 0$ such that $\mathbb{E}[\exp(t\varepsilon)] \leqslant \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$.*

Our theoretical studies also hold for a fixed design of X, with trivial modifications. We prefer to introduce the random design here to make the statements more concise and interpretable, especially for the comparison in Section 3.3.

Assumption 2 (Normal random design for covariates)**.** *The rows of the data matrix X are i.i.d. from $N(0, \Theta)$, where $1/c_1 \leqslant \gamma_{\min}(\Theta) \leqslant \gamma_{\max}(\Theta) \leqslant c_1$ for some constant $c_1 > 0$.*

Assumption 3 (Dimension of the group structure)**.** *The predefined group structure G satisfies $d_{\max} \leqslant c_2 n$ for some constant $c_2 > 0$. In addition, we assume $\log m \ll n$.*

The following theorem establishes the ℓ_2 estimation error bounds for the two estimators.

Theorem 2. *Given G and its induced $\mathscr{G}$ according to Algorithm 2, define $h_{\min}^g = \min_{j \in G_g} h_j$ and $h_{\max}^g = \max_{j \in G_g} h_j$. Let $\delta \in (0,1)$ be a scalar that might depend on n. Under Assumptions 1, 2 and 3, for $\hat{\beta}^G$ and $\hat{\beta}^{\mathscr{G}}$ defined in* (14) *and* (15)*, we have the following results:*

1. *Suppose that β^* satisfies the following group sparsity condition:*

$$\overline{s_g}(\beta^*) \lesssim \frac{n}{\log m + d_{\max}} \cdot \frac{\min\limits_{g \in [m]} (w_g^2 h_{\min}^g)}{\max\limits_{g \in \overline{S}} (w_g^2 h_{\max}^g)}. \tag{16}$$

When $\lambda_n = \frac{c'\sigma}{\min\limits_{g \in [m]} (w_g^2 h_{\min}^g)} \sqrt{\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta}$ for some constant $c' > 0$, we have

$$\left\| \hat{\beta}^G - \beta^* \right\|_2^2 \lesssim \sigma^2 \cdot \frac{\left(\sum\limits_{g \in \overline{S}} w_g^2 \right) \cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g \in [m]} \left(w_g^2 h_{\min}^g \right)} \cdot \left(\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta \right). \tag{17}$$

with probability at least $1 - e^{-c_3 n \delta}$ for constant $c_3 > 0$.

2. *Suppose that β^* satisfies the group sparsity condition:*

$$\overline{s_{\mathscr{g}}}(\beta^*) \lesssim \frac{n}{\log \mathscr{m} + \mathscr{d}_{\max}} \cdot \frac{\min\limits_{\mathscr{g}\in[\mathscr{m}]}(w_{\mathscr{g}}^2)}{\max\limits_{\mathscr{g}\in S}(w_{\mathscr{g}}^2)}. \tag{18}$$

When $\lambda_n = \frac{c'\sigma}{\min\limits_{\mathscr{g}\in[\mathscr{m}]} w_{\mathscr{g}}}\sqrt{\frac{\mathscr{d}_{\max}}{n} + \frac{\log \mathscr{m}}{n} + \delta}$ for some constant $c' > 0$, we have

$$\left\|\hat{\beta}^{\mathscr{G}} - \beta^*\right\|_2^2 \lesssim \sigma^2 \cdot \frac{\sum\limits_{\mathscr{g}\in\{F^{-1}(g)\}_{g\in S}} w_{\mathscr{g}}^2}{\min\limits_{\mathscr{g}\in[\mathscr{m}]}\left(w_{\mathscr{g}}^2\right)} \cdot \left(\frac{\mathscr{d}_{\max}}{n} + \frac{\log \mathscr{m}}{n} + \delta\right). \tag{19}$$

with probability at least $1 - e^{-c_4 n\delta}$ for constant $c_4 > 0$.

The error bound in (17) subsumes the non-overlapping group lasso error bound as a particular instance. When the groups in G are disjoint, the reduced form of (17) matches the bounds studied in Huang and Zhang (2010); Lounici et al. (2011); Negahban et al. (2012); Wainwright (2019). The main difference in the context of overlapping groups is the necessity to account for the overlapping degree and the extension of sparsity requirements to augmented groups. The conditions specified in (16) and (18) relate to the cardinality of the augmented group support set (the number of non-zero groups in non-overlapping group structure). Although the conditions in (16) and (18) may initially appear distinct, they generally converge to a similar requirement in many typical cases, which can lead to an informative comparison between the two bounds in (17) and (19). The following results can characterize this.

Assumption 4. *Assume the predefined group structure G and its induced group structure $\mathscr{G}$ satisfy $\max\{d_{\max}, m\} \asymp \max\{\mathscr{d}_{\max}, \mathscr{m}\}$.*

Proposition 3. *Suppose that $\max_{g\in\overline{S}} |F^{-1}(g)|$ is bounded by a constant. Under Assumption 4, the following inequality holds:*

$$\frac{\sum\limits_{\mathscr{g}\in F^{-1}(S)} w_{\mathscr{g}}^2}{\min\limits_{\mathscr{g}\in[\mathscr{m}]}\left(w_{\mathscr{g}}^2\right)} \cdot \left(\frac{\mathscr{d}_{\max}}{n} + \frac{\log \mathscr{m}}{n} + \delta\right) \lesssim \frac{\left(\sum\limits_{g\in\overline{S}} w_g^2\right) \cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)} \cdot \left(\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta\right).$$

This implies that the error bound for the estimator $\hat{\beta}^G$ in (17) also serves as an upper bound for the error associated with the estimator $\hat{\beta}^{\mathscr{G}}$.

The quantity $|F^{-1}(g)|$ is the number of groups in $\mathscr{G}$ that has intersect with G_g. Proposition 3 requires that every G_g such that $G_g \cap \textit{supp}(\beta^*) \neq \emptyset$ is partitioned into bounded number of non-overlapping groups. On the other hand, Assumption 4 requires that the maximum of two quantities — the maximum group size and the number of groups in the given group structure G — should have the same order as those in the induced structure $\mathscr{G}$. The above requirement always holds for interlocking groups with similar groups and overlap sizes (see Figure 1). More importantly, we can always assess the assumption directly on data by calculating the group sizes and numbers for both G and $\mathscr{G}$. In Section 4.3, we evaluate five group structures from real-world gene pathways and examine the ratio of the maximum of two quantities from each G and $\mathscr{G}$. Assumption 4 looks reasonable in all of these real-world grouping structures. See details in Table 1.

3.2 Lower Bound of Estimation Error

Proposition 3 provides a comparison of the upper bounds on the estimation errors for the two estimators. While the comparison offers intuitive insights, it does not rigorously establish the statistical equivalence without the tightness of the error bounds. To strengthen our findings, we now investigate the minimax estimation error rate in linear regression models characterized by overlapping group sparsity. Our focus will be on the following class of group-wise sparse vectors:

$$\Omega(G, s_g) = \left\{ \beta : \sum_{G_g \in G} \mathbb{1}_{\{\|\beta_{G_g}\|_2 \neq 0\}} \leqslant s_g \right\}. \tag{20}$$

Following the assumptions in Cai et al. (2022), we focus on the special case of equal-size groups.

Assumption 5 (Equal-size groups)**.** *The m predefined groups of G come with equal group size d, with $m \ll p$ and $d \ll \log(p)$.*

Theorem 4 (Lower bound of estimation error)**.** *Under Assumptions 1,2 and 5, we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Omega(G, s_g)} E\|\hat{\beta} - \beta\|_2^2 \gtrsim \frac{\sigma^2 \left(s_g (d + \log(\frac{m}{s_g})) \right)}{n}. \tag{21}$$

Combining Theorem 2 and Theorem 4, we can see that both estimators attain the minimax error rate and are statistically equivalent, as demonstrated by the following corollary:

Corollary 1. *Under Assumptions 1–4, if $h_{\max}^{G_{\overline{S}}} \asymp 1$, both $\hat{\beta}^G$ and $\hat{\beta}^{\mathscr{G}}$ attain the minimax estimation rate specified in* (21)*.*

3.3 Support Recovery Consistency

We now analyze the support recovery consistency of $\hat{\beta}^G$ and $\hat{\beta}^{\mathscr{G}}$. We begin by introducing more quantities for our analysis. For any $\beta \in \mathbb{R}^p$, we define the mapping $r^G(\beta) : \mathbb{R}^p \to \mathbb{R}^p$ as follows:

$$r^G(\beta)_j = \begin{cases} \beta_j \sum\limits_{g \in \mathsf{G}_{\mathit{supp}(\beta)}, G_g \cap j \neq \emptyset} \frac{w_g}{\|\beta_{G_g \cap \mathit{supp}(\beta)}\|_2}, & \text{if } j \in \mathit{supp}(\beta), \\ 0, & \text{if } j \notin \mathit{supp}(\beta). \end{cases} \tag{22}$$

The quantity $r^G(\beta)$ is closely related to subgradients of the penalty and is used for determining optimality conditions. In the lasso case, $r^G(\beta)$ is the sign vector, which is exactly the lasso penalty. When focusing on β^*, we write $\mathbf{S} = \mathit{supp}(\beta^*)$, $\mathbf{r}^G = r^G(\beta^*)$, and $\beta^*_{\min} = \min \left\{ |\beta^*_j|; \beta^*_j \neq 0 \right\}$.

Our analysis essentially follows the strategy in Jenatton et al. (2011a). The major difference is that we study the problem with a more tailored setup for the random design rather than the fixed design as in Jenatton et al. (2011a). Using random designs, as discussed before, is helpful to compare the two estimators $\hat{\beta}^G$ and $\hat{\beta}^{\mathscr{G}}$. We now introduce additional assumptions for studying the pattern consistency, which can be seen as the population-level counterpart of the assumptions in Jenatton et al. (2011a).

Assumption 1' (Gaussian noise for the response variable)**.** *Under model* (13)*, the coordinates of ε are i.i.d from $N(0, \sigma^2)$.*

Assumption 6 (Irrepresentable condition)**.** *For any $\beta \in \mathbb{R}^p$, define:*

$$\phi^{\mathsf{c}}_{\mathbf{S}}(\beta_{\mathbf{S}^c}) = \sum_{g \in [m] \setminus \mathsf{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S}^c \cap G_g}\|_2,$$

and its dual norm:

$$(\phi^{\mathsf{c}}_{\mathbf{S}})^*[u] = \sup_{\phi^{\mathsf{c}}_{\mathbf{S}}(\beta_{\mathbf{S}^c}) \leq 1} \beta_{\mathbf{S}^c}^\top u.$$

Assume that there exists $\tau \in (0, 2/3]$, such that

$$(\phi^{\mathsf{c}}_{\mathbf{S}})^*[\Theta_{\mathbf{S}^c\mathbf{S}}\Theta^{-1}_{\mathbf{SS}}\mathbf{r}_{\mathbf{S}}] \leqslant 1 - \frac{3\tau}{2}. \tag{23}$$

Assumption 1' is widely used to study support recovery consistency of linear regression. For example, in addition to Jenatton et al. (2011a), it is also used in Zhao and Yu (2006); Wainwright (2009, 2019). Assumption 6 is the population-level version of the irrepresentable condition as discussed in Zhao and Yu (2006) and Wainwright (2019).

Theorem 5. *Suppose that Assumption 1', Assumption 2 and Assumption 6 hold. Under model* (13)*, assume the support of β^* is compatible with the overlapping group lasso penalty, such that the zero positions are given by an exact union of groups in G. Mathematically, that means:*

$$[p] \setminus \{ \bigcup_{G_g \cap \mathbf{S} = \emptyset} G_g \} = \mathbf{S}. \tag{24}$$

1. If

$$\log(p - |\mathbf{S}|) \geqslant |\mathbf{S}|,$$

$$\lambda_n |\mathbf{S}|^{\frac{1}{2}} \lesssim \min\Big\{ \frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S^c}}}{A_{\mathbf{S}} \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \Big\}, \tag{25}$$

$$n \gtrsim \max\Big\{ \frac{\sigma^2 \log(p - |\mathbf{S}|)}{a^2_{\mathbf{S}^c} \lambda_n^2}, \frac{\max_{j \in \mathbf{S}}\{(\beta^*_j)^2\} \log(p - |\mathbf{S}|)}{a^2_{\mathbf{S}^c} \lambda_n^2} \Big\}, \tag{26}$$

where $a_{\mathbf{S}} = \min_{g \in \mathsf{G}_{\mathbf{S}}} \frac{w_g}{d_g}$, $a_{\mathbf{S}^c} = \min_{g \in \mathsf{G}_{\mathbf{S^c}}} \frac{w_g}{d_g}$, and $A_{\mathbf{S}} = h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g \in \mathsf{G}_{\mathbf{S}}} w_g \|u\|_1$.

Then for the overlapping group lasso estimator $\hat{\beta}^G$, we have:

$$\begin{aligned} \mathbb{P}\Big(\mathit{supp}(\hat{\beta}^G) \neq \mathbf{S} \Big) \leqslant & 8 \exp\Big(-\frac{n}{2} \Big) + \exp\Big(-\frac{n a^2_{\mathbf{S}} \tau^2 \gamma_{\min}(\Theta_{\mathbf{SS}})}{4 \|\mathbf{r}_{\mathbf{S}}\|_2^2 \gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)} \Big) \\ & + \exp\Big(-\frac{n \lambda_n^2 \tau^2 a^2_{\mathbf{S}^c}}{144 \sigma^2} \Big) + 2|\mathbf{S}| \exp\Big(-\frac{n c^2(\mathbf{S}, G)}{2\sigma^2} \Big) \end{aligned} \tag{27}$$

with

$$c(\mathbf{S}, G) \asymp \min\Big\{ \frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S^c}}}{A_{\mathbf{S}} \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \Big\}.$$

2. *Furthermore, if* $\max_{g\in\mathsf{G}_{\mathbf{S}}} F^{-1}(g) \asymp 1$ *for the proposed estimator* $\hat{\beta}^{\mathscr{G}}$*, then the following property holds:*

$$\begin{aligned}\mathbb{P}\Big(\mathit{supp}(\hat{\beta}^{\mathscr{G}}) \neq \mathbf{S}\Big) \leqslant & 8\exp\Big(-\frac{n}{2}\Big) + \exp\Big(-\frac{na_{\mathbf{S}}^2\tau^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\left\|\mathbf{r}_{\mathbf{S}}^{\mathscr{G}}\right\|_2^2\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\Big)\\ &+\exp\Big(-\frac{n\lambda_n^2\tau^2a_{\mathbf{S}^c}^2}{144\sigma^2}\Big)+2|\mathbf{S}|\exp\Big(-\frac{nc^2(\mathbf{S},\mathscr{G})}{2\sigma^2}\Big),\end{aligned} \tag{28}$$

with

$$c(\mathbf{S},\mathscr{G}) \asymp \min\Big\{\frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min}a_{\mathbf{S^c}}}{A_{\mathbf{S}}\sum\limits_{\mathscr{g}\in\mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}}\sqrt{|\mathscr{G}_{\mathscr{g}}\cap\mathbf{S}|}}\Big\}.$$

The conditions involved in the above theorem can be seen as the population-level counterparts of those used in Jenatton et al. (2011a) for the overlapping group lasso estimator under the fixed design. As an illustration of the conditions, in the lasso context, (25) and (26) reduce to the typical scaling of $n \approx \log p$ and $\lambda_n \approx \sigma(\log p/n)^{1/2}$. Together with the requirements on the sample size $|\mathbf{S}|\log(p-|\mathbf{S}|)$ and on $\beta^*_{\min}$, they match the requirements in Wainwright (2009) for the support recovery by the lasso regression. For non-overlapping group lasso estimators, our assumptions align with the conditions outlined in Wainwright (2019, Corollary 9.27) under the random design.

Theorem 5 shows that both estimators consistently identify the support of the group sparse regression coefficients. Compared to the previous study of the overlapping group lasso estimator of Jenatton et al. (2011a), we switch to the random design of X, because such a setting renders a common basis for the comparison of the two estimators directly. Specifically, comparing (27) and (28), as well as the common conditions, we can see that the two estimators give comparable performance in support recovery with respect to the sampling complexity.

4. Comparison of Computational Complexity

In the previous section, we have shown that the proposed penalty induces a class of estimators statistically equivalent to the original overlapping group lasso estimator. In this section, we demonstrate the advantage of our proposed estimator in computational complexity. Specifically, solving (15) admits a lower complexity compared with solving (13).

As previously mentioned, the most common strategy for solving the overlapping group lasso problem is proximal-based methods (Jenatton et al., 2011b; Yuan et al., 2011; Chen et al., 2012). These algorithms involve an outer loop implementing gradient-based steps and an inner loop executing the proximal operator (3), as studied in detail by Chen et al. (2012); Yan and Bien (2017). According to Chen et al. (2012), the per-iteration time complexity for the proximal step is $O(\sum_{g\in[m]} d_g)$, and the proximal gradient method outer loops render a convergence rate of $O(1/\epsilon)$ in scenarios with overlapping groups, where ϵ denotes the desired accuracy.

In contrast, the proposed penalty converts the optimization of the overlapping group lasso problem to a non-overlapping group lasso problem. For any available proximal gradient algorithm, as the groups are disjoint, the proximal operator in (3) can be computed in closed form with the complexity of $O(p)$ for each iteration (Yuan and Lin, 2006), which gives a substantial reduction compared with the overlapping group lasso, especially when the groups in the original structure heavily overlap. Moreover, in non-overlapping scenarios, the outer loop enjoys an improved convergence rate of $O(1/\sqrt{\epsilon})$ (Liu et al., 2009a; Mairal et al., 2010). Therefore, solving (15) by proximal gradient methods enjoys better efficiency in both per-iteration complexity and number of iterations.

Furthermore, even more efficient strategies (Friedman et al., 2010; Qin et al., 2013; Yang and Zou, 2015) exists for solving the non-overlapping group lasso problem compared to proximal gradient methods. These methods offer further improvements in computational complexity. However, as far as we know, these improvement options are unavailable for solving the overlapping group lasso problem. Hence, the proposed method can enjoy the benefits of these more efficient strategies, further amplifying its computational advantage.

5. Simulation

In this section, we assess the performance of the proposed estimator to demonstrate our claimed properties. At a high level, we use simulation experiments to show that the proposed estimator based on (7) delivers similar statistical performance to the overlapping group lasso estimator while offering significantly better computational efficiency. Our estimator achieves this primarily because of the tightest separable relaxation property in Theorem 1, which can be attributed to two designs of the norm (7): the induced partition $\mathscr{G}$ and the corresponding overlapping-based weights w. Therefore, in our simulation experiments, we will also evaluate the effects of these two designs by comparing the proposed estimator with other benchmark estimators. In Sections 5.1–5.3, we evaluate the performance of the proposed estimator and compare it with the weighted lasso estimator with overlapping-based weights, as discussed in (9), under various configurations. This sequence of experiments will highlight the importance of our proposed partition $\mathscr{G}$. In Section 5.4, we compare the proposed estimator with two other group lasso estimators that use the same $\mathscr{G}$ but with overlapping-ignorant weights, under the same set of configurations. The results will demonstrate the importance of using the proposed overlapping-based weights w.

Two MATLAB-based solvers are employed for the overlapping group lasso problems. The first solver, FoGLasso (Yuan et al., 2011), is from the SLEP package (Liu et al., 2009b). It can handle general overlapping group structures. The second solver, from the SPAM package (Mairal et al., 2014), is designed to solve the overlapping group lasso problem when the groups can be represented by tree structures, as formally defined in Section 5.2. Therefore, the SPAM solver is used only for the experiment in Section 5.2. The SLEP solver is more general, but using the two solvers can provide a more thorough evaluation across multiple implementations. For a fair comparison, the SLEP and SPAM package solvers were also applied to solve lasso and non-overlapping group lasso estimators in our benchmark set to ensure that the timing comparison implementation is consistent.

As an important note, SLEP is widely acknowledged as one of the most efficient solvers for overlapping group lasso problems (Yuan et al., 2011; Chen et al., 2012). However, for non-overlapping group lasso problems, alternative solvers, such as Yang and Zou (2015), may offer much better computational efficiency. For example, Yang and Zou (2015) reported that their solver is about 10–30 times faster than the SLEP package when solving non-overlapping group lasso problems. This enhanced efficiency is possible because of the separability in non-overlapping groups, a feature not available for overlapping problems. For a fair comparison to avoid implementation bias, we use SLEP to solve our proposed estimator. Therefore, the computational advantage demonstrate here will be conservative. In practice, with the better solvers used, our method would enjoy an even more substantial computational advantage over the original overlapping group lasso than what is reported in the experiments.

Evaluation criterion. For each configuration, we generate 50 independent replicates and report the average results. The performance assessment is conducted in three aspects:

- **Regularization path computing time.** We start by determining the regularization path through a line search to identify two pivotal values: $\lambda_{\max}$ and $\lambda_{\min}$. The search for $\lambda_{\max}$ starts at 10^8 and decreases by multiplying by 0.9 at each iteration, stopping when at least one variable is selected. Conversely, the search for $\lambda_{\min}$ starts at 10^{-8} and increases by multiplying by 1.1 each time, until a value is found that does not retain all variables. We then select 50 values in log scale within the range $[\lambda_{\min}, \lambda_{\max}]$, compute the entire regularization path, and record the computation time as a performance metric. This evaluation of computing time mimics the most practical situation where the whole regularization path is solved for tuning purposes.

- **Relative ℓ_2 estimation error**: From the entire regularization path, we select the smallest relative estimation error, defined as $\|\hat{\beta} - \beta^*\|_2/\|\beta^*\|_2$, as the estimation error for the method. This serves as the measure of the ideally tuned performance.

- **Support discrepancy**: From the entire regularization path, we select the smallest support discrepancy, defined as $|\{i \in [p] : |\text{sign}(\hat{\beta}_i)| \neq |\text{sign}(\beta_i^*)|\}|/p$. Such a (normalized) Hamming distance is commonly used as a performance metric for support recovery (Grave et al., 2011; Jenatton et al., 2011a) to quantify the accuracy of pattern selection.

5.1 Interlocking Group Structure

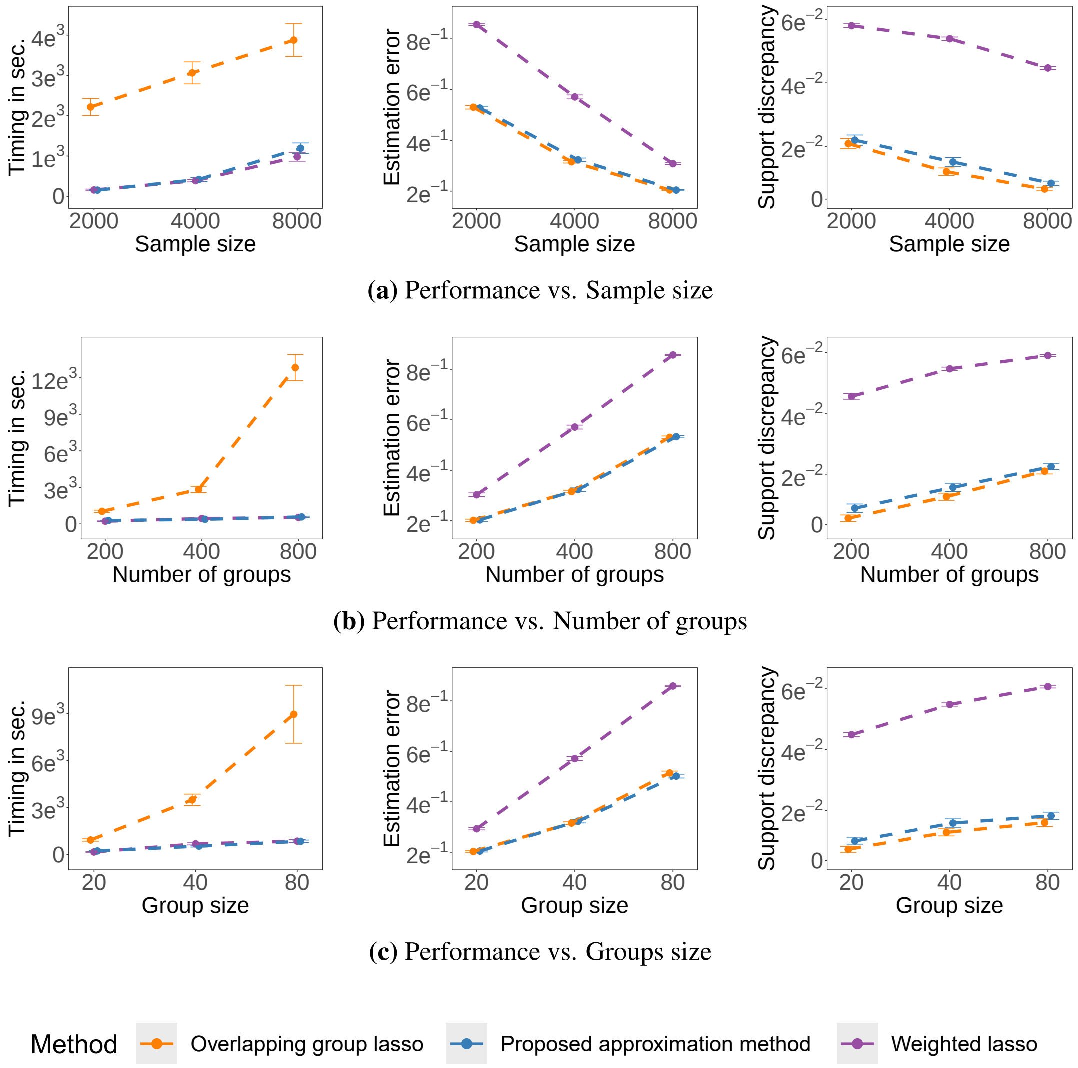
In the first set of experiments, we evaluate the performances under interlocking group structure (Figure 1a). This group structure exhibits a relatively low degree of overlap and is frequently used for evaluating overlapping group lasso methods (Yuan et al., 2011; Chen et al., 2012). Specifically, we set m interlocked groups with d variables in each group and $0.2d$ variables in each intersection. For example, $G_1 = \{1, \cdots, 10\}, G_2 = \{8, 9, \cdots, 17\}, \cdots, G_{10} = \{33, 34, \cdots, 42\}$ when $m = 5$ and $d = 10$. We will vary m and d to evaluate their impacts on the performance.

Following the strategy of Yan and Bien (2017), we generate the data matrix X from a Gaussian distribution $N(0, \Theta)$, where Θ is determined to match the correlations within the specified group structure. Initially, we construct a matrix $\tilde{\Theta}$ as follows:

$$\tilde{\Theta}_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } \beta_i \text{ and } \beta_j \text{ belong to different groups in } G, \\ 0.6, & \text{if } \beta_i \text{ and } \beta_j \text{ are in the same group in } \mathscr{G}, \\ 0.36, & \text{if } \beta_i \text{ and } \beta_j \text{ are in the same group in } G \text{ but different groups in } \mathscr{G}, \end{cases}$$

and then Θ is derived as the projection of $\tilde{\Theta}$ onto the set of symmetric positive definite matrices with a minimum eigenvalue of 0.1. Such strong within-group correlation patterns have also been used in Zhao et al. (2009a); Yang and Zou (2015).

We generate β^* by initially sampling its p coordinates from $N(10, 16)$, then randomly flipping signs of the covariates and randomly setting 90% of the groups to zero. This setup is consistent with the settings in Bach (2008); Friedman et al. (2010); Huang and Zhang (2010). The response variable Y is generated according to $Y = X\beta^* + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, and we set the signal-to-noise ratio to 3 following Yang and Zou (2015). The group weight in the overlapping group lasso problem is $w_g = \sqrt{d_g}$, as is usually used in practice. For all methods, we employ the absolute difference in function values between iterations as the stopping criterion, with a tolerance set at 10^{-5}.



(a) Performance vs. Sample size

(b) Performance vs. Number of groups

(c) Performance vs. Groups size

Figure 3: Regularization path computing time, ℓ_2 estimation error, and support discrepancy under different configurations of interlocking groups. (a) Varying n with fixed $m = 400$ and $d = 40$ (p = 12808); (b) Varying m with fixed $n = 4000$ and $d = 40$; (c) Varying d with fixed $n = 4000$ and $m = 400$.

Figure 3 presents the average computation times, estimation errors, and support discrepancy along with 95% confidence intervals (CIs). The results highlight the significant computational advantage of the proposed method over the original overlapping group lasso. Specifically, our method is 5–20 times faster than the original overlapping group lasso.

Even though the overlap within the interlocking group structure is not severe, solving the overlapping group lasso problem carries a more substantial computational burden due to the non-separable structure within its penalty term. Computational time escalates with larger sample sizes, an increased number of variables, and larger group sizes, emphasizing the substantial computational disadvantage of the overlapping group lasso as the problem scales up. In contrast, our proposed method consistently achieves accuracy similar to the overlapping group lasso estimator in both the estimation error and support discrepancy. This consistency in performance, observed across a spectrum of configurations, serves as an empirical confirmation of the validity of our theoretical findings.

On the other hand, the weighted lasso approximation is slightly faster than our method, which is expected from the optimization perspective. However, the weighted lasso approximation exhibits much higher errors compared to both the overlapping group lasso estimator and our estimator across all configurations. This reveals that the weighted gives a poor approximation to the overlapping group lasso. This shortcoming arises because the weighted lasso fails to leverage the group information, unlike the induced groups $\mathscr{G}$ used in our estimator.

In summary, our proposed estimator achieves comparable statistical performance to the original overlapping group lasso estimator while significantly enhancing computational efficiency. In contrast, despite its computational efficiency, the weighted lasso yields notably poor estimations, rendering it a noncompetitive alternative for approximating the original problems.

5.2 Nested Tree Structure of Overlapping Groups

In the second set of experiments, we evaluate the performance of the estimators under a configuration of the tree-group structures introduced in Jenatton et al. (2011b), described below.

Definition 1. *A set of groups $G = \{G_1, \cdots, G_m\}$ is said to be tree-structured in $[p]$ if $\cup_{g\in[m]} G_g = [p]$ and if for all $g, g' \in [m]$. $G_g \cap G_{g'} \neq \emptyset$ implies either $G_g \subset G_{g'}$ or $G_{g'} \subset G_g$.*

In particular, we consider the nested tree structure, a special case of tree-structured groups where all groups are nested. This configuration is interesting as it represents an extreme setting of overlapping groups – the overlapping degree is maximized in a certain sense, and we hope to evaluate the methods under this extreme scenario. The nested tree structure was also used in (Kim and Xing, 2012; Nowakowski et al., 2023). In this experiment, we use the SPAM solver, specifically designed for tree group structures, to provide a thorough evaluation across different implementations.

We consider the following nested group configuration: 800 groups $G = \{G_1, \ldots, G_{800}\}$ are established, where $G_g \subset G_{g+1}$ and $|G_g| = g \times 4$, for $g = 1, \cdots, 800$, with a total of $p = 3200$ variables. The sample size varies from 600 to 2400. The data matrix X is generated from $N(0, \Theta)$, where Θ is generated by first constructing the matrix $\tilde{\Theta}$ as:

$$\tilde{\Theta}_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0.6, & \text{if } \beta_i \text{ and } \beta_j \text{ belong to the same group in } \mathscr{G}, \\ 0.36, & \text{if } \beta_i \text{ and } \beta_j \text{ are in the same group in } G \text{ but in different groups in } \mathscr{G}, \end{cases} \quad .$$

and then projecting $\tilde{\Theta}$ onto the set of symmetric positive definite matrices with minimum eigenvalue 0.1. The generative process for β^* and y remains nearly identical as before, where the only difference is that the first 90% of the groups are set to zero following the hierarchical structure. The group weights are set to $w_g = 1/d_g$ as suggested in Nowakowski et al. (2023). For a fair comparison of the two solvers, in this experiment, we adopt the stopping criterion provided in the SPAM package (Mairal et al., 2014) with a convergence tolerance 10^{-5}.

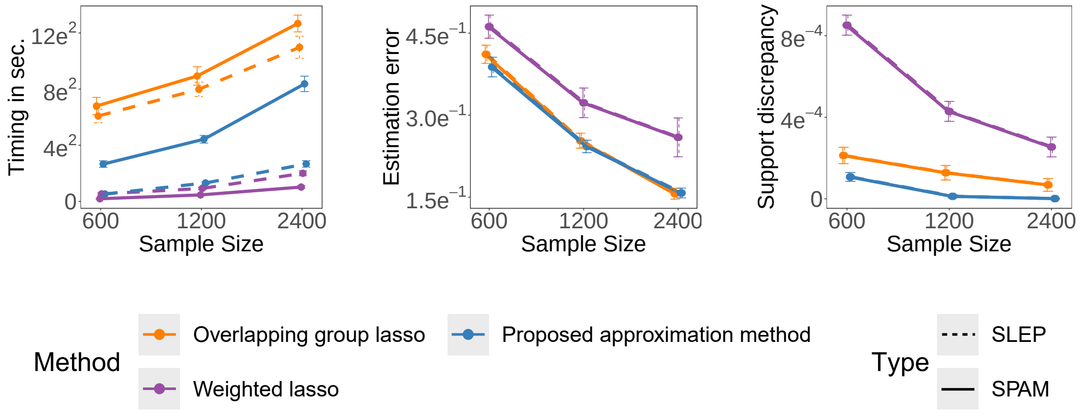


Figure 4: Regularization path computing time, ℓ_2 estimation error, and support discrepancy across various sample sizes under the nested tree group structure.

Figure 4 shows the performance of the three methods using both solvers. SLEP is generally faster than SPAM, but the two solvers give consistent conclusions about the estimators. As studied by Jenatton et al. (2011b), solving the overlapping group lasso problem becomes highly efficient under such a nested group structure because, under a tree structure, a single iteration over all groups is adequate to obtain the exact solution of the proximal operator. Our timing results support this statement. Compared with the previous setting, the timing advantage of our method is reduced. However, our method is still at least twice as fast as the overlapping group lasso. In terms of estimation error and support discrepancy, our proposed estimator consistently delivers similar results compared to the overlapping group lasso estimator. The comparison with the weighted lasso remains similar to the previous experiment; while the lasso estimator computes quickly, it continues to offer a very poor approximation.

In summary, solving overlapping group lasso problems exhibits efficiency when applied to tree structures. However, even in such cases, our proposed estimator maintains reasonable computational advantage and similar statistical estimation performance compared to the original overlapping group lasso estimator.

5.3 Group Structures Based on Real-world Gene Pathways

The previous two sets of experiments are based on human-designed group structures. To better reflect realistic situations, in this set of experiments, we use five gene pathway sets from the Molecular Signatures Database (Subramanian et al., 2005) as group structures, summarized in Table 1. Each gene pathway represents a collection of genes united by common biological characteristics. These pathways have been widely considered in studies of cancer and biological mechanisms (Menashe et al., 2010; Yuan et al., 2011; Livshits et al., 2015; Chen et al., 2020).

In particular, this data set can be used to assess the empirical applicability of Assumption 4 in our theory. The last column of Table 1 shows the ratio between $\max\{\mathscr{m}, \mathscr{d}_{\max}\}$ and $\max\{m, d_{\max}\}$. All values are within the range of [2,6], indicating that the two terms can be treated as terms in the same order.

Pathway databases	$\bar{d}/\text{sd}(d)$	$\bar{h}/\text{sd}(h)$	p	$\max\{\boldsymbol{m}, \boldsymbol{d}_{\max}\}/$ $\max\{m, d_{\max}\}$
BioCarta (Kong et al., 2006)	15.4/ 8.71	3.25/ 5.56	1129	2.35
PID (Schaefer et al., 2008)	38.51/ 19.59	3.28/ 5.09	2297	5.95
KEGG (Kanehisa et al., 2015)	58.48/ 47.36	2.58/ 3.39	4207	3.61
WIKI (Slenter et al., 2017)	38.17/ 44.10	4.35/ 7.70	6242	4.94
Reactome (Gillespie et al., 2021)	45.31/ 54.10	8.78/ 13.26	8331	2.35

Table 1: Summary information for the gene pathways: The mean and standard deviation of both the group size ($\bar{d}/\text{sd}(d)$), the overlapping degree ($\bar{h}/\text{sd}(h)$), the number of genes (p), and the ratio required in Assumption 4.

We use the gene expression data from Van De Vijver et al. (2002) as the covariate matrix X, which can be accessed through the R package `breastCancerNKI` (Schroeder et al., 2021). This design matrix has 295 observations and 24481 genes. We perform gene filtering for each gene pathway set to exclude genes not defined within any pathways, a data processing step commonly used in similar studies (Jacob et al., 2009; Chen et al., 2012; Lee and Xing, 2014). The data-generating procedure for β^* and y remains almost the same as before, except that we use a much sparser model because of the smaller sample size of the data. Specifically, we randomly sample $0.05m$ active groups and set the coefficients in other groups to zero. The weights in overlapping group lasso are set to $\sqrt{d_g}$.

Group Structure	Overlapping group lasso	Weighted lasso	The proposed approximation
BioCarts	67.18 [62.28, 72.08]	6.22 [5.99, 6.45]	16.03 [15.17, 16.89]
KEGG	287.27 [267.18, 307.36]	28.77 [26.42, 31.12]	48.32 [45.12, 51.52]
PID	445.99 [420.56, 471.42]	10.27 [9.74, 10.80]	31.25 [29.43, 33.07]
WIKI	1279.22 [1214.34, 1344.10]	63.56 [57.36, 69.76]	132.79 [121.82, 143.76]
Reactome	3739.97 [3569.27, 3910.67]	116.34 [106.32, 126.36]	194.61 [181.31, 207.91]

Table 2: Comparison of the average computing time (in seconds) and the corresponding 95% confidence intervals for each pathway group structure.

Group Structure	Overlapping group lasso	Lasso	Proposed approximation
BioCarts	0.22 [0.20, 0.24]	0.28 [0.24, 0.32]	0.25 [0.22, 0.28]
KEGG	0.52 [0.47, 0.57]	0.80 [0.76, 0.84]	0.54 [0.51, 0.57]
PID	0.23 [0.21, 0.25]	0.50 [0.44, 0.56]	0.25 [0.23, 0.28]
WIKI	0.55 [0.49, 0.61]	0.65 [0.58, 0.72]	0.55 [0.49, 0.61]
Reactome	0.66 [0.63, 0.69]	0.85 [0.83, 0.87]	0.65 [0.62, 0.68]

Table 3: Comparison of the relative ℓ_2 estimation errors and the corresponding 95% confidence intervals for each group structure.

Group Structure	Overlapping group lasso	Lasso	Proposed approximation
BioCarts	0.041 [0.039, 0.043]	0.043 [0.040, 0.046]	0.041 [0.039, 0.043]
KEGG	0.023 [0.021, 0.025]	0.026 [0.024, 0.028]	0.023 [0.021, 0.025]
PID	0.033 [0.031, 0.035]	0.033 [0.031, 0.035]	0.033 [0.031, 0.035]
WIKI	0.013 [0.012, 0.014]	0.013 [0.011, 0.015]	0.013 [0.012, 0.014]
Reactome	0.012 [0.011, 0.013]	0.020 [0.019, 0.021]	0.012 [0.010, 0.014]

Table 4: Comparison of the support discrepancy and the corresponding 95% confidence intervals for each group structure.

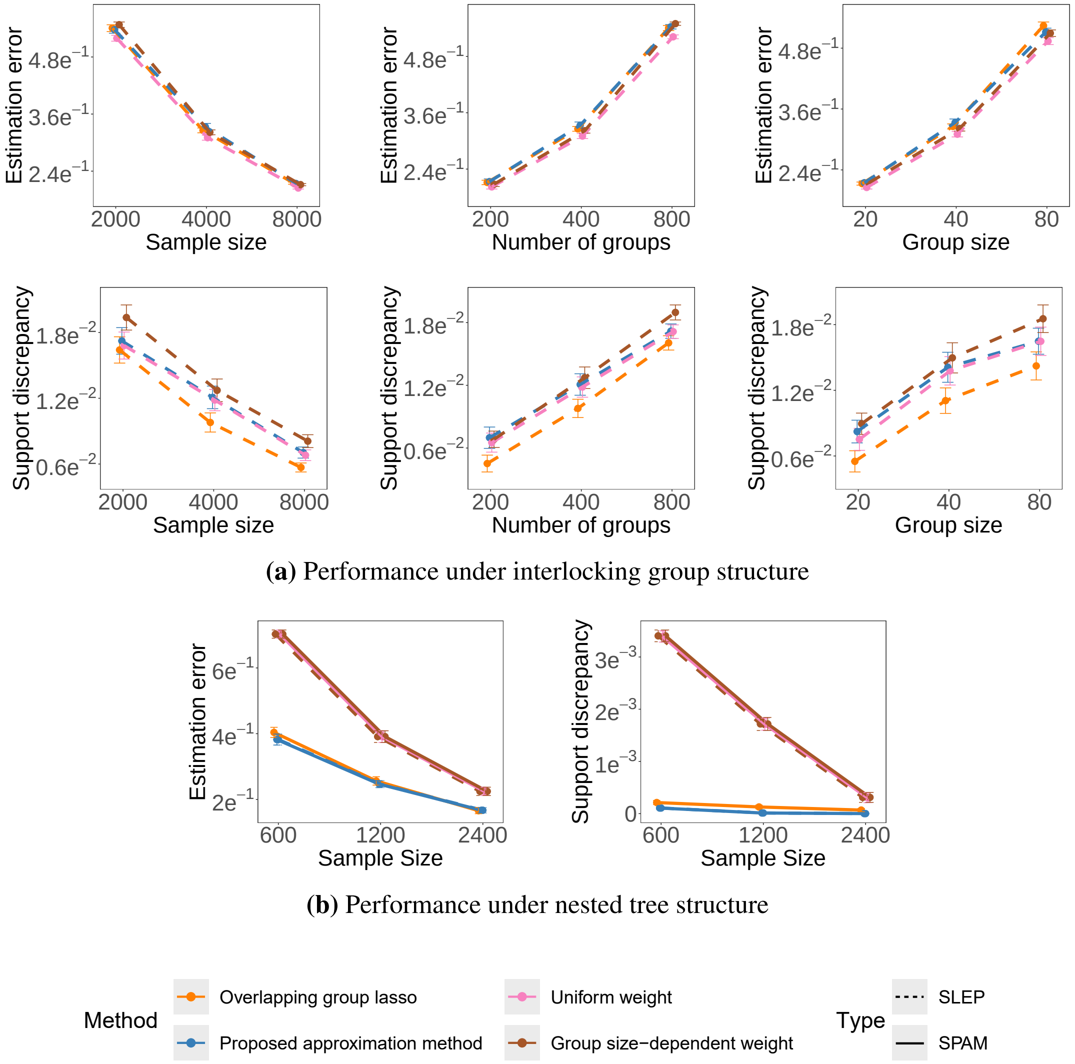
Table 2 displays the computing time, Table 3 displays the estimation error results, and Table 4 displays the support discrepancy results for the five pathway group structures. The high-level message remains consistent: Both our proposed group lasso approximation and the lasso approximation could substantially reduce the computing time. Across all settings, the proposed method reduces the computation time by 4 - 20 times and is more than 10 times faster in all settings with higher dimensions. Meanwhile, the proposed estimator delivers statistical performance similar to that of the original overlapping group lasso estimator. In contrast, the lasso approximation fails to leverage the group information effectively and yields inferior estimation results.

5.4 Comparison of the Proposed Weights against Other Weighting Choices

In addition to the partitioned groups, the overlapping-based weight defined in (6) for each partitioned group $\mathscr{g}$ is another crucial component to ensure the tightness of (7). We will demonstrate this aspect by experiments here to compare the proposed weights (6) with two other commonly used choices of weights that do not consider the original overlapping pattern: uniform weights and group size-dependent weights (Yuan and Lin, 2006), on the same induced groups $\mathscr{G}$. Specifically, uniform weighting is the setting when all groups share the same weight while the size-dependent weighting uses the weight $\sqrt{d_{\mathscr{g}}}$ if $w_g = \sqrt{d_g}$ (interlocking and gene pathway groups) and is $1/d_{\mathscr{g}}$ if $w_g = 1/d_g$ (nested groups). The comparative analysis is performed under all group structures in the previous simulations, maintaining consistent simulation settings.

Figure 5a and Figure 5b illustrate the weighting effects comparison in the settings of Figure 3 and Figure 4, respectively. Under the interlocking group structure (Figure 5a), three weighting schemes deliver similar performance in terms of estimation errors. Still, the size-dependent weighting leads to a larger support discrepancy. This interlocking group structure is not very distinctive for the three weights themes because the overlapping degree is nearly uniform. The nested group structures (Figure 5b) highlight the importance of the proposed weights more effectively. Our method significantly outperforms the other two weighting schemes and aligns well with the original overlapping group lasso estimator. The comparison of weighting designs on the gene pathway group structure is shown in Tables 5–6. The proposed estimator gives a close approximation to the original overlapping group lasso, but the other two weighting designs lead to significantly different performances in several settings.

In summary, the experiments demonstrate that the weights designed in our penalty also serve as an indispensable part of a successful approximation to the overlapping group lasso estimation, which is another aspect of the tightest separable relaxation property in Theorem 1.



(a) Performance under interlocking group structure

(b) Performance under nested tree structure

Figure 5: Regularization ℓ_2 estimation error and support discrepancy using different choices of weights. Figure 5a is an extension to Figure 3 under interlocking group structure, and Figure 5b is an extension to Figure 4 under nested tree structure.

6. Application Example: Pathway Analysis of Breast Cancer Data

In this section, we demonstrate the proposed method thorough a predictive tasks on the breast cancer tumor data, as previously used in Section 5.3. Unlike the previous simulation studies, here we use the complete data set with tumor labels for each observation. Specifically, each observation is labeled according to the status of the breast cancer tumors, with 79 classified as metastatic and 216 as non-metastatic. These labels serve as the response variable for our analysis.

Gene pathways have been widely considered to identify key gene groups in cancer studies. In particular, Yuan et al. (2011); Chen et al. (2012); Lee and Xing (2014) used overlapping group lasso techniques to exclude less significant biological pathways in cancer prediction. As a detailed

Group Structure	Proposed weight	Uniform weight	Group size-dependent weight
BioCarts	0.25 [0.22, 0.28]	0.28 [0.26, 0.30]*	0.35 [0.30, 0.40]*
KEGG	0.54 [0.51, 0.57]	0.80 [0.77, 0.83]*	0.58 [0.51, 0.65]*
PID	0.25 [0.23, 0.27]	0.24 [0.21, 0.27]	0.39 [0.36, 0.42]*
WIKI	0.55 [0.49, 0.61]	0.83 [0.80, 0.86]*	0.74 [0.67, 0.81]*
Reactome	0.65 [0.62, 0.68]	0.58 [0.55, 0.61]*	0.69 [0.63, 0.75]

Table 5: Comparative analysis of average estimation errors and the corresponding 95% confidence intervals for three weighting designs. The $*$ indicates that the error is statistically different from that of overlapping group lasso by a paired t-test.

Group Structure	Proposed weight	Uniform weight	Group size-dependent weight
BioCarts	0.041 [0.039, 0.043]	0.045 [0.042, 0.048]*	0.042 [0.039, 0.045]
KEGG	0.023 [0.021, 0.025]	0.059 [0.055, 0.063]*	0.024 [0.022, 0.026]
PID	0.033 [0.031, 0.035]	0.037 [0.035, 0.039]*	0.030 [0.027, 0.033]*
WIKI	0.013 [0.012, 0.014]	0.025 [0.023, 0.027]*	0.013 [0.012, 0.014]
Reactome	0.012 [0.010, 0.014]	0.010 [0.008, 0.012]*	0.022 [0.021, 0.023]*

Table 6: Comparative analysis of average support discrepancy and the corresponding 95% confidence intervals for three weighting designs. The $*$ indicates that the value is statistically different from that of overlapping group lasso by a paired t-test.

example, Chen et al. (2012) leveraged the overlapping group lasso penalty to pinpoint biologically meaningful gene groups. Their analysis identified several groups of genes associated with essential biological functions, such as protease activity, protease inhibitors, and nicotine and nicotinamide metabolism, which turned out to be important breast cancer markers (Ma and Kosorok, 2010). This evidence highlights the potential of using the overlapping group lasso penalty in cancer analysis. On the other hand, another way to incorporate gene pathway information in such analysis is to retain genes by entire pathways. Jacob et al. (2009) used the latent overlapping group lasso penalty to achieve this while Mairal and Yu (2013) introduced an ℓ_∞ variant further. The success of all these previous studies reveals the potential of the gene pathway information in cancer prediction. They also show that the proper way to use the pathways (e.g., either eliminating-by-group, as in overlapping group lasso, or including-by-group, as in latent overlapping group lasso) highly depends on the dataset and genes.

In our analysis, we use regularized logistic regression to build a classifier with several penalties: the overlapping group lasso penalty (OGL), our proposed group lasso approximation penalty (Proposed approximation), the standard lasso penalty (Tibshirani, 1996), the latent overlapping group lasso penalty (LOG) (Jacob et al., 2009), and the ℓ_∞ latent overlapping group lasso penalty (Mairal and Yu, 2013). As mentioned in previous sections, our focus is not on justifying the overlapping group lasso should be used. Instead, **our primary objective is to demonstrate that when an overlapping group lasso penalty is used, our method provides a good approximation to the overlapping group lasso (with a much faster computation) across various pathway sets** (Table 1), regardless whether or not the overlapping group lasso penalty is the best option for the problem.

Two additional aspects can also be evaluated as by-products from our analysis. First, as the lasso penalty does not consider the pathway information, comparing the performance of the group-based penalty and the lasso penalty in this problem would verify whether a specific gene pathway set contains predictive grouping information for breast cancer tumor type. Second, by assessing the predictive performances among the overlapping group lasso classifier and the latent overlapping group lasso classifiers, we can verify whether a specific gene pathway set is more suitable for eliminating-by-group or including-by-group strategies for prediction.

Method Database	OGL	Lasso	Proposed approximation
BioCarts	732	26	75
KEGG	2468	102	225
PID	1231	41	107
WIKI	5172	170	395
Reactome	11356	321	1186

Table 7: Computing time (in seconds) under different pathway databases.

Method Database	OGL	Lasso	Proposed approximation	LOG	LOG ∞
BioCarts	0.7103	0.6989	0.7242	0.6888	0.6995
KEGG	0.7021	0.6862	0.7081	0.7390	0.7333
PID	0.7475	0.7004	0.7301	0.6881	0.6891
WIKI	0.6862	0.7282	0.6893	0.7149	0.7207
Reactome	0.6921	0.7301	0.7053	0.7463	0.7438

Table 8: Predictive AUC results of the three methods under different pathway databases.

We adopt the evaluation procedure of Lee and Xing (2014), where we randomly split the data set into 200 training observations and 95 test observations. All methods are tuned by 5-fold cross-validation on the training data. We calculate the area under the receiver operating characteristic (AUC) curve, a commonly used metric for classifying accuracy (Hanley and McNeil, 1982), on the test data. The total time for the entire cross-validation process is recorded as the computation time. The experiment is repeated 100 times independently. Table 7 and Table 8 show the average computing time and AUC, respectively. The results can be summarized as follows:

- First and foremost, the proposed estimator acts as an effective and computationally efficient approximation for the overlapping group lasso estimator. The results evidently support this claim. The proposed estimator delivers predictive performance that is (the most) similar to the overlapping group lasso estimator across various pathway datasets while significantly reducing the computing time by roughly ten times.

- Second, the lasso classifier performs best only on the WIKI pathway set, suggesting that the pathways in the WIKI database might not be sufficiently informative for cancer prediction.

- Third, the superiority of either the overlapping group lasso regularizations or the latent overlapping group lasso regularizations depends on the specific group information. Among the four pathway sets with useful group information, the overlapping group lasso delivers superior predictive performance for the Biocarts and PID databases, while the latent overlapping group lasso classifiers provide better predictions on the KEGG and Reactome databases.

As a remark, while our evaluation is based on prediction accuracy, it is not the only criterion to determine if a method is proper for the dataset. For example, Mairal and Yu (2013) found that neither the overlapping group lasso model nor the latent overlapping group lasso model outperformed simple ridge regularization in prediction. The value of structured penalties also lies in their ability to identify potentially more interpretable genes, depending on the biological interpretations.

7. Discussion

We have introduced a separable penalty as an approximation to the group lasso penalty when groups overlap. The penalty is designed by partitioning the original overlapping groups into disjoint subgroups and reweighing the new groups according to the original overlapping pattern. The penalty is the tightest separable relaxation of the overlapping group lasso among all ℓ_{q_1}/ℓ_{q_2} norms. We have also shown that for linear problems, the proposed estimator is statistically equivalent to the original overlapping group lasso estimator but enjoys significantly faster computation for large-scale problems.

Several interesting directions could be considered for future research. The overlapping group lasso penalty presents a variable selection by eliminating variables by entire groups. A counterpart selection procedure can include variables by entire groups, which is achieved by the latent overlapping group lasso (Jacob et al., 2009). This penalty also suffers from a non-separability computational bottleneck. It would be valuable to investigate whether a similar approximation strategy could be designed to boost the computational performance in this scenario. More generally, the introduced concept of "tightest separable relaxation" might be a promising direction for optimizing non-separable functions. Studying the more general form and corresponding properties of this concept may generate fundamental insights about optimization.

Acknowledgments

The work is partly supported by the NSF grant DMS-2015298 and the 3-Caveliers award from the University of Virginia. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota and the Research Computing at the University of Virginia for providing resources that contributed to the research results reported within this paper. We appreciate the insightful feedback and comments from the editor and reviewer, which significantly improved the paper.

References

E. Austin, W. Pan, and X. Shen. A new semiparametric approach to finite mixture of regressions using penalized regression via fusion. Statistica Sinica, 30(2):783, 2020.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research, 9(6), 2008.

S. Basu, A. Shojaie, and G. Michailidis. Network granger causality with inherent grouping structure. The Journal of Machine Learning Research, 16(1):417–453, 2015.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1):183–202, 2009.

D. P. Bertsekas. Nonlinear programming. Journal of the Operational Research Society, 48(3): 334–334, 1997.

S. Boyd, N. Parikh, E. Chu, B.Peleato, and J.Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine learning, 3:1–122, 2011.

T. T. Cai, A. R. Zhang, and Y. Zhou. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. IEEE Transactions on Information Theory, 2022.

F. Campbell and G. I. Allen. Within group variable selection through the exclusive lasso. Electronic Journal of Statistics, 11(2):4220–4257, 2017.

J. Chen, C. Liu, J. Cen, T. Liang, J. Xue, H. Zeng, Z. Zhang, G. Xu, C. Yu, Z. Lu, et al. Kegg-expressed genes and pathways in triple negative breast cancer: Protocol for a systematic review and data mining. Medicine, 99(18), 2020.

X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing. Smoothing proximal gradient method for general structured sparse regression. The Annals of Applied Statistics, 6(2):719–752, 2012.

J. Cheng, T. Li, E. Levina, and J. Zhu. High-dimensional mixed graphical models. Journal of Computational and Graphical Statistics, 26, 2017.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(2):373–397, 2014.

A. Dedieu. An error bound for lasso and group lasso in high dimensions. arXiv:1912.11398, 2019.

W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. Proceedings of the SPIE, 2013.

J. H. Friedman, T. J. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. arXiv: Statistics Theory, 2010.

M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Ragueneau, Y. Haider, B. May, V. Shamovsky, J. Weiser, T. Brunson, N. Sanati, L. Beckman, X. Shao, A. Fabregat, K. Sidiropoulos, J. Murillo, G. Viteri, J. Cook, S. Shorser, G. Bader, E. Demir, C. Sander, R. Haw, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase 2022. Nucleic Acids Research, 50(D1):D687–D692, 11 2021.

R. L. Graham, D. E. Knuth, and O. Patashnik. Concrete Mathematics: A Foundation for Computer Science. Addison-Wesley, Reading, MA, second edition, 1994. ISBN 0201558025 9780201558029 0201580438 9780201580433 0201142368 9780201142365.

E. Grave, G. R. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. Advances in Neural Information Processing Systems, 24, 2011.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology, 143(1):29–36, 1982.

J. Huang and T. Zhang. The benefit of group sparsity. The Annals of Statistics, 38(4):1978–2004, 2010. ISSN 00905364, 21688966. URL `http://www.jstor.org/stable/20744481`.

L. Jacob, G. Obozinski, and J. Vert. Group lasso with overlap and graph lasso. Proceedings of the 26th Annual International Conference on Machine Learning,ICML, 09:433–440, 2009.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. The Journal of Machine Learning Research, 12:2777–2824, 2011a.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. The Journal of Machine Learning Research, 12:2297–2334, 2011b.

M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research, 44(D1):D457–D462, 10 2015.

S. Kim and E. P. Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. The Annals of Applied Statistics, 6:1095–1117, 2012.

S. W. Kong, W. T. Pu, and P. J. Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics, 22(19):2373–2380, 2006.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. Annals of Statistics, pages 1302–1338, 2000.

S. Lee and E. Xing. Screening rules for overlapping group lasso. arXiv:1410.6880, 2014.

S. Lee and E. P. Xing. Leveraging input and output structures for joint mapping of epistatic and marginal eqtls. Bioinformatics, 28(12):i137–i146, 2012.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. The Annals of Applied Statistics, 2(1):245–263, 2008.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research, 10(10), 2009a.

J. Liu, S. Ji, and J. Ye. SLEP: Sparse Learning with Efficient Projections. Arizona State University, 2009b. URL `http://www.public.asu.edu/~jye02/Software/SLEP`.

A. Livshits, A. Git, G. Fuks, C. Caldas, and E. Domany. Pathway-based personalized analysis of breast cancer expression data. Molecular oncology, 9(7):1471–1483, 2015.

P.-L. Loh. High-dimensional statistics with systematically corrupted data. University of California, Berkeley, 2014.

K. Lounici, M. Pontil, S. V. D. Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. The Annals of Statistics, 39(4):2164–2204, 2011.

S. Ma and M. R. Kosorok. Detection of gene pathways with predictive power for breast cancer prognosis. BMC bioinformatics, 11(1):1–11, 2010.

J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. Journal of Machine Learning Research, 14(8), 2013.

J. Mairal, R. Jenatton, F. Bach, and G. R. Obozinski. Network flow algorithms for structured sparsity. Advances in Neural Information Processing Systems, 23, 2010.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, R. Jenatton, and G. Obozinski. Spams: A sparse modeling software, v2. 3. URL http://spams-devel. gforge. inria. fr/downloads. html, 2014.

L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(1):53–71, 2008.

I. Menashe, D. Maeder, M. Garcia-Closas, J. D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D. J. Hunter, S. J. Chanock, P. S. Rosenberg, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer research, 70(11):4453–4459, 2010.

K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee. Node-based learning of multiple gaussian graphical models. The Journal of Machine Learning Research, 15(1):445–488, 2014.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statistical science, 27 (4):538–557, 2012.

Y. Nesterov. Gradient methods for minimizing composite functions. Mathematical programming, 140(1):125–161, 2013.

S. Nowakowski, P. Pokarowski, W. Rejchel, and A. Sołtys. Improving group lasso for high-dimensional categorical data. In International Conference on Computational Science, pages 455–470. Springer, 2023.

Z. Qin, K.Scheinberg, and D. Goldfarb. Efficient block-coordinate descent algorithms for the grouplasso. Mathematical Programming Computation, 5(2), 2013.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(5):1009–1030, 2009.

C. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. Buetow. Pid: The pathway interaction database. Nature Precedings, 3, 08 2008. doi: 10.1038/npre.2008.2243.1.

M. Schroeder, B. Haibe-Kains, A. Culhane, C. Sotiriou, G. Bontempi, and J. Quackenbush. breastCancerNKI: Genexpression dataset published by van't Veer et al. [2002] and van de Vijver et al. [2002] (NKI)., 2021. URL `http://compbio.dfci.harvard.edu/`. R package version 1.32.0.

D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. Digles, F. Ehrhart, P. Giesbertz, M. Kalafati, M. Martens, R. Miller, K. Nishida, L. Rieswijk, A. Waagmeester, L. M. T. Eijssen, C. T. Evelo, A. R. Pico, and E. L. Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research, 46(D1):D661–D667, 11 2017. ISSN 0305-1048.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, 2005.

A. Tank, E. B. Fox, and A. Shojaie. An efficient admm algorithm for structural break detection in multivariate time series. arXiv preprint arXiv:1711.08392, 2017.

D. A. Tarzanagh and G. Michailidis. Estimation of graphical models through structured norm minimization. Journal of machine learning research, 18(1), 2018.

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

M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine, 347(25):1999–2009, 2002.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 constrained quadratic programming (lasso). IEEE transactions on information theory, 55(5): 2183–2202, 2009.

M. J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

S. Xiang, X. Shen, and J. Ye. Efficient nonconvex sparse group feature selection via continuous and discrete optimization. Artificial Intelligence, 224:28–50, 2015. ISSN 0004-3702.

X. Yan and J. Bien. Hierarchical sparse modeling: A choice of two group lasso formulations. Statistical Science, 32(4):531–560, 2017.

C. Yang, X. Wan, Q. Yang, H. Xue, and W. Yu. Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. BMC bioinformatics, 11(1):1–11, 2010.

J. Yang and J. Peng. Estimating time-varying graphical models. Journal of Computational and Graphical Statistics, 29(1):191–202, 2020.

Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. Statistics and Computing, 25(6):1129–1141, 2015.

G. Yu and J. Bien. Learning local dependence in ordered data. The Journal of Machine Learning Research, 18(1):1354–1413, 2017.

L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. Advances in Neural Information Process Systems, pages 352–360, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. Annals of Statistics, 37(6A):3468–3497, 2009a.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. Annals of Statistics, 37(6A):3468–3497, 2009b.

Appendix A. Notation Summary

Indices:	
$[z]$	index set $\{1, ..., z\}$
G_g	index set of g^{th} group
G_S	collection of non-zero groups, $\bigcup_{g \in S(\beta)} G_g$
$G_{\overline{S}}$	$\bigcup_{g \in \overline{S(\beta)}} G_g$
β_j	the j^{th} element of β
β_{G_g}	sub-vector of β indexed by G_g
$\beta_{M(S)}$	projection of β onto $M(S)$
$A_{,T}$	sub-matrix consisting of the columns indexed by T
Parameters:	
H	a diagonal matrix, $diag(\frac{1}{h_1}, \cdots, \frac{1}{h_p})$
$\mathbf{G}$	group structure matrix, $\mathbf{G}_{gj} = 1$ iff $\beta_j \in G_g$
d_g	group size, $d_g = \sum_{j \in [p]} \mathbf{G}_{gj}$
$d_{\max}$	maximum group size, $d_{\max} = \max_{g \in [m]} d_g$
h_j	overlap degree, $h_j = \sum_{g \in [m]} \mathbf{G}_{gj}$
$h^g_{\max}$	maximum overlap degree in G_g, $h^g_{\max} = \max_{j \in G_g} h_j$
$h^g_{\min}$	minimum overlap degree in G_g, $\min_{j \in G_g} h_j$
$h_{\max}$	maximum overlap degree, $h_{\max} = \max_{j \in [p]} h_j$
$\hbar_{\mathscr{g}}$	overlap degree of $\mathscr{G}_{\mathscr{g}}$, $h_{\{j \mid j \in \mathscr{G}_{\mathscr{g}}\}}$
σ	parameter in the sub-Gaussian distribution
s_g	number of non-zero groups $\lvert S \rvert$
$\overline{s_g}$	number of groups in the argument group support set $\lvert \overline{S} \rvert$
κ	parameter controls convexity
Definitions:	
$\phi(\beta)$	group lasso norm, $\sum\limits_{g \in [m]} w_g \left\lVert \beta_{G_g} \right\rVert_2$,
$\phi^*(\beta)$	dual norm of $\phi(\beta)$, $\max\limits_{g \in [m]} \frac{1}{w_g} \left\lVert (H\beta)_{G_g} \right\rVert_2$
$F(\mathscr{g}) \subseteq [m]$	overlapping groups which include the variables in $\mathscr{G}_{\mathscr{g}}$
$F^{-1}(g) \subseteq [\mathscr{m}]$	non-overlapping groups that were partitioned from G_g
$\lVert\beta_{\{G,w\}}\rVert_{q_1,q_2}$	ℓ_{q_1,q_2} norm, $\left\{ \sum\limits_{g \in [m]} w_g \left(\sum\limits_{j \in G_g} \lvert\beta_j\rvert^{q_2} \right)^{\frac{q_1}{q_2}} \right\}^{\frac{1}{q_1}}$
$supp(\beta)$	support set, $\{j \in \{1, \cdots, p\} \mid \beta_j \neq 0\}$
$S(\beta)$	group support set, $\{g \in \{1, \cdots, m\} \mid G_g \cap supp(\beta) \neq \varnothing\}$
$\overline{S(\beta)}$	$\{g = \{1, \cdots, m\} \mid G_g \cap G_{S(\beta)} \neq \varnothing\}$
$M(S)$	$\{\beta \in \mathbb{R}^p \mid \beta_j = 0$ for all $j \in (G_S)^c\}$
$M^{\perp}(S)$	$\{\beta \in \mathbb{R}^p \mid \beta_j = 0$ for all $j \in G_S\}$
$\Omega(G, s_g)$	$\left\{\beta : \sum\limits_{G_g \in G} \mathbb{1}_{\{\lVert\beta_{G_g}\rVert_2 \neq 0\}} \leqslant s_g\right\}$
$J_G(\beta)$	$[p] \backslash \{ \bigcup_{G_g \cap supp(\beta) = \emptyset} G_g \}$.
$\mathsf{G}_{J_G(\beta)}$	$\{g \in [m] \mid G_g \cap J_G(\beta) \neq \emptyset\}$
$\mathsf{G}_{J_G(\beta)^c}$	$\{g \in [m] \mid G_g \cap J_G(\beta)^c \neq \emptyset\}$

Table 9: Mathematical notations in the paper.

Appendix B. Uniqueness of the Overlapping Group Lasso Problem

The group lasso penalization problems (14) and (15) are generally convex, but may not be strictly convex. The uniqueness of these problems has been studied by Jenatton et al. (2011a). Here we introduce their results for completeness. Note that our theoretical properties in Section 3 do not rely on such uniqueness.

Lemma 6. *(see Jenatton et al., 2011a, Proposition 1) If the gram matrix $Q = X^\top X/n$ is invertible, or if there exists $g \in [m]$ such that $G_g = [p]$, then the optimization problem specified in* (14)*, with $\lambda_n > 0$, is guaranteed to have a unique solution. The same property holds for problem* (15) *with G replaced by $\mathscr{G}$.*

Appendix C. Additional Theoretical Results

To begin with, we introduce our proposed upper bound for the dual norm of the overlapping group lasso penalty.

Proposition 1. *Recall that $\phi^G(\beta)$ is the overlapping group lasso penalty defined in* (1)*. Let ϕ^* be the dual norm of $\phi^G(\beta)$. Then the sharp upper bound for ϕ^* is:*

$$\max_{g\in[m]} \frac{1}{w_g} \left\|(H\beta)_{G_g}\right\|_2,$$

where H is a diagonal matrix with diagonals $(\frac{1}{h_1}, \cdots, \frac{1}{h_p})$.

Assumption 7. *Under model* (13)*, we assume*

1. *(Sub-Gaussian noises) The coordinates of ε are i.i.d zero mean sub-Gaussian random variable denote with parameter σ, which means that there exist $\sigma > 0$ such that*

$$E[e^{t\varepsilon})] \leqslant \frac{e^{\sigma^2 t^2}}{2}, \ \textit{for all} \ \ t \in \mathbb{R}.$$

2. *(Group normalization condition) $\sqrt{\gamma_{\max}(X_{G_g}^\top X_{G_g}/n)} \leqslant c$ for some constant c.*

3. *(Restricted strong convexity condition) For some $\kappa > 0$,*

$$\frac{\left\|X\left(\bar{\beta}-\beta^*\right)\right\|_2^2}{n} \geqslant \kappa \left\|\bar{\beta}-\beta^*\right\|_2^2, \ \textit{for all} \ \bar{\beta} \in \left\{\beta \mid \phi\big(\left(\beta-\beta^*\right)_{M^\perp(\bar{S})}\big) \leqslant 3\phi\big(\left(\beta-\beta^*\right)_{M(\bar{S})}\big)\right\}.$$

Remark: The assumption requires an upper bound for the quadratic form associated with each group. This type of assumption is commonly used for developing the upper estimation error bound for non-overlapping group lasso (Huang and Zhang, 2010; Lounici et al., 2011; Negahban et al., 2012; Dedieu, 2019; Wainwright, 2019). Additionally, the restricted curvature conditions have been well discussed by Wainwright (2019). The curvature κ in Assumption 7 is a parameter measuring the convexity. Generally speaking, the restricted curvature conditions state the loss function is locally strongly convex in a neighborhood of ground truth and thus guarantees that a small distance between the estimate and the true parameter implies the closeness in the loss function. However, such a strong convexity condition cannot hold in the high-dimensional setting. So, we focus on a restrictive set of estimates. Restricted curvature conditions are milder than the group-based RIP conditions used in (Huang and Zhang, 2010; Dedieu, 2019), which require that all submatrices up to a certain size are close to isometries (Wainwright, 2019). Based on Assumption 7, Theorem 7 gives ℓ_2 norm estimation upper error bound for overlapping group lasso.

Theorem 7. *Define* $h^g_{\min} = \min_{j \in G_g} h_j$, $d_{\max} = \max_{g \in [m]} d_g$, *and* $\mathscr{d}_{\max} = \max_{\mathscr{g} \in [\mathscr{m}]} \mathscr{d}_g$. *Suppose that Assumption 7 holds. Then for any* $\delta \in [0, 1]$,

1. *with* $\lambda_n = \frac{8c\sigma}{\min_{g\in[m]}\left(w_g^2 h^g_{\min}\right)}\sqrt{\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta}$, *the following bound holds for* $\hat{\beta}^G$ *in* (14)

$$\left\|\hat{\beta}^G - \beta^*\right\|_2^2 \lesssim \frac{\sigma^2}{\kappa^2} \cdot \frac{\left(\sum_{g\in\overline{S}} w_g^2\right) \cdot h_{\max}^{G_{\overline{S}}}}{\min_{g\in[m]}\left(w_g^2 h^g_{\min}\right)} \cdot \left(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\right). \tag{29}$$

   with probability at least $1 - e^{-2n\delta}$.

2. *with* $\lambda_n = \frac{8c\sigma}{\min_{g\in[m]} \mathscr{w}_g}\sqrt{\frac{\mathscr{d}_{\max}\log 5}{n} + \frac{\log m}{n} + \delta}$, *the following bound holds for* $\hat{\beta}^{\mathscr{G}}$ *in* (15)

$$\left\|\hat{\beta}^{\mathscr{G}} - \beta^*\right\|_2^2 \lesssim \frac{\sigma^2}{\kappa^2} \cdot \frac{\sum_{\mathscr{g}\in F^{-1}(S)} \mathscr{w}_{\mathscr{g}}^2}{\min_{\mathscr{g}\in[\mathscr{m}]}\left(\mathscr{w}_{\mathscr{g}}^2\right)} \cdot \left(\frac{\mathscr{d}_{\max}\log 5}{n} + \frac{\log \mathscr{m}}{n} + \delta\right). \tag{30}$$

Following the framework in Negahban et al. (2012); Wainwright (2019), we further study the applicability of the restricted curvature conditions in terms of a random design matrix. Given a group structure G, Theorem 7 is developed based on the assumption that the fixed design matrix X satisfies the restricted curvature condition. In practice, verifying that a given design matrix X satisfies this condition is difficult. Indeed, developing methods to "certify" design matrices this way is one line of ongoing research (Wainwright, 2019). However, it is possible to give high-probability results based on the following assumptions.

Theorem 8. *Under Assumptions 1,2, and 3, we have*

1. *with probability at least* $1 - e^{-c'n}$, $\max_{g\in[m]} \sqrt{\gamma_{\max}(X_{G_g}^\top X_{G_g}/n)} \leqslant c$ *for some constants* $c, c' > 0$, *as long as* $\log m = o(n)$.

2. *the restricted strong convexity condition, which is*

$$\frac{\left\|X\left(\bar{\beta} - \beta^*\right)\right\|_2^2}{n} \geqslant \kappa \left\|\bar{\beta} - \beta^*\right\|_2^2, \text{ for all } \bar{\beta} \in \left\{\beta \mid \phi\big(\left(\beta - \beta^*\right)_{M^\perp(\overline{S})}\big) \leqslant 3\phi\big(\left(\beta - \beta^*\right)_{M(\overline{S})}\big)\right\}.$$

   hold with probability at least $1 - \frac{e^{-\frac{n}{32}}}{1 - e^{-\frac{n}{64}}}$ *for some constant* $\kappa > 0$.

Appendix D. Proofs

D.1 Proof of Theorem 1

Lemma 9. *For any norm* $||\cdot_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ *satisfying the conditions in* (11)*, the following two statements hold:*

1. *For any* $\mathscr{g} \in [m]$*, there exists a* $\tilde{g} \in [|\tilde{G}|]$ *such that* $\mathscr{G}_{\mathscr{g}} \subseteq \tilde{G}_{\tilde{g}}$*.*
2. *For any* $\tilde{g} \in [|\tilde{G}|]$*, there exists a* $\mathscr{g} \in [m]$ *such that* $\tilde{G}_{\tilde{g}} = \mathscr{G}_{\mathscr{g}}$*.*

Proof Based on Lemma 9, if a norm $||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ satisfies (11), then it must be that $\tilde{G} = \mathscr{G}$. Consequently, any disparity between $||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ and our proposed norm could only be due to differences in weights or the values of q_1 or q_2. Consequently, for any β with non-zero elements solely in the $\mathscr{g}$th group $\mathscr{G}_{\mathscr{g}}$, we have:

$$\sum_{g\in[m]} w_g||\beta_{G_g}||_2 = \sum_{g\in[m]} \Big(\sum_{g\in F(\mathscr{g})} w_g\Big)||\mathscr{G}_{\mathscr{g}}||_2 \leqslant ||\beta_{\{\mathscr{G},\tilde{w}\}}||_{q_1,q_2} \leqslant \sum_{\mathscr{g}\in[m]} \mathscr{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_2, \tag{31}$$

which further implies that

$$(\tilde{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_{q_2}^{q_1})^{\frac{1}{q_1}} = \mathscr{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_2.$$

By setting one element in $\mathscr{G}_{\mathscr{g}}$ to 1, and other elements to 0, it follows that $\tilde{w}_{\mathscr{g}} = \mathscr{w}_{\mathscr{g}}$. Since this holds for any group in $\mathscr{G}$, we have $\tilde{w} = \mathscr{w}$.

From (31), it is evident that $(\mathscr{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_{q_2}^{q_1})^{\frac{1}{q_1}} = \mathscr{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_2$ for any β with non-zero elements only in $\mathscr{G}_{\mathscr{g}}$. This suggests that $q_1 = 1$ and $q_2 = 2$. Therefore, the existing norm $||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2}$ does not satisfy the second condition in (11). ■

D.1.1 PROOF OF LEMMA 9

Proof We begin by proving the first item. Recall that $\mathbb{G}$ represents the space of all possible partitions of $[p]$. Given that $\tilde{G} \in \mathbb{G}$, for an arbitrary $\mathscr{g} \in [m]$, suppose $\mathscr{G}_{\mathscr{g}} \nsubseteq \tilde{G}_{\tilde{g}}$ for any $\tilde{g}$. Then, we can identify the smallest set T such that:

$$\mathscr{G}_{\mathscr{g}} \subseteq \bigcup_{\tilde{g}\in T} \tilde{G}_{\tilde{g}}.$$

Let $T = \{t_1, t_2, \cdots, t_{|T|}\}$. Select one element $\beta_j \in \mathscr{G}_{\mathscr{g}} \cap \tilde{G}_{t_1}$ and another $\beta_k \in \mathscr{G}_{\mathscr{g}} \cap \tilde{G}_{t_2}$. Since both β_j and β_k belong to $\mathscr{G}_{\mathscr{g}}$, if an original group includes β_j, it also contains β_k. Let β be a vector where only β_j and β_k are non-zero, then we have:

$$\begin{aligned}\sum_{g\in[m]} w_g||\beta_{G_g}||_2 = (\sum_{\{g|\beta_j\in G_g\}} w_g)\sqrt{\beta_j^2+\beta_k^2} &\leqslant ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} \\ &\leqslant \sum_{\mathscr{g}\in[m]} \mathscr{w}_{\mathscr{g}}||\beta_{\mathscr{G}_{\mathscr{g}}}||_2 = (\sum_{\{g|\beta_j\in G_g\}} w_g)\sqrt{\beta_j^2+\beta_k^2},\end{aligned}$$

which further leads to

$$||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} = \Big((\tilde{w}_{t_1}|\beta_j|)^{q_1}+(\tilde{w}_{t_2}|\beta_k|)^{q_1}\Big)^{\frac{1}{q_1}} = \tilde{w}_{t_1}^{\frac{1}{q_1}}|\beta_j|+\tilde{w}_{t_2}^{\frac{1}{q_1}}|\beta_k| = \Big(\sum_{\{g|\beta_j\in G_g\}} w_g\Big)\sqrt{\beta_j^2+\beta_k^2},$$

for any $0 \leqslant q_1, q_2 \leqslant \infty$. However, by setting

$$\begin{cases}\beta_j = \beta_k = 1, \beta_{\{[p]\backslash\{j,k\}\}} = 0 & \text{if } w_{t_1}^{\frac{1}{q_1}} + w_{t_2}^{\frac{1}{q_1}} \neq \sqrt{2}\left(\sum_{\{g|\beta_j \in G_g\}} w_g\right) \\ \beta_j = 2, \beta_k = 1, \beta_{\{[p]\backslash\{j,k\}\}} = 0 & \text{if } w_{t_1}^{\frac{1}{q_1}} + w_{t_2}^{\frac{1}{q_1}} = \sqrt{2}\left(\sum_{\{g|\beta_j \in G_g\}} w_g\right)\end{cases},$$

we arrive at a contradiction. Thus, we demonstrate that if a norm $|| \cdot_{\{\tilde{G},w\}} ||_{q_1,q_2}$ exists, then each group in $\tilde{G}$ is a union of groups in $\mathscr{G}$.

Now, we continue to prove the second item. Given that the first part establishes each group in $\tilde{G}$ is a union of groups in $\mathscr{G}$, consider a specific group $\tilde{g} \in [|\tilde{G}|]$. Assume there is an index set $V \subseteq [\mathscr{m}]$ such that $\tilde{G}_{\tilde{g}} = \bigcup_{\mathscr{g}\in V} \mathscr{G}_{\mathscr{g}}$ with $|V| > 1$. Denote $V = \{\mathscr{v}_1, \cdots, \mathscr{v}_{|V|}\}$. We analyze two scenarios:

- **Case I:** $\nexists a \in [m]$ s.t. $(\mathscr{G}_{\mathscr{v}_1} \cup \mathscr{G}_{\mathscr{v}_2}) \subseteq G_a$.
- **Case II:** $\exists a \in [m]$ s.t. $(\mathscr{G}_{\mathscr{v}_1} \cup \mathscr{G}_{\mathscr{v}_2}) \subseteq G_a$.

Under case I, if only $\mathscr{G}_{\mathscr{v}_1}$ and $\mathscr{G}_{\mathscr{v}_2}$ have non-zero values in β, we obtain:

$$\begin{aligned}\sum_{g\in[m]} w_g ||\beta_{G_g}||_2 &= \Big(\sum_{g\in F(\mathscr{v}_1)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_1}}} + \Big(\sum_{g\in F(\mathscr{v}_2)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_2}}} \\ &\leqslant ||\beta_{\{\tilde{G},\tilde{w}\}}||_{q_1,q_2} \leqslant \sum_{\mathscr{g}\in[\mathscr{m}]} \mathscr{w}_{\mathscr{g}} ||\beta_{\mathscr{G}_{\mathscr{g}}}||_2 \\ &= \mathscr{w}_{\mathscr{v}_1}\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_1}}} + \mathscr{w}_{\mathscr{v}_2}\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_2}}} \\ &= \Big(\sum_{g\in F(\mathscr{v}_1)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_1}}} + \Big(\sum_{g\in F(\mathscr{v}_2)} w_g\Big)\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_2}}},\end{aligned}$$

which leads to

$$\mathscr{w}_{\mathscr{v}_1}\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_1}}} + \mathscr{w}_{\mathscr{v}_2}\sqrt{\beta^2_{\mathscr{G}_{\mathscr{v}_2}}} = \tilde{w}_{\tilde{g}}\Big(\sum_{j\in\tilde{G}_{\tilde{g}}} |\beta_j|^{q_2}\Big)^{\frac{1}{q_2}} = \tilde{w}_{\tilde{g}}\Big(\sum_{j\in\{\mathscr{G}_{\mathscr{v}_1}\cup\mathscr{G}_{\mathscr{v}_2}\}} |\beta_j|^{q_2}\Big)^{\frac{1}{q_2}}.$$

This equation does not hold by picking $j \in \mathscr{G}_{\mathscr{v}_1}, k \in \mathscr{G}_{\mathscr{v}_2}$, and setting

$$\begin{cases}\beta_j = \beta_k = 1, \beta_{\{[p]\backslash\{j,k\}\}} = 0 & \text{if } \mathscr{w}_{\mathscr{v}_1} + \mathscr{w}_{\mathscr{v}_2} \neq \tilde{w}_{\tilde{g}} \cdot 2^{\frac{1}{q_2}} \\ \beta_j = 2, \beta_k = 1, \beta_{\{[p]\backslash\{j,k\}\}} = 0 & \text{if } \mathscr{w}_{\mathscr{v}_1} + \mathscr{w}_{\mathscr{v}_2} = \tilde{w}_{\tilde{g}} \cdot 2^{\frac{1}{q_2}}\end{cases}.$$

Therefore, $|V| > 1$ cannot happen.

Under case II, let $\beta_j \in \mathscr{G}_{\mathscr{v}_1}$ and $\beta_k \in \mathscr{G}_{\mathscr{v}_2}$. Define β^j as the vector with 1 at the j-th element and 0 elsewhere, and β^k as the vector with 1 at the k-th element and 0 elsewhere, with $j \neq k$.

When $\beta = \beta^j$, we have:

$$\sum_{g\in[m]} w_g ||\beta_{G_g}||_2 = \Big(\sum_{g\in F(\mathscr{v}_1)} w_g\Big) \leqslant \tilde{w}_{\tilde{g}} \leqslant \sum_{\mathscr{g}\in[\mathscr{m}]} \mathscr{w}_{\mathscr{g}} ||\beta_{\mathscr{G}_{\mathscr{g}}}||_2 = \mathscr{w}_{\mathscr{v}_1},$$

indicating that $\tilde{w}_{\tilde{g}} = \mathscr{w}_{\mathscr{v}_1}$ for all q_1, q_2. Similarly, for $\beta = \beta^k$, we have:

$$\sum_{g \in [m]} w_g ||\beta_{G_g}||_2 = \Big(\sum_{g \in F(\mathscr{v}_2)} w_g \Big) \leqslant \tilde{w}_{\tilde{g}} \leqslant \sum_{\mathscr{g} \in [\mathscr{m}]} \mathscr{w}_{\mathscr{g}} ||\beta_{\mathscr{G}_{\mathscr{g}}}||_2 = \mathscr{w}_{\mathscr{v}_2},$$

indicating that $\tilde{w}_{\tilde{g}} = \mathscr{w}_{\mathscr{v}_2}$ for all q_1, q_2.

If $\mathscr{w}_{\mathscr{v}_1} \neq \mathscr{w}_{\mathscr{v}_2}$, then such a weight assignment is not feasible. Assuming $\mathscr{w}_{\mathscr{v}_1} = \mathscr{w}_{\mathscr{v}_2} = w_{\tilde{g}} = k$, then for any β with non-zero values only in $\mathscr{G}_{\mathscr{v}_1}$, we have $w_{\tilde{g}} ||\beta_{\mathscr{G}_{\mathscr{v}_1}}||_2 = (w_{\tilde{g}} ||\beta_{\mathscr{G}_{\mathscr{v}_1}}||_{q_2}^{q_1})^{\frac{1}{q_1}}$, implying that if a norm satisfies (11), it must be an ℓ_1/ℓ_2 norm.

Since $\mathscr{G}_{\mathscr{v}_1}$ and $\mathscr{G}_{\mathscr{v}_2}$ are different groups, there is at least one original group that contains variables in $\mathscr{G}_{\mathscr{v}_1}$ but not in $\mathscr{G}_{\mathscr{v}_2}$, and vice versa. Taking β with non-zero values in both $\mathscr{G}_{\mathscr{v}_1}$ and $\mathscr{G}_{\mathscr{v}_2}$, we have:

$$\sum_{g \in [m]} k ||\beta_{G_g}||_2 > k ||\beta_{\mathscr{G}_{\mathscr{v}_1}} \cup \beta_{\mathscr{G}_{\mathscr{v}_2}}||_2 = ||\beta_{\{\tilde{G}, \tilde{w}\}}||_{1,2},$$

which is a contradiction. Hence, in both cases, $|V| > 1$ is not possible, implying that there exists a $\mathscr{g} \in [\mathscr{m}]$ such that $\tilde{G}_{\tilde{g}} = \mathscr{G}_{\mathscr{g}}$. ■

D.2 Proof of Theorem 2

Proof We begin by examining the bound for the estimator $\hat{\beta}^G$. Considering a fixed design matrix X and a group structure G that comply with Assumption 7, and selecting an appropriate λ_n, Theorem 7 asserts that both inequalities (17) and (19) hold with a probability of at least $1 - e^{-2n\delta}$.

Under Assumptions 1,2, and 3, Theorem 8 establishes that Assumption 7 is valid with a probability of at least $1 - e^{-c_2 n \delta^2} - \frac{e^{-\frac{n}{32}}}{1 - e^{\frac{n}{64}}}$, where c_2 is a positive constant.

Considering these two theorems together, we conclude that under Assumptions 1,2, and 3, both (17) and (19) are satisfied with a probability of at least $1 - e^{-c_2 n \delta^2} - e^{-2n\delta} - \frac{e^{-\frac{n}{32}}}{1 - e^{\frac{n}{64}}}$. This probability can be further bounded below by $1 - e^{-c' n \delta}$ for some suitable constant c'.

The bound for $\hat{\beta}^{\mathscr{G}}$ can be directly derived, noting that it represents a group lasso estimator with group $\mathscr{G}$ and weights $\mathscr{w}$.

■

D.3 Proof of Corollary 3

Proof Assuming that $\max\{d_{\max}, m\} \asymp \max\{\mathscr{d}_{\max}, \mathscr{m}\}$, then we have

$$\Big(\frac{\mathscr{d}_{\max} \log 5}{n} + \frac{\log \mathscr{m}}{n} + \delta \Big) \asymp \Big(\frac{d_{\max} \log 5}{n} + \frac{\log m}{n} + \delta \Big).$$

Let $\mathscr{w}_{\mathscr{g}} = \sum_{g \in F(\mathscr{g})} w_g$, by the Cauchy–Schwarz inequality, we have

$$\mathscr{w}_{\mathscr{g}}^2 = \Big(\sum_{g \in F(\mathscr{g})} w_g \Big)^2 \leqslant \mathscr{h}_{\mathscr{g}} \Big(\sum_{g \in F(\mathscr{g})} w_g^2 \Big).$$

Therefore,

$$\sum_{\mathscr{g}\in F^{-1}(S)} \mathscr{w}_{\mathscr{g}}^2 \leqslant \sum_{\mathscr{g}\in F^{-1}(S)} \hbar_{\mathscr{g}}\big(\sum_{g\in F(\mathscr{g})} w_g^2\big) \leqslant h_{\max}^{G_{\overline{S}}}\Big(\sum_{\mathscr{g}\in F^{-1}(S)}\sum_{g\in F(\mathscr{g})} w_g^2\Big).$$

Let's introduce k_g as the number of non-overlapping groups from G into which the gth group is partitioned in the new structure $\mathscr{G}$. We also define K as the maximum number of such partitions, i.e., $K = \max_g k_g$ and $K \leqslant \infty$. Now we want to show that

$$\sum_{\mathscr{g}\in F^{-1}(S)}\sum_{g\in F(\mathscr{g})} w_g^2 \leqslant \sum_{g\in\overline{S}} k_g w_g^2.$$

Recall the definition of $F^{-1}(S)$ as:

$$F^{-1}(S) = \{\mathscr{g} \mid \mathscr{g} \in F^{-1}(g), g \in S\}.$$

For each $\mathscr{g} \in F^{-1}(g)$ that also belongs to $F^{-1}(S)$, we add w_g^2 to the summation. Therefore, the maximum contribution from each original group g to the sum $\sum\limits_{\mathscr{g}\in F^{-1}(S)}\sum\limits_{g\in F(\mathscr{g})} w_g^2$ is $k_g w_g^2$.

Given that

$$\{g | g \in F(\mathscr{g}) \text{ and } \mathscr{g} \in F^{-1}(S)\} = \overline{S},$$

we have

$$h_{\max}^{G_{\overline{S}}}\Big(\sum_{\mathscr{g}\in F^{-1}(S)}\sum_{g\in F(\mathscr{g})} w_g^2\Big) \leqslant h_{\max}^{G_{\overline{S}}}\sum_{g\in\overline{S}} k_g w_g^2 \leqslant h_{\max}^{G_{\overline{S}}} K \sum_{g\in\overline{S}} w_g^2.$$

On the other hand, we have

$$\begin{aligned}\min_{\mathscr{g}\in[m]}\left(\mathscr{w}_{\mathscr{g}}^2\right) &= \min_{\mathscr{g}\in[m]}\Big(\sum_{g\in F(\mathscr{g})} w_g\Big)^2 \geqslant \min_{\mathscr{g}\in[m]}\left(\sum_{g\in F(\mathscr{g})}\min_{g\in[m]}\{w_g\}\right)^2 \\ &\geqslant \min_{\mathscr{g}\in[m]}\left(h_{\min}^{g}\min_{g\in[m]}\{w_g\}\right)^2 = \left(h_{\min}^{g}\min_{g\in[m]}\{w_g\}\right)^2 \geqslant \min_{g\in[m]}\left(w_g^2 h_{\min}^{g}\right).\end{aligned}$$

Therefore,

$$\frac{\sum\limits_{\mathscr{g}\in F^{-1}(S)} \mathscr{w}_{\mathscr{g}}^2}{\min\limits_{\mathscr{g}\in[m]}\left(\mathscr{w}_{\mathscr{g}}^2\right)} \leqslant \frac{K\Big(\sum\limits_{g\in\overline{S}} w_g^2\Big)\cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^{g}\right)}.$$

Consequently, if K is upper bounded by a constant, then

$$\frac{\sigma^2}{\kappa^2}\cdot\frac{\sum\limits_{\mathscr{g}\in F^{-1}(S)} \mathscr{w}_{\mathscr{g}}^2}{\min\limits_{\mathscr{g}\in[m]}\left(\mathscr{w}_{\mathscr{g}}^2\right)}\cdot\left(\frac{\mathscr{d}_{\max}\log 5}{n} + \frac{\log \mathscr{m}}{n} + \delta\right) \lesssim \frac{\sigma^2}{\kappa^2}\cdot\frac{\Big(\sum\limits_{g\in\overline{S}} w_g^2\Big)\cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^{g}\right)}\cdot\left(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\right).$$

■

D.4 Proof of Proposition 1

Proof Let H_{G_g} be the sub-matrix of H consisting of the columns indexed by G_g. Let u_{G_g}, v_{G_g} be the sub-vectors of u, v indexed by G_g respectively. Given two vectors $u, v \in \mathbb{R}^p$, we have

$$\begin{aligned}
\phi^*(v) &= \sup_{\phi(u)\leqslant 1} \left\{u^T v\right\} = \sup_{\phi(u)\leqslant 1} \{u_1 v_1 + u_2 v_2 + \cdots + u_p v_p\} \\
&= \sup_{\phi(u)\leqslant 1} \left\{\frac{v_1}{h_1} \cdot h_1 \cdot u_1 + \cdots + \frac{v_p}{h_p} \cdot h_p \cdot u_p\right\} \\
&= \sup_{\phi(u)\leqslant 1} \left\{\sum_{g=1}^{m} \left(H_{G_g} v_{G_g}\right)^T u_{G_g}\right\} = \sup_{\phi(u)\leqslant 1} \left\{\sum_{g=1}^{m} \frac{\left((Hv)_{G_g}\right)}{w_g} \cdot w_g \cdot u_{G_g}\right\} \\
&\leqslant \sup_{\phi(u)\leqslant 1} \left\{\sum_{g=1}^{m} \frac{\left\|(Hv)_{G_g}\right\|_2}{w_g} \cdot \left\|w_g u_{G_g}\right\|_2\right\} \leqslant \left(\max_{g\in[m]} \frac{1}{w_g} \cdot \left\|(Hv)_{G_g}\right\|_2\right) \cdot \phi(u) \\
&\leqslant \max_{g\in[m]} \frac{1}{w_g} \cdot \left\|(Hv)_{G_g}\right\|_2 ,
\end{aligned}$$

where the first inequality is achieved by using Cauchy's inequality.

Let $g_0 = \arg\max_{g\in[m]} \frac{1}{w_g}\left\| (Hv)_{G_g} \right\|_2$ and $h_{\max}^{g_0} = 1$. Define $u \in \mathbb{R}^p$ as:

$$u_j = \begin{cases} 0 \; \mathit{for} \;\; j \notin G_{g_0} \\ \frac{1}{w_{g_0}} \cdot \frac{v_j}{{h_j}^2} \cdot \frac{1}{\left\|(Hv)_{G_{g_0}}\right\|_2} \;\; \mathit{for} \;\; j \in G_{g_0}, \end{cases}$$

then we have

$$\begin{aligned}
\phi\left(u\right) &= \sum_{g=1}^{m} w_g \left\|u_{G_g}\right\|_2 = w_{g_0} \cdot \frac{1}{w_{g_0}} \cdot \frac{1}{\left\| (Hv)_{G_{g_0}} \right\|_2} \cdot \sqrt{\sum_{j\in G_{g_0}} \frac{{v_j}^2}{{h_j}^4}} \\
&= \frac{1}{\left\| (Hv)_{G_{g_0}} \right\|_2} \sqrt{\sum_{j\in G_{g_0}} \frac{{v_j}^2}{{h_j}^2}} = 1,
\end{aligned}$$

where the last equality holds due to the fact that $h_j = 1$ for any $j \in G_{g_0}$, and we also have

$$\begin{aligned}
u^T v &= \frac{1}{w_{g_0}} \frac{1}{\left\| (Hv)_{G_{g_0}} \right\|_2} \cdot \sum_{j\in G_{g_0}} \frac{{v_j}^2}{{h_j}^2} = \frac{1}{w_{g_0}} \frac{1}{\left\| (Hv)_{G_{g_0}} \right\|_2} \cdot \left\| (Hv)_{G_{g_0}} \right\|_2^2 \\
&= \frac{1}{w_{g_0}} \left\| (Hv)_{G_{g_0}} \right\|_2 = \max_{g\in[m]} \frac{1}{w_{g_0}} \left\| (Hv)_{G_g} \right\|_2 = \phi^*\left(v\right).
\end{aligned}$$

Therefore, this is a sharp bound.

■

D.5 Proof of Theorem 7

Proof This section mostly follow the proof in Wainwright (2019, Chap. 14). For simplicity, we write $S = S(\beta^*)$ and $\overline{S} = \overline{S}(\beta^*)$. From the optimality of $\hat{\beta}^G$, we have

$$
\begin{aligned}
0 &\geqslant \frac{1}{n}\big\|Y - X\hat{\beta}\big\|_2^2 - \frac{1}{n}\big\|Y - X\beta^*\big\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&= \frac{1}{n}\left(Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} - Y^TY + 2Y^TX\beta^* - \beta^{*T}X^TX\beta^*\right) + \lambda_n\left(\phi(\hat{\beta}) - \phi\left(\beta^*\right)\right) \\
&= \frac{1}{n}\left((2X^TX\beta^* - 2X^TY)^T(\hat{\beta} - \beta^*) + (\hat{\beta} - \beta^*)^TX^TX(\hat{\beta} - \beta^*)\right) + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&= \Big\langle \bigtriangledown \frac{\big\|Y - X\beta^*\big\|_2^2}{n}, \big(\hat{\beta} - \beta^*\big)\Big\rangle + \frac{\big\|X\big(\hat{\beta} - \beta^*\big)\big\|_2}{n} + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&\geqslant \Big\langle \bigtriangledown \frac{\big\|Y - X\beta^*\big\|_2^2}{n}, \big(\hat{\beta} - \beta^*\big)\Big\rangle + \kappa\big\|\big(\hat{\beta} - \beta^*\big)\big\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right) \\
&\geqslant -\Big|\Big\langle \bigtriangledown \frac{\big\|Y - X\beta^*\big\|_2^2}{n}, \big(\hat{\beta} - \beta^*\big)\Big\rangle\Big| + \kappa\big\|\big(\hat{\beta} - \beta^*\big)\big\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right),
\end{aligned}
$$

where the penultimate step is valid due to the assumption of restrictive strong convexity.

By applying Holder's inequality with the regularizer ϕ and its dual norm ϕ^*, we have

$$
\Big|\Big\langle \bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}, \big(\hat{\beta} - \beta^*\big)\Big\rangle\Big| \leqslant \phi^*\Big(\bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}\Big)\phi\big(\hat{\beta} - \beta^*\big). \tag{32}
$$

Next, we have

$$
\begin{aligned}
\phi(\hat{\beta}) = \phi\left(\beta^* + (\hat{\beta} - \beta^*)\right) &= \phi\left(\beta^*_{M(S)} + \beta^*_{M^\perp(S)} + (\hat{\beta} - \beta^*)_{M(\overline{S})} + (\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) \\
&\geqslant \phi\left(\beta^*_{M(S)} + (\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi(\beta^*_{M^\perp(S)}) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) \\
&= \phi(\beta^*_{M(S)}) + \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi(\beta^*_{M^\perp(S)}) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right).
\end{aligned}
$$

The inequality holds by applying the triangle inequality on $\phi(\hat{\beta})$, and the last step holds by applying Lemma 11. Consequently, we have

$$
\begin{aligned}
\phi(\hat{\beta}) - \phi\left(\beta^*\right) &\geqslant \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) - 2\phi(\beta^*_{M^\perp(S)}) \\
&= \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right),
\end{aligned} \tag{33}
$$

where $\phi\left(\beta^*_{M^\perp(S)}\right) = 0$ as $\beta^*_{M^\perp(S)}$ is a zero vector.

Based on (32) and (33), we have

$$\begin{aligned}
&\frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right)\\
&\geqslant -\left|\left\langle \bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right) \right\rangle\right| + \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right)\\
&\geqslant \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right) - \left|\left\langle \bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}, \left(\hat{\beta} - \beta^*\right) \right\rangle\right|\\
&\geqslant \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right) - \phi^*\Big(\bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}\Big)\phi\left(\hat{\beta} - \beta^*\right)\\
&\geqslant \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right) - \frac{\lambda_n}{2}\phi\left(\hat{\beta} - \beta^*\right),
\end{aligned}$$

where the last step is valid because Lemma 10 implies that we can guarantee $\lambda_n \geqslant 2\phi^*\left(\bigtriangledown \frac{\|Y - X\beta^*\|_2^2}{n}\right)$ with high probability by taking appropriate λ_n. Moreover, Lemma 12 implies that

$$\hat{\beta} \in \left\{\beta \in \mathbb{R}^p \mid \phi\left((\beta - \beta^*)_{M^\perp(\overline{S})}\right) \leqslant 3\phi\left((\beta - \beta^*)_{M(\overline{S})}\right)\right\}.$$

By the triangle inequality, we have

$$\phi(\hat{\beta} - \beta^*) = \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})} + (\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) \leqslant \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) + \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right),$$

and hence we have

$$\begin{aligned}
&\frac{1}{n}\left\|Y - X\hat{\beta}\right\|_2^2 - \frac{1}{n}\left\|Y - X\beta^*\right\|_2^2 + \lambda_n\left(\phi(\hat{\beta}) - \phi(\beta^*)\right)\\
&\geqslant \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right) - \frac{\lambda_n}{2}\phi\left(\hat{\beta} - \beta^*\right)\\
&\geqslant \kappa\left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2 + \lambda_n\left(\phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right) - \phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\right)\\
&\qquad\qquad - \frac{\lambda_n}{2}\left(\phi\left((\hat{\beta} - \beta^*)_{M(S)}\right) + \phi\left((\hat{\beta} - \beta^*)_{M^\perp(\overline{S})}\right)\right)\\
&\geqslant \kappa\left\|\hat{\beta} - \beta^*\right\|_2^2 + \frac{\lambda_n}{2}\left(\phi(\hat{\beta} - \beta^*)_{M^\perp(\overline{S})} - 3\phi(\hat{\beta} - \beta^*)_{M(\overline{S})}\right)\\
&\geqslant \kappa\left\|\hat{\beta} - \beta^*\right\|_2^2 - \frac{3\lambda_n}{2}\phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right).
\end{aligned}$$

By definition, we have $\phi\left((\hat{\beta} - \beta^*)_{M(\overline{S})}\right) = \sum_{g \in \overline{S}} w_g \left\| \left(\hat{\beta} - \beta^*\right)_{G_g} \right\|_2$, and by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\sum_{g \in \overline{S}} w_g \left\|\left(\hat{\beta} - \beta^*\right)_{G_g}\right\|_2 &\leqslant \sqrt{\sum_{g \in \overline{S}} w_g^2} \cdot \sqrt{h_{\max}^{G_{\overline{S}}} \cdot \max_{g \in \overline{S}} \left\|\left(\hat{\beta} - \beta^*\right)_{G_g}\right\|_2^2}\\
&\leqslant \sqrt{\sum_{g \in \overline{S}} w_g^2} \cdot \sqrt{h_{\max}^{G_{\overline{S}}} \cdot \left\|\left(\hat{\beta} - \beta^*\right)\right\|_2^2}\\
&= \sqrt{\sum_{g \in \overline{S}} w_g^2} \cdot \sqrt{h_{\max}^{G_{\overline{S}}}} \left\|\left(\hat{\beta} - \beta^*\right)\right\|_2.
\end{aligned}$$

On the other hand, since $\kappa \left\|\hat{\beta} - \beta^*\right\|_2^2 - \frac{3\lambda_n}{2}\sqrt{\sum\limits_{g\in\overline{S}} w_g^2} \cdot \sqrt{h_{\max}^{G_{\overline{S}}}} \left\|\left(\hat{\beta} - \beta^*\right)\right\|_2 \leqslant 0$, we have

$$\begin{aligned}
\left\|\hat{\beta} - \beta^*\right\|_2^2 &\leqslant \frac{9\lambda_n^2}{4\kappa^2}\sum_{g\in\overline{S}} w_g^2 \cdot h_{\max}^{G_{\overline{S}}} \\
&\leqslant \frac{9}{4\kappa^2} \cdot \frac{64c^2\sigma^2 \sum\limits_{g\in\overline{S}} w_g^2 \cdot h_{\max}(\overline{S})}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)} \cdot \left(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\right) \\
&\leqslant \frac{144c^2\sigma^2}{\kappa^2} \cdot \frac{\sum\limits_{g\in\overline{S}} w_g^2 \cdot h_{\max}^{G_{\overline{S}}}}{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)} \cdot \left(\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta\right).
\end{aligned}$$

$\blacksquare$

D.5.1 LEMMAS FOR THE PROOF OF THEOREM 7

In these lemmas, we abbreviate $\hat{\beta}^G$ by $\hat{\beta}$.

Lemma 10. *Under the Assumption 7 and* (2)*, taking*

$$\lambda_n = \frac{8c\sigma}{\sqrt{\min\limits_{g\in[m]}\left(w_g^2 h_{\min}^g\right)}}\sqrt{\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta} \quad \textit{for some } \delta \in [0,1],$$

then $\mathbb{P}\left(\lambda_n \geqslant 2\phi^*(\frac{X^\top\varepsilon}{n})\right) \geqslant 1 - e^{-2n\delta}$.

Proof of Lemma 10 Let $V_{i\cdot g} = -\varepsilon_i\left(\frac{X_{ig_1}}{h_{g_1}w_g}, \frac{X_{ig_2}}{h_{g_2}w_g}, \ldots, \frac{X_{ig_{d_g}}}{h_{g_{d_g}}w_g}\right) \in \mathbb{R}^{d_g}$. According to the variational form of ℓ_2 norm, we have $\frac{1}{n}\left\|\sum_{i=1}^n V_{i\cdot g}\right\|_2 = \sup\limits_{u\in S^{d_g-1}}\left\langle u, \frac{1}{n}\sum_{i=1}^n V_{i\cdot g}\right\rangle$, where S^{d_g-1} is the Euclidean sphere in $\mathbb{R}^{d_g}$. Also, for any vector $u \in S^{d_g-1}$ and $t \in \mathbb{R}$, we have

$$\begin{aligned}
\frac{1}{n}\log\mathbb{E}\Big(e^{t\left\langle u, \sum\limits_{i=1}^n V_{i\cdot g}\right\rangle}\Big) &= \frac{1}{n}\log\mathbb{E}\Big(e^{t\sum\limits_{j=1}^{d_g} u_j \sum\limits_{i=1}^n V_{i\cdot g_j}}\Big) = \frac{1}{n}\log\mathbb{E}\Big(e^{t\sum\limits_{i=1}^n\left(\sum\limits_{j=1}^{d_g} u_j V_{i\cdot g_j}\right)}\Big) \\
&= \frac{1}{n}\log\mathbb{E}\Big(e^{-t\sum\limits_{i=1}^n\left(\sum\limits_{g=1}^{d_g}\frac{u_j X_{ig_j}\varepsilon_i}{h_{g_j}w_g}\right)}\Big) = \frac{1}{n}\log\mathbb{E}\Big(e^{-t\sum\limits_{i=1}^n \varepsilon_i\left(\sum\limits_{j=1}^{d_g}\frac{u_j X_{ig_j}}{h_{g_j}w_g}\right)}\Big).
\end{aligned}$$

Since $\{\epsilon_i\}_{i=1}^n$ are i.i.d zero mean sub-Gaussian random variables with parameter σ, let $u = (u_1, \cdots, u_{d_g})^T \in \mathbb{R}^{d_g \times 1}$, $X_{i,g} = (X_{ig_1}, \cdots X_{ig_{d_g}})^T \in \mathbb{R}^{d_g \times 1}$, then we have

$$\begin{aligned}
\frac{1}{n}\log\mathbb{E}\Bigg(e^{-t\sum\limits_{i=1}^{n}\varepsilon_i\left(\sum\limits_{j=1}^{d_g}\frac{u_j x_{ig_j}}{h_{g_j}w_g}\right)}\Bigg) &= \frac{1}{n}\log\mathbb{E}\Bigg(e^{-t\varepsilon_1\left(\sum\limits_{j=1}^{d_g}\frac{u_j X_{1g_j}}{h_{g_j}w_g}\right)}\Bigg) + \cdots + \frac{1}{n}\log\mathbb{E}\Bigg(e^{-t\varepsilon_n\left(\sum\limits_{j=1}^{d_g}\frac{u_j X_{ng_j}}{h_{g_j}w_g}\right)}\Bigg)\\
&\leqslant \frac{t^2\sigma^2}{2n}\Bigg(\sum_{i=1}^{n}\Big(\sum_{j=1}^{d_g}\frac{u_j X_{ig_j}}{w_g h_{g_j}}\Big)^2\Bigg) \leqslant \frac{t^2\sigma^2}{2n}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\Bigg(\sum_{i=1}^{n}\Big(\sum_{j=1}^{d_g}u_j X_{ig_j}\Big)^2\Bigg)\\
&= \frac{t^2\sigma^2}{2n}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\Bigg(\sum_{i=1}^{n}\langle u,\ X_{i,g}\rangle^2\Bigg) = \frac{t^2\sigma^2}{2n}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\Bigg(\sum_{i=1}^{n}(u^T X_{i,g}X_{i,g}^T u)\Bigg)\\
&= \frac{t^2\sigma^2}{2}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\Bigg(u^T\Big(\frac{1}{n}\sum_{i=1}^{n}X_{i,g}X_{i,g}^T\Big)u\Bigg)\\
&= \frac{t^2\sigma^2}{2}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\Bigg(u^T\frac{X_{G_g}^T X_{G_g}}{n}u\Bigg)\\
&\leqslant \frac{t^2\sigma^2}{2}\frac{1}{w_g^2\left(h_{\min}^g\right)^2}\Bigg(\gamma_{\max}\big(\frac{X_{G_g}^T X_{G_g}}{n}\big)\Bigg).
\end{aligned}$$

By Assumption 7, we have $\gamma_{\max}(\frac{X_{G_g}^T X_{G_g}}{n}) \leqslant c^2$. Combining this with the previous proof, we have $\frac{1}{n}\log\mathbb{E}\big(e^{t\langle u, \sum_{i=1}^{n}V_{i\cdot g}\rangle}\big) \leqslant c^2t^2\sigma^2/2w_g^2\left(h_{\min}^g\right)$. Therefore, the random variable $\langle u, \sum\limits_{i=1}^{n}V_{i\cdot g}\rangle$ is the sub-Gaussian with the parameter at most $\sqrt{c^2\sigma^2/w_g^2\left(h_{\min}^g\right)}$, and by properties of sub-Gaussian variables, we have

$$\log\mathbb{P}\Big(\Big\langle u, \sum_{i=1}^{n}V_{i\cdot g}\Big\rangle \geqslant \frac{\lambda_n}{4}\Big) \leqslant -\frac{\lambda_n^2 w_g^2 h_{\min}^g}{32C^2\sigma^2}.$$

We can find a $\frac{1}{2}$ covering of S^{d_g-1} in Euclidean norm:$\{u^1, u^2, \ldots, u^N\}$ with $N \leq 5^{d_g}$, recall that $\frac{1}{n}\left\|\sum_{i=1}^{n}V_{i\cdot g}\right\|_2 = \frac{1}{n}\sup\limits_{u\in S^{d_g-1}}\langle u, \sum\limits_{i=1}^{n}V_{i\cdot g}\rangle$, so that for any $u \in S^{d_g-1}$, we can find a $u^{q(u)} \in \{u^1, \ldots, u^N\}$, such that $\left\|u^{q(u)} - u\right\|_2 \leqslant \frac{1}{2}$, and

$$\begin{aligned}
\frac{1}{n}\sup_{u\in S^{d_g-1}}\Big\langle u, \sum_{i=1}^{n}V_{i\cdot g}\Big\rangle &= \frac{1}{n}\sup_{u\in S^{d_g-1}}\Big(\Big\langle u - u^{q(u)}, \sum_{i=1}^{n}V_{i\cdot g}\Big\rangle + \Big\langle u^{q(u)}, \sum_{i=1}^{n}V_{i\cdot g}\Big\rangle\Big)\\
&\leqslant \frac{1}{n}\sup_{u\in S^{d_g-1}}\Big\langle u - u^{q(u)}, \sum_{i=1}^{n}V_{i\cdot g}\Big\rangle + \frac{1}{n}\max_{q\in[N]}\Big\langle u^q, V_{i\cdot g}\Big\rangle.
\end{aligned}$$

By applying the Cauchy-Schwarz inequality, we have

$$\frac{1}{n}\sup_{u\in S^{d_g-1}}\Big\langle u - u^{q(u)}, \sum_{i=1}^{n}V_{i\cdot g}\Big\rangle \leqslant \frac{\left\|u - u^{q(u)}\right\|_2}{n}\Big\|\sum_{i=1}^{n}V_{i\cdot g}\Big\|_2 \leqslant \frac{1}{2n}\Big\|\sum_{i=1}^{n}V_{i\cdot g}\Big\|_2.$$

Hence, we obtain $\frac{1}{n}\Big\|\sum\limits_{i=1}^{n} V_{i\cdot g}\Big\|_2 \leqslant \frac{1}{2n}\Big\|\sum\limits_{i=1}^{n} V_{i\cdot g}\Big\|_2 + \frac{1}{n}\max\limits_{q\in[N]}\Big\langle u^q, \sum\limits_{i=1}^{n} V_{i\cdot g}\Big\rangle$, which indicates that

$$\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \leqslant 2\max_{q\in[N]}\Big\langle u^q, \frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\rangle.$$

Consequently, we can express the probability as

$$\begin{aligned}
\mathbb{P}\Big(\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \geqslant \frac{\lambda_n}{2}\Big) &\leqslant \mathbb{P}\Big(\max_{q\in[N]}\Big\langle u^q, \frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\rangle \geqslant \frac{\lambda_n}{4}\Big)\\
&\leqslant \sum_{q=1}^{N}\mathbb{P}\Big(\Big\langle u^q, \frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\rangle \geqslant \frac{\lambda_n}{4}\Big)\\
&\leqslant N\exp\Big(-\frac{n\lambda_n^2 w_g^2 h_{\min}^g}{32C^2\sigma^2}\Big) \leqslant \exp\Big(-\frac{n\lambda_n^2 w_g^2 h_{\min}^g}{32C^2\sigma^2} + d_g\log 5\Big),
\end{aligned}$$

and by setting $\lambda_n = \frac{8C\sigma}{\sqrt{\min\limits_{g\in[m]}(w_g^2 h_{\min}^g)}}\sqrt{\frac{d_{\max}\log 5}{n} + \frac{\log m}{n} + \delta}$, we get

$$\begin{aligned}
\mathbb{P}\Big(\max_{g\in[m]}\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \geqslant \frac{\lambda_n}{2}\Big) &\leqslant \sum_{g=1}^{m}\mathbb{P}\Big(\frac{1}{n}\Big\|\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2 \geqslant \frac{\lambda_n}{2}\Big)\\
&\leqslant \exp\Big(-\frac{n\lambda_n^2}{32C^2\sigma^2}\min_{g\in[m]}(w_g^2 h_{\min}^g) + d_{\max}\log 5 + \log m\Big)\\
&\leqslant \exp\{-2n\delta\}.
\end{aligned}$$

From proposition 1, we have

$$\phi^*\Big(\frac{X^\top\varepsilon}{n}\Big) \leqslant \max_{g\in[m]}\frac{1}{w_g}\Big\|\Big(\frac{HX^\top\varepsilon}{n}\Big)_{G_g}\Big\|_2 = \max_{g\in[m]}\frac{1}{w_g}\Big\|\frac{1}{n}\sum_{i=1}^{n} -\varepsilon_i\Big(\frac{X_{ig_1}}{h_{g_1}}, \cdots, \frac{X_{ig_{d_g}}}{h_{g_{d_g}}}\Big)\Big\|_2 = \max_{g\in[m]}\Big\|\frac{1}{n}\sum_{i=1}^{n} V_{i\cdot g}\Big\|_2.$$

Therefore, $\mathbb{P}\left(\lambda_n \geqslant 2\phi^*(\frac{X^\top\varepsilon}{n})\right) \geqslant 1 - e^{-2n\delta}$.

■

Lemma 11. *The group lasso regularizer* (1) *is decomposable with respect to the pair* $\{M(S), M^\perp(\overline{S})\}$. *That is,* $\phi(a+b) = \phi(a) + \phi(b)$, *for all* $a \in M(S)$ *and for all* $b \in M^\perp(\overline{S})$.

Proof of Lemma 11 By definition, we have

$$\begin{aligned}
\phi(a+b) &= \sum_{g=1}^{m} w_g\left\|(a+b)_{G_g}\right\|_2 = \sum_{g\in M(\overline{S})} w_g\left\|(a+b)_{G_g}\right\|_2 + \sum_{g\notin M(\overline{S})} w_g\left\|(a+b)_{G_g}\right\|_2\\
&= \sum_{g\in M(\overline{S})} w_g\left\|a_{G_g}\right\|_2 + \sum_{g\in M^\perp(\overline{S})} w_g\left\|b_{G_g}\right\|_2 = \sum_{g\in M(S)} w_g\left\|a_{G_g}\right\|_2 + \sum_{g\in M^\perp(\overline{S})} w_g\left\|b_{G_g}\right\|_2\\
&= \phi(a) + \phi(b).
\end{aligned}$$

■

Lemma 12. *If $\lambda_n \geqslant 2\phi^*\left(\frac{X^T\varepsilon}{n}\right)$, then $\phi\left(\left(\hat\beta-\beta^*\right)_{M^\perp(\overline{S})}\right) \leqslant 3\phi\left(\left(\hat\beta-\beta^*\right)_{M(\overline{S})}\right)$.*

Proof of Lemma 12 (see Wainwright, 2019, proposition 9.13) From equation (33), we have

$$\phi(\hat\beta)-\phi\left(\beta^*\right) \geqslant \phi\left(\left(\hat\beta-\beta^*\right)_{M^\perp(\overline{S})}\right)-\phi\left(\left(\hat\beta-\beta^*\right)_{M(\overline{S})}\right),$$

On the other hand, by the convexity of the cost function, we have

$$\frac{1}{n}\big\|Y-X\hat\beta\big\|_2^2-\frac{1}{n}\big\|Y-X\beta^*\big\|_2^2 \geqslant \Big\langle \bigtriangledown \frac{\big\|Y-X\beta^*\big\|_2^2}{n}, \left(\hat\beta-\beta^*\right)\Big\rangle \geqslant -\Big\langle \bigtriangledown \frac{\big\|Y-X\beta^*\big\|_2^2}{n}, \left(\hat\beta-\beta^*\right)\Big\rangle.$$

By applying Holder's inequality with the regularizer ϕ and its dual norm ϕ^*, we have

$$\left|\Big\langle \bigtriangledown \frac{\|Y-X\beta^*\|_2^2}{n}, \left(\hat\beta-\beta^*\right)\Big\rangle\right| \leqslant \phi^*\bigg(\bigtriangledown \frac{\|Y-X\beta^*\|_2^2}{n}\bigg)\phi\left(\hat\beta-\beta^*\right).$$

Therefore,

$$\begin{aligned}\frac{1}{n}\big\|Y-X\hat\beta\big\|_2^2-\frac{1}{n}\big\|Y-X\beta^*\big\|_2^2 &\geqslant -\Big\langle \bigtriangledown \frac{\|Y-X\beta^*\|_2^2}{n}, \left(\hat\beta-\beta^*\right)\Big\rangle \geqslant -\phi^*\bigg(\bigtriangledown \frac{\|Y-X\beta^*\|_2^2}{n}\bigg)\phi\left(\hat\beta-\beta^*\right)\\ &\geqslant -\frac{\lambda_n}{2}\phi\left(\hat\beta-\beta^*\right) \geqslant -\frac{\lambda_n}{2}\left(\phi(\hat\beta-\beta^*)_{M(\overline{S})}+\phi(\hat\beta-\beta^*)_{M^\perp(\overline{S})}\right),\end{aligned}$$

and

$$\begin{aligned}0 &= \frac{1}{n}\big\|Y-X\hat\beta\big\|_2^2-\frac{1}{n}\big\|Y-X\beta^*\big\|_2^2+\lambda_n\left(\phi(\hat\beta)-\phi(\beta^*)\right)\\ &\geq \lambda_n\left(\phi\left(\left(\hat\beta-\beta^*\right)_{M^\perp(\overline{S})}\right)-\phi\left(\left(\hat\beta-\beta^*\right)_{M(\overline{S})}\right)-2\phi(\beta^*_{M^\perp(S)})\right)-\frac{\lambda_n}{2}\left(\phi(\hat\beta-\beta^*)_{M(\overline{S})}+\phi(\hat\beta-\beta^*)_{M^\perp(\overline{S})}\right)\\ &= \frac{\lambda_n}{2}\left(\phi\left(\left(\hat\beta-\beta^*\right)_{M^\perp(\overline{S})}\right)-3\phi\left(\hat\beta-\beta^*\right)_{M(\overline{S})}\right)\right),\end{aligned}$$

from which the claim follows. ■

D.6 Proof of Theorem 8

Two lemmas are used in this proof:

Lemma 13 (Theorem 6.5 in (Wainwright, 2019)). *Let $|||.|||_2$ be the spectral norm of a matrix. There are universal constants c_2, c_3, c_4, c_5 such that, for any matrix $A \in \mathbb{R}^{n\times p}$, if all rows are drawn i.i.d from $N(0,\Theta)$, then the sample covariance matrix $\hat\Theta$ satisfies the bound*

$$\mathbb{E}\left(e^{t|||\hat\Theta-\Theta|||_2}\right) \leqslant e^{c_3\frac{t^2\theta^2}{n}+4p} \quad \textit{for all} \ \ |t| < \frac{n}{64e^2|||\Theta|||_2},$$

and hence for all $\delta \in [0,1]$

$$\mathbb{P}\Big(\frac{|||\hat\Theta-\Theta|||_2}{|||\Theta|||_2} \leqslant c_5(\sqrt{\frac{p}{n}}+\frac{p}{n})+\delta\Big) > 1-c_4e^{-c_2n\delta^2}. \tag{34}$$

Lemma 14. *Under Assumptions 1,2, and 3, and use $\rho(\Theta)$ to denote the maximum diagonal of a covariance matrix Θ. For any vector $\beta \in \mathbb{R}^p$ and a given group structure with m groups, we have*

$$\frac{\|X\beta\|_2}{\sqrt{n}} \geq \frac{1}{4}\left\|\Theta^{\frac{1}{2}}\beta\right\|_2 - 8\rho(\Theta)\left(\max_{g\in[m]}\frac{1}{w_g\sqrt{h^g_{\min}}}\right)\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\phi(\beta), \tag{35}$$

with probability at least $1-\frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{64}}}$.

Proof We first prove the first part of Theorem 8. By Lemma 13, we have

$$\mathbb{P}\Big(\frac{|||\frac{X^T_{G_g}X_{G_g}}{n} - \Theta_{G_g,G_g}|||_2}{|||\Theta_{G_g,G_g}|||_2} \leqslant c_5(\sqrt{\frac{d_g}{n}}+\frac{d_g}{n})+\delta\Big) > 1-c_4e^{-c_2n\delta^2}.$$

By the triangle inequality, since $X^T_{G_g}X_{G_g}$ is a positive semi-definite, we have

$$\begin{aligned}\gamma_{\max}(\frac{X^T_{G_g}X_{G_g}}{n}) &= |||\frac{X^T_{G_g}X_{G_g}}{n}|||_2 = |||\frac{X^T_{G_g}X_{G_g}}{n} - \Theta_{G_g,G_g}|||_2 + |||\Theta_{G_g,G_g}|||_2\\ &\leqslant (1+c_5(\sqrt{\frac{d_g}{n}}+\frac{d_g}{n})+\delta)|||\Theta_{G_g,G_g}|||_2,\end{aligned}$$

with probability at least $1-c_4e^{-c_2n\delta^2}$. Because $|||\Theta_{G_g,G_g}|||_2 \leq |||\Theta|||_2 \leqslant c_1$ for some constant c_1 and $d_g \leqslant n$, we have $\gamma_{\max}(\frac{X^T_{G_g}X_{G_g}}{n}) \leqslant c+\delta$ for some constant c, with probability at least $1-e^{-c_2n\delta^2}$. Taking the union probability for all m groups, we have

$$\max_{g\in[m]}\gamma_{\max}(\frac{X^T_{G_g}X_{G_g}}{n}) \leq c+\delta$$

with probability at least $1-\exp(-c'2n\delta^2)$ for some constant $c'>0$ as long as

$$\log m \ll n\delta^2.$$

For simplicity, we take δ as a constant.

Now we proceed to prove the second part. First note that we must have $\rho(\Theta) \leq \gamma_{\max}(\Theta) \leq c_1$ by Assumptions 1,2, and 3. By applying Minkowski inequality, we have

$$\phi(\beta) = \sum_{g=1}^m w_g\left\|\beta_{G_g}\right\|_2 \leqslant \sqrt{m}\sqrt{\sum_{g=1}^m w_g^2\left\|\beta_{G_g}\right\|_2^2} \leqslant \sqrt{m}\sqrt{\max_{g\in[m]} w_g^2h^g_{\max}\|\beta\|_2^2}.$$

Let $\beta = \beta^* - \bar{\beta}$, we now want to prove that $\phi\left(\beta_{M^\perp(\bar{S})}\right) \leqslant 3\phi\left(\beta_{M(\bar{S})}\right)$ implies $\frac{\|X\beta\|_2^2}{n} \geqslant \frac{\gamma_{\min}}{64}\|\beta\|_2^2$.
Since $\phi\left(\beta_{M^\perp(\bar{S})}\right) \leqslant 3\phi\left(\beta_{M(\bar{S})}\right)$, combining with triangle inequality, we have

$$\begin{aligned}\phi(\beta) = \phi\left(\beta_{M(\bar{S})}\right) + \phi\left(\beta_{M^\perp(\bar{S})}\right) \leqslant 4\phi\left(\beta_{M(\bar{S})}\right) &\leqslant 4\sqrt{s_g}\sqrt{\max_{g\in\bar{S}} w_g^2h^g_{\max}}\left\|\beta_{M(\bar{S})}\right\|_2\\ &\leqslant 4\sqrt{s_g}\sqrt{\max_{g\in\bar{S}} w_g^2h^g_{\max}}\|\beta\|_2.\end{aligned}$$

From Lemma 14, we have

$$
\begin{aligned}
\frac{\|X\beta\|_2}{\sqrt{n}} &\geqslant \frac{1}{4}\left\|\Theta^{\frac{1}{2}}\beta\right\|_2 - 8\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\phi(\beta)\\
&\geqslant \frac{1}{4\sqrt{c_1}}\|\beta\|_2 - 32\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\sqrt{s_g}\sqrt{\max_{g\in\bar{S}} w_g^2 h_{\max}^g}\,\|\beta\|_2\\
&\geqslant \frac{1}{64\sqrt{c_1}}\|\beta\|_2\,,
\end{aligned}
$$

where the last step is valid due to Assumption 2 and 3. ■

D.6.1 LEMMAS FOR THE PROOF OF THEOREM 8

Proof of Lemma 14 To begin with, for a vector $\beta \in \mathbb{R}^p$ with a fixed group structure, we define the set:

$$S^{p-1}(\Theta) = \left\{\beta \in \mathbb{R}^p \middle| \left\|\Theta^{\frac{1}{2}}\beta\right\|_2 = 1\right\},$$

the function:

$$g(t) = 4\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\cdot t,$$

and the event:

$$\mathscr{E}\left(S^{p-1}(\Theta)\right) = \left\{X \in \mathbb{R}^{n\times p} \middle| \inf_{\beta\in S^{p-1}(\Theta)}\frac{\|X\beta\|_2}{\sqrt{n}} + 2g(\phi(\beta)) \leqslant \frac{1}{4}\right\},$$

where $\phi(.)$ is the overlapping group lasso regularizer. In addition, given $0 \leqslant r_\ell \leqslant r_u$, we define the set

$$\mathbb{K}(r_\ell, r_u) = \left\{\beta \in S^{p-1}(\Theta) \middle| g(\phi(\beta)) \in [r_\ell, r_u]\right\},$$

and the event:

$$\mathscr{A}(r_\ell, r_u) = \left\{X \in \mathbb{R}^{n\times p} \middle| \inf_{\beta\in\mathbb{K}(r_\ell,r_u)}\frac{\|X\beta\|_2}{\sqrt{n}} \leqslant \frac{1}{2} - r_u\right\}.$$

Now we introduce two additional lemmas:

Lemma 15. *For $\upsilon = \frac{1}{4}$, we have $\mathscr{E} \subseteq \mathscr{A}(0,\upsilon) \cup \left(\bigcup_{\ell=1}^{\infty}\mathscr{A}\left(2^{\ell-1}\upsilon, 2^\ell\upsilon\right)\right)$.*

Lemma 16. *For any pair (r_ℓ, r_u), where $0 \leqslant r_\ell \leqslant r_u$, we have $\mathbb{P}(\mathscr{A}(r_\ell, r_u)) \leqslant e^{-\frac{n}{32}}e^{-\frac{n}{8}r_u^2}$.*

Based on Lemma 15 and Lemma 16, we have

$$\mathbb{P}(X \in \mathscr{E}) \leqslant \mathbb{P}(\mathscr{A}(0,\upsilon)) + \sum_{\ell=1}^{\infty}\mathbb{P}\left(\mathscr{A}(2^{\ell-1}\upsilon, 2^\ell\upsilon)\right) \leqslant e^{-\frac{n}{32}}\left\{\sum_{t=0}^{\infty}e^{-\frac{n}{8}2^{2\ell}\upsilon^2}\right\}.$$

Since $\upsilon = \frac{1}{4}$ and $2^{2\ell} \geqslant 2\ell$, we have

$$\mathbb{P}(X \in \mathscr{E}) \leqslant e^{-\frac{n}{32}}\sum_{\ell=0}^{\infty}e^{-\frac{n}{8}2^{2\ell}\upsilon^2} \leqslant e^{-\frac{n}{32}}\sum_{\ell=0}^{\infty}e^{-n\frac{\ell}{4}\upsilon^2} \leqslant \frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{64}}}.$$

We now get the upper bound of $\mathbb{P}\left(X \in \mathscr{E}\right)$. We next show that the bound in (35) always hold on the complementary set $\mathscr{E}^c$. If $X \notin \mathscr{E}$, based on the definition of $\mathscr{E}$, we have $\inf_{\beta \in S^{p-1}(\Theta)} \frac{|X\beta\|_2}{\sqrt{n}} \geqslant \frac{1}{4} - 2g\left(\phi(\beta)\right)$. That is $\forall \beta \in S^{p-1}(\Theta)$. $\frac{\|X\beta\|_2}{\sqrt{n}} \geqslant \frac{1}{4} - 2g\left(\phi(\beta)\right)$. Therefore, for any $\beta' \in \{\beta' \in \mathbb{R} | \frac{\beta'}{\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2} \in S^{p-1}(\Theta)\}$, we have

$$\frac{\left\|X\frac{\beta'}{\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2}\right\|_2}{\sqrt{n}} \geqslant \frac{1}{4} - 2g\Big(\phi\Big(\frac{\beta'}{\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2}\Big)\Big)$$

$$\frac{\left\|X\beta'\right\|_2}{\sqrt{n}} \geqslant \frac{1}{4}\left\|\Theta^{\frac{1}{2}}\beta'\right\|_2 - 2g\Big(\phi(\beta')\Big),$$

where we finish the proof by substituting the definition of $g(\phi(\beta))$. ■

Proof of Lemma 15 By definition, $\mathbb{K}(0, \upsilon) \cup \left(\bigcup_{\ell=1}^{\infty} \mathbb{K}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)\right)$ is a cover of $S^{p-1}(\Theta)$. Therefore, for any β, it either belongs to $\mathbb{K}(0, \upsilon)$ or $\mathbb{K}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)$, which leads to the following two cases:

Case 1 If $\beta \in \mathbb{K}(0, \upsilon)$, by definition, we have $g\left(\phi(\beta)\right) \in [0, \upsilon]$ and

$$\frac{\|X\beta\|_2}{\sqrt{n}} \leqslant \frac{1}{4} - 2g\left(\phi(\beta)\right) \leqslant \frac{1}{4} = \frac{1}{2} - \upsilon.$$

Therefore, the event $\mathscr{A}(0, \upsilon)$ must happen in this case.

Case 2: If $\beta \notin \mathbb{K}(0, \upsilon)$, we must have $\beta \in \mathbb{K}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)$ for some $\ell = 1, 2, \cdots$, and moreover

$$\frac{\|X\beta\|_2}{\sqrt{n}} \leqslant \frac{1}{4} - 2g\left(\phi(\beta)\right) \leqslant \frac{1}{4} - 2 \cdot \left(2^{\ell-1}\upsilon\right) \leqslant \frac{1}{2} - \left(2 \cdot 2^{\ell-1}\right)\upsilon \leqslant \frac{1}{2} - 2^{\ell}\upsilon.$$

So that the event $\mathscr{A}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)$ must happen. Therefore, $\mathscr{E} \subseteq \mathscr{A}(0, \upsilon) \cup \left(\bigcup_{\ell=1}^{\infty} \mathscr{A}\left(2^{\ell-1}\upsilon, 2^{\ell}\upsilon\right)\right)$.
■

Proof of Lemma 16 To prove Lemma 16, we first introduce the following lemmas:

Lemma 17 (Gordon's Inequality). *Let $\{Z_{u,v}\}_{u \in U, v \in V}$ and $\{Y_{u,v}\}_{u \in U, v \in V}$ be zero-mean Gaussian process indexed by a non-empty index set $I = U \times V$. If*

1. $\mathbb{E}\left((Z_{u,v} - Z_{u'v'})^2\right) \leq \mathbb{E}\left(\left(Y_{u,v} - Y_{u',v'}\right)^2\right)$ *for all pairs* (u, v) and $(u'\, v') \in I$.

2. $\mathbb{E}\left((Z_{u,v} - Z_{u'v})^2\right) = \mathbb{E}\left(\left(Y_{u,v} - Y_{u',v}\right)^2\right)$,

then we have $\mathbb{E}(\max_{v \in V} \min_{u \in U} Z_{u,v}) \leq \mathbb{E}(\max_{v \in V} \min_{u \in U} Y_{u,v})$.

Lemma 18. *Suppose that $\alpha = (\alpha_1, ..., \alpha_d)$, where each $\alpha_i, i \in [d]$ is a zero-mean sub-Gaussian random variable with parameter at most σ^2, then for any $t \in \mathbb{R}$, we have $\mathbb{E}\left(\exp\left(t\,\|\alpha\|_2\right)\right) \leqslant 5^d \exp\left(2t^2\sigma^2\right)$.*

Lemma 19. *Suppose that $\alpha = (\alpha_1, ..., \alpha_d)$, where each $\alpha_i, i \in [d]$ is a zero-mean sub-Gaussian random variable with parameter at most σ^2, and for a given group structure G, let $\left\|\alpha_{G_g}\right\|$ be the corresponding group norm, m be the number of groups and d_{max} be the maximum group size, then*

$$\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right) \leqslant 2\sqrt{2\sigma^2\left(\log m + d_{max}\log 5\right)}.$$

Lemma 20 (Theorem 2.26 in (Wainwright, 2019))**.** *Let $x = (x_1, \cdots, x_n)$ be a vector of i.i.d standard Gaussian variable, and $f : \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz, with respect to the Euclidean norm, then $f(x) - \mathbb{E}f(x)$ is sub-Gaussian with parameter at most L, and hence $\mathbb{P}\left(\left(f(x) - \mathbb{E}\left[f(x)\right]\right)\right) \geqslant t] \leqslant e^{-\frac{t^2}{2L^2}}, \forall t \geqslant 0$.*

We now start to prove. First, we define and bound the random variable $T\left(r_\ell, r_u\right) = -\inf_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\|X\beta\|_2}{\sqrt{n}}$. Let S^{n-1} be a unit ball on $\mathbb{R}^n$, by the variational representation of the ℓ_2-norm, we have

$$T\left(r_\ell, r_u\right) = -\inf_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\|X\beta\|_2}{\sqrt{n}} = -\inf_{\beta\in\mathbb{K}(r_\ell,r_u)} \sup_{u\in S^{n-1}} \frac{\langle u, X\beta\rangle}{\sqrt{n}} = \sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \inf_{u\in S^{n-1}} \frac{\langle u, X\beta\rangle}{\sqrt{n}}.$$

Let $X = W\Theta^{\frac{1}{2}}$, where $W \in \mathbb{R}^{n\times p}$ is a standard Gaussian matrix, and define the transformed vector $v = \Theta^{\frac{1}{2}}\beta$, then

$$T\left(r_\ell, r_u\right) = \sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \inf_{u\in S^{n-1}} \frac{\langle u, X\beta\rangle}{\sqrt{n}} = \sup_{v\in\bar{\mathbb{K}}(r_\ell,r_u)} \inf_{u\in S^{n-1}} \frac{\langle u, Wv\rangle}{\sqrt{n}},$$

where $\bar{\mathbb{K}}\left(r_\ell, r_u\right) = \left\{v \in \mathbb{R}^p \middle| \|v\|_2 = 1, g\left(\phi(\Theta^{-\frac{1}{2}}v)\right) \in \left[r_\ell, r_u\right]\right\}$.

Define $Z_{u,v} = \frac{\langle u,Wv\rangle}{\sqrt{n}}$, since (u, v) range over a subset of $S^{n-1} \times S^{p-1}$, each variable $Z_{u,v}$ is zero-mean Gaussian with variance n^{-1}. We compare the Gaussian process $Z_{u,v}$ to the zero-mean Gaussian process $Y_{u,v}$ which defined as:

$$Y_{u,v} = \frac{\langle\zeta, u\rangle}{\sqrt{n}} + \frac{\langle\xi, v\rangle}{\sqrt{n}} \qquad \text{where } \zeta \in \mathbb{R}^n, \xi \in \mathbb{R}^p, \text{have i.i.d } N(0,1) \text{ entries.}$$

Next, we show that the $Y_{u,v}$ and $Z_{u,v}$ defined above satisfy conditions in Gordon's inequality. By definition, we have

$$\begin{aligned}
\mathbb{E}\left(Z_{u,v} - Z_{u',v'}\right)^2 = \mathbb{E}\left(\frac{\langle u, Wv\rangle}{\sqrt{n}} - \frac{\langle u', Wv'\rangle}{\sqrt{n}}\right)^2 &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p \left(u_iv_j - u_i'v_j'\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p \left(u_iv_j - u_i'v_j + u_i'v_j - u_i'v_j'\right)^2 \\
&= \frac{1}{n}\left(\|v\|_2^2\left\|u - u'\right\|_2^2 + \left\|u'\right\|_2^2\left\|v - v'\right\|_2^2 + 2\left(\|v\|_2^2 - \langle v, v'\rangle\right)\left(\langle u, u'\rangle - \|u\|_2^2\right)\right).
\end{aligned} \tag{36}$$

Since $\|v\|_2^2 \leqslant 1$, $\|u'\|_2^2 \leqslant 1$, we have $\mathbb{E}\left(Z_{u,v} - Z_{u',v'}\right)^2 \leqslant \frac{1}{n}\left(\|u-u'\|_2^2 + \|v-v'\|_2^2\right)$. On the other hand, we have

$$\begin{aligned}\mathbb{E}\left(Y_{u,v} - Y_{u',v'}\right)^2 &= \mathbb{E}\left(\frac{\langle \zeta, u-u'\rangle}{\sqrt{n}} + \frac{\langle \xi, v-v'\rangle}{\sqrt{n}}\right)^2 \\ &= \frac{1}{n}\Bigg(\sum_{i=1}^{n}\sum_{j=1}^{p}(u-u')^2 + \sum_{i=1}^{n}\sum_{j=1}^{p}(v-v')^2\Bigg) = \frac{1}{n}\Bigg(\left\|u-u'\right\|_2^2 + \left\|v-v'\right\|_2^2\Bigg).\end{aligned} \tag{37}$$

Taking equation (36) and (37) together, we have

$$\mathbb{E}\left(Z_{u,v} - Z_{u',v'}\right)^2 \leqslant \frac{1}{n}\left(\left\|u-u'\right\|_2^2 + \left\|v-v'\right\|_2^2\right) = \mathbb{E}\left(Y_{u,v} - Y_{u',v'}\right)^2.$$

If $V = V'$, then $n\mathbb{E}\left(\left(Z_{u,v} - Z_{u',v'}\right)^2\right) = \|u-u'\|_2 = n\mathbb{E}\left(\left(Y_{u,v} - Y_{u',v'}\right)^2\right)$. By applying Lemma 17, we have

$$\mathbb{E}\Bigg(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}} Z_{u,v}\Bigg) \leqslant \mathbb{E}\Bigg(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}} Y_{u,v}\Bigg).$$

Therefore,

$$\begin{aligned}\mathbb{E}\Bigg(T\left(r_\ell, r_u\right)\Bigg) &= \mathbb{E}\Bigg(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}} \frac{\langle u, Wv\rangle}{\sqrt{n}}\Bigg) \leqslant \mathbb{E}\Bigg(\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)}\inf_{u\in S^{n-1}}\left(\frac{\langle \xi, v\rangle}{\sqrt{n}} + \frac{\langle \zeta, u\rangle}{\sqrt{n}}\right)\Bigg) \\ &= \mathbb{E}\Bigg(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\left\langle \Sigma^{\frac{1}{2}}\xi, \beta\right\rangle}{\sqrt{n}}\Bigg) - \mathbb{E}\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right).\end{aligned}$$

Next, we bound these two terms. For the second term, we have $\mathbb{E}\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right) = \mathbb{E}\left(\sqrt{\frac{\xi_1^2+\ldots+\xi_n^2}{n}}\right) \geqslant \mathbb{E}\left(\frac{|\xi_1|+\ldots+|\xi_n|}{n}\right) = \sqrt{\frac{2}{\pi}}$. For the first term, we have $\mathbb{E}\Bigg(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\left\langle \Theta^{\frac{1}{2}}\xi, \beta\right\rangle}{\sqrt{n}}\Bigg) \leqslant \mathbb{E}\Bigg(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\phi(\beta)\phi^*(\Theta^{\frac{1}{2}}\xi)}{\sqrt{n}}\Bigg)$, where $\phi^*(\Theta^{\frac{1}{2}}\xi)$ is the the dual norm defined before. Since $\beta \in \mathbb{K}\left(r_\ell, r_u\right)$, $g\left(\phi(\beta)\right) \leqslant r_u$, by the definition of $g(t)$, we have

$$\phi(\beta) \leqslant \frac{r_u}{\left(4\rho(\Theta)\max_{g\in[m]}\frac{1}{w_g\sqrt{h_{\min}^g}}\sqrt{\frac{2(\log m + d_{\max}\log 5)}{n}}\right)}. \tag{38}$$

Let $\eta_{G_g} = (\Theta^{\frac{1}{2}}\xi)_{G_g}$, to bound $\mathbb{E}\left(\max_g \left\|(\Theta^{\frac{1}{2}}\xi)_{G_g}\right\|_2\right) = \mathbb{E}\left(\max_g \left\|\eta_{G_g}\right\|_2\right)$. Since $\Theta^{\frac{1}{2}}\xi \sim N(0,\Theta)$, by the properties of normal distribution, its corresponding marginal distribution of jth variable $(\Theta^{\frac{1}{2}}\xi)_j$ also follows zero mean normal distribution with covariance matrix Θ_{jj}, which is the jth diagonal elements of Θ. Therefore, any subset of $\Theta^{\frac{1}{2}}\xi$ is a zero-mean sub-Gaussian random sequence with parameters at most $\rho(\Theta)$. By (38) and Lemma 19, we have

$$\begin{aligned}
\mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{\phi(\beta)\phi^*\Theta^{\frac{1}{2}}\xi}{\sqrt{n}}\right) &\leqslant \mathbb{E}\left(\sup_{\beta\in\mathbb{K}(r_\ell,r_u)} \frac{r_u}{\left(4\rho(\Theta)\big(\max\limits_{g\in[m]} \frac{1}{w_g h_{\min}^g}\big)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)} \frac{\phi^*\left(\Theta^{\frac{1}{2}}\xi\right)}{\sqrt{n}}\right) \\
&= \frac{r_u}{\left(4\rho(\Theta)\big(\max\limits_{g\in[m]} \frac{1}{w_g h_{\min}^g}\big)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)} \mathbb{E}\left(\frac{\phi^*\left(\Theta^{\frac{1}{2}}\xi\right)}{\sqrt{n}}\right) \\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\big(\max\limits_{g\in[m]} \frac{1}{w_g h_{\min}^g}\big)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)} \mathbb{E}\left(\max_{g\in[m]} \frac{1}{\sqrt{n}w_g}\left\|H\left(\Theta^{\frac{1}{2}}\xi\right)_{G_g}\right\|_2\right) \\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\big(\max\limits_{g\in[m]} \frac{1}{w_g h_{\min}^g}\big)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)} \mathbb{E}\left(\max_{g\in[m]} \frac{1}{\sqrt{n}w_g h_{\min}^g}\left\|\left(\Theta^{\frac{1}{2}}\xi\right)_{G_g}\right\|_2\right) \\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)} \mathbb{E}\left(\left\|\max_{g\in[m]}\left(\Theta^{\frac{1}{2}}\xi\right)_{G_g}\right\|_2\right) \\
&\leqslant \frac{r_u}{\left(4\rho(\Theta)\sqrt{\frac{2(\log m+d_{\max}\log 5)}{n}}\right)} \left(2\rho(\Theta)\sqrt{(\log m + d_{\max}\log 5)\,2\sigma^2}\right) \leqslant \frac{r_u}{2}.
\end{aligned}$$

Therefore, $\mathbb{E}\left[T\left(r_\ell, r_u\right)\right] \leqslant -\sqrt{\frac{2}{\pi}} + \frac{r_u}{2}$. Next we want to bound $\mathbb{P}\left(T\left(r_\ell, r_u\right) \geqslant -\frac{1}{2} + r_u\right)$ based on the bound of this expectation. To apply Lemma 20, we first show that, the $f = T(r_l, r_u)$, a function of the random variable W is a $\frac{1}{\sqrt{n}}$-Lipschitz function and without making confusion, we denote the corresponding function as $T(W)$. For any standard Gaussian matrix W_1 and W_2, we have

$$\begin{aligned}
\Big|T(W_1) - T(W_2)\Big| &= \Big|\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \inf_{u\in S^{n-1}} \frac{\Big\langle u, W_1 v\Big\rangle}{\sqrt{n}} - \sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \inf_{u\in S^{n-1}} \frac{\Big\langle u, W_2 v\Big\rangle}{\sqrt{n}}\Big| \\
&= \Big|\sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \Big(-\frac{\big\|W_1 v\big\|_2}{\sqrt{n}}\Big) - \sup_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \Big(-\frac{\big\|W_2 v\big\|_2}{\sqrt{n}}\Big)\Big| \\
&= \Big|\Big(-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \frac{\big\|W_1 v\big\|_2}{\sqrt{n}}\Big) - \Big(-\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \frac{\big\|W_2 v\big\|_2}{\sqrt{n}}\Big)\Big| \\
&= \Big|\inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \frac{\big\|W_2 v\big\|_2}{\sqrt{n}} - \inf_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \frac{\big\|W_1 v\big\|_2}{\sqrt{n}}\Big|.
\end{aligned}$$

Suppose that $\frac{\|W_1 v_1\|_2}{\sqrt{n}} = \inf\limits_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}}$ and $\frac{\|W_2 v_2\|_2}{\sqrt{n}} = \inf\limits_{v\in\tilde{\mathbb{K}}(r_\ell,r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}}$.

- **Case I** If $\|W_1 v_1\|_2 > \|W_2 v_2\|_2$, then we have

$$\begin{aligned}|T(W_1) - T(W_2)| &= \left| \inf_{v \in \tilde{\mathbb{K}}(r_\ell, r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}} - \inf_{v \in \tilde{\mathbb{K}}(r_\ell, r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}} \right| \\ &= \frac{\|W_1 v_1\|_2 - \|W_2 v_2\|_2}{\sqrt{n}} \leqslant \frac{\|W_1 v_2\|_2 - \|W_2 v_2\|_2}{\sqrt{n}} \\ &\leqslant \frac{\|(W_1 - W_2) v_2\|_2}{\sqrt{n}} \leqslant \frac{\|W_1 - W_2\|_F}{\sqrt{n}}\end{aligned}$$

.

- **Case II** If $\|W_1 v_1\|_2 \leqslant \|W_2 v_2\|_2$, then we have

$$\begin{aligned}|T(W_1) - T(W_2)| &= \left| \inf_{v \in \tilde{\mathbb{K}}(r_\ell, r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}} - \inf_{v \in \tilde{\mathbb{K}}(r_\ell, r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}} \right| \\ &= \frac{\|W_2 v_2\|_2 - \|W_1 v_1\|_2}{\sqrt{n}} \leqslant \frac{\|W_2 v_1\|_2 - \|W_1 v_1\|_2}{\sqrt{n}} \\ &\leqslant \frac{\|(W_1 - W_2) v_1\|_2}{\sqrt{n}} \leqslant \frac{\|W_1 - W_2\|_F}{\sqrt{n}},\end{aligned}$$

where $\|.\|_F$ represent the Frobenious norm of a matrix.

Thus under the Euclidean norm, $T(W)$ is a $\frac{1}{\sqrt{n}}$-Lipschitz function. Therefore, by Lemma 19, we have

$$\mathbb{P}(T(r_l, r_u) - \mathbb{E}(T(r_l, r_u)) \geqslant t) \leqslant e^{-nt^2/2}, \forall t \geqslant 0.$$

Set $t = \sqrt{\frac{2}{\pi}} - \frac{1}{2} + \frac{r_u}{2} \geqslant \frac{1}{4} + \frac{r_u}{2}$, we have $\mathbb{E}(T(r_l, r_u)) + t \leqslant -\frac{1}{2} + r_u$ and $\mathbb{P}\left[T\left(r_\ell, r_u\right) \geqslant -\frac{1}{2} + r_u\right] \leqslant e^{-\frac{n}{32}} e^{-\frac{n}{8} r_u^2}$, which is actually the Lemma 16. ■

Proof of Lemma 18 We can find a $\frac{1}{2}$ - cover of S^{d-1}, and for any $u \in S^{d-1}$ in the Euclidean norm with cardinally at most $N \leqslant 5^d$. Suppose that there exists $u^{q(u)} \in \left\{u^1, \ldots, u^N\right\}$, such that $\left\|u^{q(u)} - u\right\|_2 \leqslant \frac{1}{2}$. By the variational representation of the ℓ_2 norm, we have

$$\|\alpha\|_2 = \max_{u \in S^{d-1}} \langle u, \alpha \rangle \leqslant \max_{q(u) \in [N]} \left\langle u^{q(u)}, \alpha \right\rangle + \frac{1}{2} \|\alpha\|_2 .$$

Therefore, $\|\alpha\|_2 \leqslant 2 \max\limits_{q(u) \in [N]} \left\langle u^{q(u)}, \alpha \right\rangle$. Consequently,

$$\begin{aligned}\mathbb{E}\left(\exp\left(t \|\alpha\|_2\right)\right) &\leqslant \mathbb{E}\left(\exp\left(2t \max_{q \in [N]} \langle u^q, \alpha \rangle\right)\right) = \mathbb{E}\left(\max_{q \in [N]} \exp\left(2t \langle u^q, \alpha \rangle\right)\right) \\ &\leqslant \sum_{q=1}^{N} \mathbb{E}\left(\exp\left(2t \langle u^q, \alpha \rangle\right)\right) \leqslant 5^d \exp\left(\frac{4t^2\sigma^2}{2}\right) \leqslant 5^d \exp\left(2t^2\sigma^2\right).\end{aligned}$$

■

Proof of Lemma 19 For any $t > 0$, by Jensen's inequality, we have

$$\begin{aligned}\exp\left(t\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right)\right) &\leqslant \mathbb{E}\left(\exp\left(t\max_g \left\|\alpha_{G_g}\right\|_2\right)\right) = \mathbb{E}\left(\max_j \exp\left(t\left\|\alpha_{G_g}\right\|_2\right)\right)\\ &\leqslant \sum_{j=1}^{m} \mathbb{E}\left(\exp\left(t\left\|\alpha_{G_g}\right\|_2\right)\right) \leqslant \sum_{j=1}^{m} 5^{d_g}\exp\left(2t^2\sigma^2\right) \leqslant m\cdot 5^{d_{\max}}\cdot\exp(2t^2\sigma^2).\end{aligned}$$

By taking log at both sides, we have $t\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right) \leqslant \log m + d_{\max}\log 5 + 2t^2\sigma^2$. Consequently, let $t = \sqrt{\frac{\log m + d_{\max}\log 5}{2\sigma^2}}$, we have $\mathbb{E}\left(\max_g \left\|\alpha_{G_g}\right\|\right) \leqslant 2\sqrt{(\log m + d_{\max}\log 5)\,2\sigma^2}$. ■

D.7 Proof of Theorem 4

The two lemmas below are integral to the proof:

Lemma 21 (Packing Number for Binary Sets)**.** *Consider a set A defined for real numbers m, s_g as*

$$A = \left\{a \in \{0,1\}^m \mid \sum_{j=1}^{m} a_j \leq s_g\right\}.$$

Then the $\sqrt{\frac{s_g}{2}}$-packing number of set $A \geqslant \frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}\cdot 2^{\frac{s_g}{2}}}$, and

$$\log\left(\frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}\cdot 2^{\frac{s_g}{2}}}\right) \asymp s_g\log(\frac{m}{s_g}).$$

Lemma 22 (Packing Number for Sparse Group Vectors)**.** *For the set $\Omega(G, s_g)$, the $\sqrt{\frac{2ds_g}{5}}$-packing number $\gtrsim \frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}\cdot 2^{\frac{s_g}{2}}}\cdot(\sqrt{2})^{ds_g}$, and*

$$\log\left(\frac{\binom{m}{s_g}-2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}\cdot 2^{\frac{s_g}{2}}}\cdot(\sqrt{2})^{ds_g}\right) \asymp s_g(d + \log(\frac{m}{s_g})).$$

Proof of Theorem 4 First, select N points $\omega^{(1)}, \dots, \omega^{(N)}$ from $\Omega(G, s_g)$ such that $\left\|\omega^{(i)} - \omega^{(j)}\right\| > \sqrt{\frac{2ds_g}{5}}$ for all distinct i, j. Clearly, $\left\|\omega^{(i)} - \omega^{(j)}\right\| \leqslant \sqrt{4s_g d}$. Define $\beta^{(i)} = r\omega^{(i)}$ for each i. This results in

$$\frac{2ks_g r^2}{5} \leq \left\|\beta^{(i)} - \beta^{(j)}\right\|_2^2 \leqslant 4s_g dr^2.$$

Next, let $y^{(i)} = X\beta^{(i)} + \varepsilon$ for $1 \leqslant i \leqslant N$. Consider the Kullback-Leibler divergence between different distribution pairs:

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

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

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

Thus, for $1 \leq i \neq j \leq N$,

$$D_{KL}\left(\left(y^{(i)}, X\right), \left(y^{(j)}, X\right)\right) = \mathbb{E}_X \frac{\left\|X\left(\beta^{(i)} - \beta^{(j)}\right)\right\|_2^2}{2\sigma^2} = \frac{n(\beta^{(i)} - \beta^{(j)})^\top \Sigma(\beta^{(i)} - \beta^{(j)})}{2\sigma^2}$$
$$\leq \frac{3c_1\left\|\beta^{(i)} - \beta^{(j)}\right\|_2^2}{2\sigma^2} \leq \frac{2c_1 ndr^2 s_g}{\sigma^2}.$$

From Lemma 22, $\log N \asymp s_g\left(d + \log\frac{m}{s_g}\right)$. Setting $\frac{\frac{ndr^2 s_g}{\sigma^2} + \log 2}{\log N} = \frac{1}{2}$, we obtain

$$r \gtrsim \sqrt{\frac{\left(d + \log\frac{m}{s_g}\right)\sigma^2}{3nd}}.$$

By generalized Fano's Lemma, $\inf\limits_{\hat\beta}\sup\limits_{\beta} \mathbb{E}\|\hat\beta - \beta\|_2 \geqslant \sqrt{\frac{2r^2 k s_g}{5}}\left(1 - \frac{\frac{ndr^2 s_g}{\sigma^2} + \log 2}{\log N}\right)$. Consequently,

$$\inf\sup \mathbb{E}\|\hat\beta - \beta\|_2^2 \geq \left(\inf\sup \mathbb{E}\|\hat\beta - \beta\|_2\right)^2 \gtrsim \frac{\sigma^2\left(s_g(d + \log(\frac{m}{s_g}))\right)}{n}.$$

■

Proof of Lemma 21 Notice that the cardinality of A is $\binom{m}{s_g}$. Denote the hamming distance between any two points $x, y \in A$ by

$$h(a, b) = |\{j : a_j \neq b_j\}|.$$

Then, for a fixed point a $\in A$,

$$\left|\left\{b \in A, h(a, b) \leq \frac{s_g}{2}\right\} = \binom{m}{\lfloor\frac{s_g}{2}\rfloor} \cdot 2^{\lfloor\frac{s_g}{2}\rfloor}\right|.$$

In fact, all elements $b \in A$ with $h(a, b) \leq \frac{s_g}{2}$ can be obtained as follows. First, take any subset $J \subset [m]$ of cardinality $\lfloor\frac{s_g}{2}\rfloor$, then set $a_j = b_j$ for $j \notin J$ and choose $b_j \in \{0, 1\}$ for $j \in J$.

Now let A_s be any subset of A with cardinality at most $T = \frac{\binom{m}{s_g} - 2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor} \cdot 2^{\frac{s_g}{2}}}$, then we have

$$|\left\{b \in A \mid \text{there exist } \ a \in A_s \text{ with } h(a, b) \leq \frac{s_g}{2}\right\} \leq (|A_s|) \cdot \binom{m}{\lfloor\frac{s_g}{2}\rfloor} \cdot 2^{\frac{s_g}{2}}| < |A|.$$

It implies that one can find an element $b \in A$ with $h(a, b) > \frac{s_g}{2}$ for all $a \in A_s$. Therefore one can construct a subset A_s with $|A_s| \geq T$ and the property $h(a, b) > \frac{s_g}{2}$ for any two distinct elements $a, b \in A_s$.

On the other hand, $h(a,b) > \frac{s_g}{2}$ implies $\|a-b\| > \sqrt{\frac{s_g}{2}}$. Therefore, there exist at least T points in A such that the distance between any two points is greater than $\sqrt{\frac{s_g}{2}}$.

Moreover, since $\frac{\binom{m}{s_g}}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}} = \frac{\lfloor\frac{s_g}{2}\rfloor!(m-\lfloor\frac{s_g}{2}\rfloor)!}{s_g!(m-s_g)!} = \frac{(m-s_g+1)\cdots m-\lfloor\frac{s_g}{2}\rfloor}{(\lfloor\frac{s_g}{2}\rfloor+1)\cdots s_g} = \prod_{j=1}^{\lceil\frac{s_g}{2}\rceil}\frac{m-s_g+j}{\lfloor\frac{s_g}{2}\rfloor+j}$, we have

$$\left(\frac{m-\lfloor\frac{s_g}{2}\rfloor}{2s_g}\right)^{\lfloor\frac{s_g}{2}\rfloor} \leqslant \frac{\binom{m}{s_g}}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}2^{\frac{s_g}{2}}} \leqslant \left(\frac{m-s_g+1}{\lceil s_g\rceil}\right)^{\lceil\frac{s_g}{2}\rceil},$$

and therefore we can find C_1, C_2, such that $C_1 s_g\log(\frac{m}{s_g}) \leqslant \log T \leqslant C_2 s_g\log(\frac{m}{s_g})$, so that

$$\log\left(\frac{\binom{m}{s_g}-\ 2}{\binom{m}{\lfloor\frac{s_g}{2}\rfloor}\cdot 2^{\frac{s_g}{2}}}\right) \asymp s_g\log(\frac{m}{s_g}).$$

■

Proof of Lemma 22 Given a group support $a \in A$, define $k_a = \left|\left\{i \mid i \in (\bigcup_{\{g|a_g=0\}} G_g)^c\right\}\right|$, and the set

$$\Omega^{(a)} = \left\{\omega \in \mathbb{R}^p \mid \omega_i = 0 \text{ if } i \in \bigcup_{\{g|a_g=0\}} G_g, \omega_i \in \{-1,1\} \text{ if } i \in \Big(\bigcup_{\{g|a_g=0\}} G_g\Big)^c\right\}.$$

Notice that $\Omega^{(a)} \subseteq \Omega(G, s_g)$, and $|\Omega^{(a)}| = 2^{k_a}$. Also denote the hamming distance between $x, y \in \Omega^{(a)}$ by

$$h(x,y) = |\{j : x_j \neq y_j\}|.$$

Then for any fixed $x \in \Omega_G^{(a)}$, we have

$$\left|\{y \in \Omega^{(a)}, h(x,y) \leq \frac{k_a}{10}\}\right| = \sum_{j=0}^{\lfloor\frac{k_a}{10}\rfloor}\binom{k_a}{j}.$$

Let $\Omega_s^{(a)}$ be any subset of $\Omega^{(a)}$ with cardinality at most $N^{(a)} = \frac{2^{k_a}-2}{\sum_{j=0}^{\lfloor\frac{k_a}{10}\rfloor}\binom{k_a}{j}}$. Then,

$$\left|\{y \in \Omega^{(a)} \mid \exists x \in \Omega_s^{(a)} \text{ with } h(x,y) \leq \frac{k_a}{10}\}\right| < |\Omega^{(a)}|.$$

On the other hand, $h(x,y) > \frac{k_a}{10}$ implies $\|x-y\| \geq \sqrt{\frac{2k_a}{5}}$. Thus, there are at least $N^{(a)}$ points in $\Omega^{(a)}$ with pairwise distances greater than $\sqrt{\frac{2k_a}{5}}$. From the results in Graham et al. (1994, Chap. 9),

$$\sum_{j\leq\lfloor\frac{k_a}{10}\rfloor}\binom{k_a}{j} < \frac{9}{8}\binom{k_a}{\lfloor\frac{k_a}{10}\rfloor} \leq \frac{9}{8}(10e)^{\frac{k_a}{10}} \leq \frac{9}{8}2^{\frac{k_a}{2}}.$$

Consequently, we have $N^{(a)} > \frac{8}{9} 2^{\frac{k_a}{2}} \gtrsim (\sqrt{2})^{k_a}$.

The value of k_a depends on the predefined groups and group support a and spans a range from 0 to $s_g d$. Lemma 22 seeks a lower bound for all conceivable overlapping patterns, necessitating an analysis of the maximum value of k_a.

Furthermore, according to Lemma 21, we can identify at least T points in A where the distance between any two points exceeds $\sqrt{\frac{s_g}{2}}$. For $\{a_1, \cdots, a_T\}$ group supports, if there is a group structure such that we could find at least $\frac{8}{9}(\sqrt{2})^{s_g d}$ on each group support, and the distance between every pair of these points is greater than $\sqrt{\frac{2 s_g d}{5}}$, then Lemma 22 is proved.

Considering m non-overlapping groups, $k_a = s_g d$ for each group support a. In addition, given any two group support a, b with $\|a - b\| > \sqrt{\frac{s_g}{2}}$, $\|x - y\| > \sqrt{\frac{d s_g}{2}} > \sqrt{\frac{2 d s_g}{5}}$ for any $x \in \Omega^{(a)}$ and $y \in \Omega^{(b)}$. Thus, considering all possible overlapping patterns, we can find at least $\frac{\binom{m}{s_g} - 2}{\binom{m}{\lfloor \frac{s_g}{2} \rfloor} \cdot 2^{\frac{s_g}{2}}} \cdot \frac{8}{9}(\sqrt{2})^{d s_g}$ point in $\Omega(G, s_g)$, such that the distance between every pair of points is greater than $\sqrt{\frac{2 d s_g}{5}}$.

■

D.8 Proof of Theorem 5

This proof consists of parts: Parts I-IV dedicated to Theorem 5.1, and Part V is for Theorem 5.2. To be more specific, Part I provides some additional concepts, Part II introduces the reduced problem, Part III shows the successful selection of the correct pattern under favorable conditions, and Part IV establishes that certain conditions are satisfied with high probability.

D.8.1 PART I

Recall that $\mathbf{S} = \mathit{supp}(\beta^*)$. With $\mathbf{S}$, we define the norm $\phi_{\mathbf{S}}$ for any $\beta \in \mathbb{R}^p$ as

$$\phi_{\mathbf{S}}(\beta_{\mathbf{S}}) = \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S} \cap G_g}\|_2,$$

along with its dual norm $(\phi_{\mathbf{S}})^*[u] = \sup_{\phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leq 1} \beta_{\mathbf{S}}^\top u$. Similarly, for $\mathbf{S}^c = [p] \setminus \mathbf{S}$, we define the norm $\phi_{\mathbf{S}}^c$ for any $\beta \in \mathbb{R}^p$ as

$$\phi_{\mathbf{S}}^c(\beta_{\mathbf{S}^c}) = \sum_{g \in [m] \setminus \mathsf{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S}^c \cap G_g}\|_2,$$

accompanied by its corresponding dual norm $(\phi_{\mathbf{S}}^c)^*[u] = \sup_{\phi_{\mathbf{S}}^c(\beta_{\mathbf{S}^c}) \leq 1} \beta_{\mathbf{S}^c}^\top u$.

We also introduce equivalence parameters $a_{\mathbf{S}}, A_{\mathbf{S}}, a_{\mathbf{S}^c}, A_{\mathbf{S}^c}$ as follows:

$$\forall \beta \in \mathbb{R}^p,\ a_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1 \leqslant \phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leqslant A_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1, \tag{39}$$

$$\forall \beta \in \mathbb{R}^p,\ a_{\mathbf{S}^c} \|\beta_{\mathbf{S}^c}\|_1 \leqslant \phi_{\mathbf{S}}^c(\beta_{\mathbf{S}^c}) \leqslant A_{\mathbf{S}^c} \|\beta_{\mathbf{S}^c}\|_1. \tag{40}$$

We now study the equivalence parameters from two aspects. First, since

$$\sup_{a_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1 \leqslant 1} \beta_{\mathbf{S}}^\top u \geqslant \sup_{\phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leqslant 1} \beta_{\mathbf{S}}^\top u \geqslant \sup_{A_{\mathbf{S}} \|\beta_{\mathbf{S}}\|_1 \leqslant 1} \beta_{\mathbf{S}}^\top u,$$

by the definition of dual norm, we have

$$\forall u \in \mathbb{R}^{|\mathbf{S}|}, A_{\mathbf{S}}^{-1}\|u\|_\infty \leqslant (\phi_{\mathbf{S}})^*[u] \leqslant a_{\mathbf{S}}^{-1}\|u\|_\infty. \tag{41}$$

Similarly, by order-reversing,

$$\forall u \in \mathbb{R}^{|\mathbf{S}^c|}, A_{\mathbf{S}^c}^{-1}\|u\|_\infty \leqslant (\phi_{\mathbf{S}}^c)^*[u] \leqslant a_{\mathbf{S}^c}^{-1}\|u\|_\infty. \tag{42}$$

Second, by the Cauchy-Schwarz inequality, for any $\beta \in \mathbb{R}^p$ and $g \in \mathsf{G}_{\mathbf{S}}$,

$$\frac{w_g}{\sqrt{d_g}}\|\beta_{\mathbf{S}\cap G_g}\|_1 \leqslant w_g\|\beta_{\mathbf{S}\cap G_g}\|_2 \leqslant \max_{g\in\mathsf{G}_{\mathbf{S}}} w_g\|\beta_{\mathbf{S}\cap G_g}\|_1.$$

Consequently, we have

$$\min_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g}{\sqrt{d_g}}\|\beta_{\mathbf{S}}\|_1 \leqslant \phi_{\mathbf{S}}(\beta_{\mathbf{S}}) \leqslant h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g\in\mathsf{G}_{\mathbf{S}}} w_g\|\beta_{\mathbf{S}}\|_1,$$

Therefore, we can set $a_{\mathbf{S}} = \min_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g}{\sqrt{d_g}}$ and $A_{\mathbf{S}} = h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g\in\mathsf{G}_{\mathbf{S}}} w_g$. With an trivial extension, we can set $a_{\mathbf{S}^c} = \min_{g\in\mathsf{G}_{\mathbf{S^c}}} w_g/\sqrt{d_g}$.

D.8.2 PART II

From the full problem to the reduced problem

Recall that the group lasso estimator in (14) is defined as

$$\hat{\beta}^G = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n\phi^G(\beta). \tag{43}$$

Now we write $\phi^G(\beta) = \phi(\beta)$ and $L(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2$ for ease of notation. Following Jenatton et al. (2011a); Wainwright (2009), we consider the following restricted problem

$$\begin{aligned}\hat{\beta}^R &= \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p,\beta_{\mathbf{S}^c}=0} L(\beta) + \lambda_n\phi(\beta) = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p,\beta_{\mathbf{S}^c}=0} L(\beta) + \lambda_n \sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g \left\|\beta_{\mathbf{S}\cap G_g}\right\|_2 \\ &:= \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p,\beta_{\mathbf{S}^c}=0} L(\beta) + \lambda_n\phi_{\mathbf{S}}(\beta_{\mathbf{S}}).\end{aligned} \tag{44}$$

Let $L_{\mathbf{S}}(\beta_{\mathbf{S}}) = \frac{1}{2n}\|Y - X_{\mathbf{S}}\beta_{\mathbf{S}}\|_2^2$. Due to the restriction of $\hat{\beta}^R$, we can obtain $\hat{\beta}^R$ by first solving the following reduced problem

$$\begin{aligned}\hat{\beta}_{\mathbf{S}} &= \operatorname*{arg\,min}_{\beta_{\mathbf{S}}\in\mathbb{R}^{|\mathbf{S}|}} \frac{1}{2n}\|Y - X_{\mathbf{S}}\beta_{\mathbf{S}}\|_2^2 + \lambda_n \sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g \left\|\beta_{\mathbf{S}\cap G_g}\right\|_2 \\ &= \operatorname*{arg\,min}_{\beta_{\mathbf{S}}\in\mathbb{R}^{|\mathbf{S}|}} L_{\mathbf{S}}(\beta_{\mathbf{S}}) + \lambda_n\phi_{\mathbf{S}}(\beta_{\mathbf{S}}),\end{aligned} \tag{45}$$

and then padding $\hat{\beta}_{\mathbf{S}}$ with zeros on $\mathbf{S}^c$. In addition,

$$\begin{aligned}L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) &= \frac{1}{2n}\|Y - X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\|_2^2 \\ &= \frac{1}{2n}\left(Y^\top Y - 2Y^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} + (X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\right) \\ &= \frac{1}{2n}\left(Y^\top Y - 2(X\beta^* + \epsilon)^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} + (X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\right) \\ &= \frac{1}{2n}\left(Y^\top Y - 2(X_{\mathbf{S}}\beta_{\mathbf{S}}^*)^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} - 2\epsilon^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} + (X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}})^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}}\right),\end{aligned}$$

and consequently,

$$\begin{aligned}\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) &= \frac{1}{n}X_{\mathbf{S}}^\top X_{\mathbf{S}}\hat{\beta}_{\mathbf{S}} - \frac{1}{n}X_{\mathbf{S}}^\top X_{\mathbf{S}}\beta_{\mathbf{S}}^* - \frac{1}{n}\epsilon^\top X_{\mathbf{S}} \\ &:= Q_{\mathbf{SS}}(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}^*) - q_{\mathbf{S}},\end{aligned} \tag{46}$$

where $Q = \frac{1}{n}X^\top X$, $q = \frac{1}{n}\sum\limits_{i=1}^{n}\epsilon_i x_i$.

D.8.3 PART III

Part III mostly follows the proof in Theorem 7 of Jenatton et al. (2011a). Here we aim to show that *supp*$(\hat{\beta}^G) = \mathbf{S}$ under certain conditions.

To begin with, Given $\beta \in \mathbb{R}^p$, we define $J^G(\beta)$ as:

$$J^G(\beta) = [p] \setminus \Big\{ \bigcup_{G_g \cap supp(\beta) = \emptyset} G_g \Big\}.$$

$J^G(\beta)$ is called the adapted hull of the support of β in Jenatton et al. (2011a). For simplicity, we write $J^G(\beta) = J(\beta)$. Notice that by assumption we have

$$J(\beta^*) = [p] \setminus \Big\{ \bigcup_{G_g \cap supp(\beta^*) = \emptyset} G_g \Big\} = \mathbf{S}.$$

Now we consider the reduced problem (45), and we want to show that for all $g \in \mathsf{G}_{\mathbf{S}}$, $\left\|\hat{\beta}_{\mathbf{S}\cap G_g}\right\|_\infty > 0$. That is, no active group is missing.

Lemma 23. *(see Jenatton et al., 2011a, Lemma 14)*

For the loss $L(\beta)$ and norm ϕ in (43), *$\hat{\beta} \in \mathbb{R}^p$ is a solution of*

$$\min_{\beta \in \mathbb{R}^p} L(\beta) + \lambda_n \phi(\beta) \tag{47}$$

if and only if

$$\begin{cases} \nabla L(\hat{\beta})_{J(\hat{\beta})} + \lambda_n r(\hat{\beta})_{J(\hat{\beta})} = \mathbf{0} \\ (\phi^c_{J(\hat{\beta})})^* \left[\nabla L(\hat{\beta})_{J(\hat{\beta})^c}\right] \leqslant \lambda_n. \end{cases} \tag{48}$$

In addition, the solution $\hat{\beta}$ satisfies

$$\phi^*[\nabla L(\hat{\beta})] \leqslant \lambda_n. \tag{49}$$

As $\hat{\beta}_{\mathbf{S}}$ is the solution of (45), Equation (49) in Lemma 23 implies that

$$(\phi_{\mathbf{S}})^* \left[\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})\right] \overset{(46)}{=} (\phi_{\mathbf{S}})^* \left[Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right] \leqslant \lambda_n. \tag{50}$$

By the property of the equivalent parameters, we have

$$A_{\mathbf{S}}^{-1}\left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right\|_\infty \overset{(41)}{\leqslant} (\phi_{\mathbf{S}})^* \left[Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right] \overset{(50)}{\leqslant} \lambda_n. \tag{51}$$

If

$$\lambda_n \leqslant \frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3|\mathbf{S}|^{\frac{1}{2}}A_{\mathbf{S}}}, \tag{52}$$

and

$$\|q_{\mathbf{S}}\|_\infty \leqslant \frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3|\mathbf{S}|^{\frac{1}{2}}}, \tag{53}$$

then we have

$$\begin{aligned}
\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty &= \left\|Q^{-1}_{\mathbf{SS}} Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right)\right\|_\infty \\
&\leqslant \|Q^{-1}_{\mathbf{SS}}\|_{\infty,\infty}\left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right)\right\|_\infty \\
&\leqslant |\mathbf{S}|^{\frac{1}{2}}\gamma_{\max}\left(Q^{-1}_{\mathbf{SS}}\right)\left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right)\right\|_\infty \\
&\leqslant |\mathbf{S}|^{\frac{1}{2}}\gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\left(\left\|Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta_{\mathbf{S}}\right) - q_{\mathbf{S}}\right\|_\infty + \|q_{\mathbf{S}}\|_\infty\right) \\
&\overset{(51)}{\leqslant} |\mathbf{S}|^{\frac{1}{2}}\gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\left(\lambda_n A_{\mathbf{S}} + \|q_{\mathbf{S}}\|_\infty\right) \\
&\leqslant |\mathbf{S}|^{\frac{1}{2}}\gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\lambda_n A_{\mathbf{S}} + |\mathbf{S}|^{\frac{1}{2}}\gamma^{-1}_{\min}\left(Q_{\mathbf{SS}}\right)\|q_{\mathbf{S}}\|_\infty \\
&\leqslant \frac{2}{3}\beta^*_{\min}.
\end{aligned} \tag{54}$$

If there exist a group $g \in \mathsf{G}_{\mathbf{S}}$ such that $\left\|\hat{\beta}_{\mathbf{S}\cap G_g}\right\|_\infty < \frac{\beta^*_{\min}}{3}$, then

$$\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty > \beta^*_{\min} - \frac{\beta^*_{\min}}{3} = \frac{2\beta^*_{\min}}{3}.$$

Thus, (54) implies that for all $g \in \mathsf{G}_{\mathbf{S}}$,

$$\left\|\hat{\beta}_{\mathbf{S}\cap G_g}\right\|_\infty > \frac{\beta^*_{\min}}{3} > 0. \tag{55}$$

Secondly, we want to show that $\hat{\beta}^R$ solves problem (43). As $\hat{\beta}^R$ is obtained by padding $\hat{\beta}_{\mathbf{S}}$ with zeros on $\mathbf{S}^c$,

$$\begin{aligned}
J(\hat{\beta}^R) = [p] \setminus \left\{\bigcup_{G_g \cap supp(\hat{\beta}^R) = \emptyset} G_g\right\} &= [p] \setminus \left\{\bigcup_{G_g \cap supp(\hat{\beta}_{\mathbf{S}}) = \emptyset} G_g\right\} \\
&\overset{(54)}{=} [p] \setminus \left\{\bigcup_{G_g \cap \mathbf{S} = \emptyset} G_g\right\} = \mathbf{S}.
\end{aligned}$$

From Lemma 23 we know that $\hat{\beta}^R$ is the optimal for problem (43) if and only if

$$\nabla L(\hat{\beta}^R)_{\mathbf{S}} + \lambda_n r(\hat{\beta}^R)_{\mathbf{S}} = \mathbf{0}, \tag{56}$$

and

$$(\phi^c_{\mathbf{S}})^*\left[\nabla L(\hat{\beta}^R)_{\mathbf{S}^c}\right] \leqslant \lambda_n. \tag{57}$$

We now verify the condition in (56). Since

$$\begin{aligned} L(\hat{\beta}^R) &= \frac{1}{2n}\|Y - X\hat{\beta}^R\|_2^2 \\ &= \frac{1}{2n}\left(Y^\top Y - 2(X\beta^*)^\top X\hat{\beta}^R - 2\epsilon^\top X\hat{\beta}^R + (X\hat{\beta}^R)^\top X\hat{\beta}^R\right), \end{aligned}$$

we have

$$\begin{aligned} \nabla L(\hat{\beta}^R)_{\mathbf{S}} &= \left[\frac{1}{n}X^\top X\left(\hat{\beta}^R - \beta^*\right) - \frac{1}{n}\epsilon^\top X\right]_{\mathbf{S}} \\ &= \left[Q\left(\hat{\beta}^R - \beta^*\right)\right]_{\mathbf{S}} - q_{\mathbf{S}} = Q_{\mathbf{SS}}\left(\hat{\beta}^R - \beta^*\right)_{\mathbf{S}} - q_{\mathbf{S}} \\ &= Q_{\mathbf{SS}}\left(\hat{\beta}^R_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right) - q_{\mathbf{S}} = Q_{\mathbf{SS}}\left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right) - q_{\mathbf{S}} \\ &= \nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}). \end{aligned} \tag{58}$$

On the other hand, as $\hat{\beta}^R$ is obtained by padding $\hat{\beta}_{\mathbf{S}}$ with zeros on $\mathbf{S}^c$, we have

$$\lambda_n r(\hat{\beta}^R)_{\mathbf{S}} = \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}).$$

Because $\hat{\beta}_{\mathbf{S}}$ is the optimal for problem (45), (48) in Lemma 23 implies that

$$\nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) + \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) \overset{(46)}{=} Q_{\mathbf{SS}}(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}) - q_{\mathbf{S}} + \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) = \mathbf{0}. \tag{59}$$

Thus, (56) holds as

$$\nabla L(\hat{\beta}^R)_{\mathbf{S}} + \lambda_n r_{\mathbf{S}}(\hat{\beta}^R) = \nabla L_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) + \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) \overset{(59)}{=} \mathbf{0}. \tag{60}$$

Now we continue to show (57). Notice that

$$\left(\hat{\beta}^R - \beta^*\right)_{\mathbf{S}} \overset{(58)}{=} \left(\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right) \overset{(59)}{=} Q^{-1}_{\mathbf{SS}}(q_{\mathbf{S}} - \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})). \tag{61}$$

Let $q_{\mathbf{S}^c|\mathbf{S}} = q_{\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}q_{\mathbf{S}}$, we have

$$\begin{aligned} \nabla L(\hat{\beta}^R)_{\mathbf{S}^c} &\overset{(58)}{=} \left(Q(\hat{\beta}^R - \beta^*)\right)_{\mathbf{S}^c} - q_{\mathbf{S}^c} = Q_{\mathbf{S}^c\mathbf{S}}(\hat{\beta}^R - \beta^*)_{\mathbf{S}} - q_{\mathbf{S}^c} \\ &\overset{(61)}{=} Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}\left(q_{\mathbf{S}} - \lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}})\right) - q_{\mathbf{S}^c} \\ &= -Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}\lambda_n r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) + Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}q_{\mathbf{S}} - q_{\mathbf{S}^c} \\ &= -\lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right) - \lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q^{-1}_{\mathbf{SS}}r_{\mathbf{S}}(\beta^*_{\mathbf{S}}) - q_{\mathbf{S}^c|\mathbf{S}}. \end{aligned} \quad . \tag{62}$$

The previous expression leads us to study the difference of $r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})$. We now introduce the following lemma.

Lemma 24. *(see Jenatton et al., 2011a, Lemma 12)*

For any $J \subset [p]$, let u_J and v_J be two nonzero vectors in $\mathbb{R}^{|J|}$, and define the mapping $r_J : \mathbb{R}^{|J|} \mapsto \mathbb{R}^{|J|}$ such that

$$r_J\left(\beta_J\right)_j = \beta_j \sum_{g \in \mathsf{G}_J, G_g \cap j \neq \phi} \frac{\omega_g}{\left\|\beta_{J \cap G_g}\right\|_2}.$$

Then there exists $\xi_J = t_0 u_J + (1-t_0) v_J$ *for some* $t_0 \in (0,1)$*, such that*

$$\|r_J\left(u_J\right) - r_J(v_J)\|_1 \leqslant \|u_J - v_J\|_\infty \left(\sum_{j\in J}\sum_{g\in \mathsf{G}_J} \frac{w_g \mathbb{1}_{\{j\in G_g\}}}{\left\|\xi_{J\cap G_g}\right\|_2} + \sum_{j\in J}\left(\sum_{k\in J}\sum_{g\in \mathsf{G}_J} \frac{|\xi_j||\xi_k| w_g^4 \mathbb{1}_{\{j,k\in G_g\}}}{\left\|\xi_{J\cap G_g}\right\|_2^3}\right)\right).$$

Lemma 24 implies that

$$\left\|r_{\mathbf{S}}(\hat\beta_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_1 \leqslant \left\|\hat\beta_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty \left(\sum_{j\in\mathbf{S}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g \mathbb{1}_{\{j\in G_g\}}}{\left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2} + \sum_{j\in\mathbf{S}}\sum_{k\in\mathbf{S}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{(w_g)^4 \mathbb{1}_{\{j,k\in G_g\}} |\tilde\beta_j||\tilde\beta_k|}{w_g^3 \left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2^3}\right), \tag{63}$$

where $\tilde\beta = t_0\hat\beta_{\mathbf{S}} + (1-t_0)\,\beta^*_{\mathbf{S}}$.

To find an upper bound of the right-hand side. Recall that (54) implies that $\left\|\hat\beta_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty \leqslant \frac{2}{3}\beta^*_{\min}$, so we have

$$\begin{aligned} \left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2 &\geqslant \sqrt{|\mathbf{S}\cap G_g|}\min\{|\tilde\beta|_j \mid \tilde\beta_j \neq 0\} \\ &\geqslant \sqrt{|\mathbf{S}\cap G_g|}(\beta^*_{\min} - t_0\left\|\hat\beta_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty) \\ &\geqslant \sqrt{|\mathbf{S}\cap G_g|}(\beta^*_{\min} - \left\|\hat\beta_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty) \\ &\geqslant \sqrt{|\mathbf{S}\cap G_g|}\frac{\beta^*_{\min}}{3}. \end{aligned}$$

Consequently, the first term could be upper bounded by

$$\sum_{j\in\mathbf{S}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g \mathbb{1}_{\{j\in G_g\}}}{\left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2} = \sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g |\mathbf{S}\cap G_g|}{\left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2} \leqslant \frac{3}{\beta^*_{\min}} \sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g \sqrt{|\mathbf{S}\cap G_g|}.$$

On the other hand, the Cauchy-Schwarz inequality gives

$$\left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_1^2 \leqslant |\mathbf{S}\cap G_g| \left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2^2.$$

Thus, the second term could also be upper bounded by

$$\begin{aligned} \sum_{j\in\mathbf{S}}\sum_{k\in\mathbf{S}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{(w_g)^4 \mathbb{1}_{\{j,k\in G_g\}} |\tilde\beta_j||\tilde\beta_k|}{w_g^3 \left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2^3} &= \sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g^4 \left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_1^2}{w_g^3 \left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2^3} \\ &\leqslant \sum_{g\in\mathsf{G}_{\mathbf{S}}} \frac{w_g |\mathbf{S}\cap G_g|}{\left\|\tilde\beta_{\mathbf{S}\cap G_g}\right\|_2} \\ &\leqslant \frac{3}{\beta^*_{\min}} \sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g \sqrt{|\mathbf{S}\cap G_g|}. \end{aligned}$$

Let $c_2 = \frac{6}{\beta^*_{\min}} \sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g \sqrt{|\mathbf{S}\cap G_g|}$, then (63) implies

$$\left\|r_{\mathbf{S}}(\hat\beta_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_1 \leqslant c_2 \left\|\hat\beta_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty.$$

If

$$\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\|_{2,\infty} \leqslant 3, \tag{64}$$

then we have

$$\begin{aligned}
\left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right\|_\infty &= \left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right\|_\infty \\
&\leqslant \left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\right\|_{\infty,2}\left\|Q_{\mathbf{SS}}^{-\frac{1}{2}}\right\|_2\left\|r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_2 \\
&\leqslant 3\gamma_{\max}(Q_{\mathbf{SS}}^{-\frac{1}{2}})\left\|r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right\|_\infty \\
&\leqslant 3\gamma_{\min}^{-\frac{1}{2}}(Q_{\mathbf{SS}})c_2\left\|\hat{\beta}_{\mathbf{S}} - \beta^*_{\mathbf{S}}\right\|_\infty \\
&\overset{(54)}{\leqslant} 3c_2\gamma_{\min}^{-\frac{1}{2}}(Q_{\mathbf{SS}})\,|\mathbf{S}|^{\frac{1}{2}}\gamma_{\min}^{-1}(Q_{\mathbf{SS}})\,(\lambda_n A_{\mathbf{S}} + \|q_{\mathbf{S}}\|_\infty) \\
&= 3\frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|\mathbf{S}\cap G_g|}\gamma_{\min}^{-\frac{3}{2}}(Q_{\mathbf{SS}})\,|\mathbf{S}|^{\frac{1}{2}}\,(\lambda_n A_{\mathbf{S}} + \|q_{\mathbf{S}}\|_\infty)\,.
\end{aligned}$$

If the following conditions are satisfied:

$$a_{\mathbf{S}^c}^{-1}\frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|\mathbf{S}\cap G_g|}\gamma_{\min}^{-\frac{3}{2}}(Q_{\mathbf{SS}})\,|\mathbf{S}|^{\frac{1}{2}}\lambda_n A_{\mathbf{S}} \leqslant \frac{\tau}{12}, \tag{65}$$

$$a_{\mathbf{S}^c}^{-1}\frac{6}{\beta^*_{\min}}\sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|\mathbf{S}\cap G_g|}\gamma_{\min}^{-\frac{3}{2}}(Q_{\mathbf{SS}})\,|\mathbf{S}|^{\frac{1}{2}}\,\|q_{\mathbf{S}}\|_\infty \leqslant \frac{\tau}{12}, \tag{66}$$

$$(\phi^c_{\mathbf{S}})^*[Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\mathbf{r}_{\mathbf{S}}] \leqslant 1-\tau, \tag{67}$$

$$(\phi^c_{\mathbf{S}})^*[q_{\mathbf{S}^c|\mathbf{S}}] \leqslant \frac{\lambda_n\tau}{2}, \tag{68}$$

then we have

$$\begin{aligned}
(\phi^c_{\mathbf{S}})^*\left[\nabla L(\hat{\beta}^R)_{\mathbf{S}^c}\right] &\overset{(62)}{=} (\phi^c_{\mathbf{S}})^*\left[\lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right) + \lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}r_{\mathbf{S}}(\beta^*_{\mathbf{S}}) - q_{\mathbf{S}^c|\mathbf{S}}\right] \\
&\leqslant (\phi^c_{\mathbf{S}})^*\left[\lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right] + (\phi^c_{\mathbf{S}})^*\left[\lambda_n Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right] + (\phi^c_{\mathbf{S}})^*\left[-q_{\mathbf{S}^c|\mathbf{S}}\right] \\
&\leqslant \lambda_n\,(\phi^c_{\mathbf{S}})^*\left[Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right] + \lambda_n(1-\tau) + \frac{\lambda_n\tau}{2} \\
&\overset{(42)}{\leqslant} \lambda_n a\,(\mathbf{S}^c)^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}\left(r_{\mathbf{S}}(\hat{\beta}_{\mathbf{S}}) - r_{\mathbf{S}}(\beta^*_{\mathbf{S}})\right)\right\|_\infty + \lambda_n - \frac{\lambda_n\tau}{2} \\
&\leqslant \frac{\lambda_n\tau}{4} + \frac{\lambda_n\tau}{4} + \lambda_n - \frac{\lambda_n\tau}{2} \leqslant \lambda_n,
\end{aligned}$$

which is (57). Because (56) and (57) are satisfied, Lemma 23 implies that $\hat{\beta}^R$ is the optimal. Thus,

$$\mathit{supp}(\hat{\beta}^G) = \mathit{supp}(\hat{\beta}^R) = \mathbf{S}.$$

D.8.4 PART IV

The results in Part III depend on conditions (52), (53), (64), (65), (66), (67), and (68), which are summarized as follows:

$$\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\|_{2,\infty} \leqslant 3, \tag{69}$$

$$\lambda_n|\mathbf{S}|^{\frac{1}{2}} \leqslant \min\left\{\frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3A_{\mathbf{S}}}, \frac{\tau\gamma_{\min}^{\frac{3}{2}}(Q_{\mathbf{SS}})a_{\mathbf{S}^c}\beta^*_{\min}}{72A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\}, \tag{70}$$

$$(\phi^c_{\mathbf{S}})^*[Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}\mathbf{r}_{\mathbf{S}}] \leqslant 1-\tau, \tag{71}$$

$$(\phi^c_{\mathbf{S}})^*[q_{\mathbf{S}^c|\mathbf{S}}] \leqslant \frac{\lambda_n\tau}{2}, \tag{72}$$

$$\|q_{\mathbf{S}}\|_\infty \leqslant \min\left\{\frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3A_{\mathbf{S}}}, \frac{\tau\gamma_{\min}^{\frac{3}{2}}(Q_{\mathbf{SS}})a_{\mathbf{S}^c}\beta^*_{\min}}{72A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\}. \tag{73}$$

In Part IV, we want to make sure that these conditions hold with high probability.

Condition (69)

To begin with, for any matrix $A\in\mathbb{R}^{m\times n}$, the Cauchy-Schwarz inequality implies that

$$\begin{aligned}
\|A\|_{2,\infty} = \sup_{\|u\|_2\leqslant 1}\|Au\|_\infty &= \sup_{\|u\|_2\leqslant 1}\max_{i\in[m]}\left(\sqrt{\sum_{j\in[n]}A_{ij}u_j}\right)\\
&\leqslant \sup_{\|u\|_2\leqslant 1}\max_{i\in[m]}\left(\sqrt{\sum_{j\in[n]}A_{ij}^2}\sqrt{\sum_{j\in[n]}u_j^2}\right)\\
&\leqslant \max_{i\in[m]}\left(\sqrt{\sum_{j\in[n]}A_{ij}^2}\right) \leqslant \max_{i\in[m]}\left\{\sqrt{\operatorname{diag}(AA^\top)}\right\}.
\end{aligned}$$

Recall that $Q=\frac{1}{n}X^\top X$. Let $A=Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}$, we have

$$\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\|_{2,\infty} \leqslant \max\{\sqrt{\operatorname{diag}(Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c})}\}.$$

Using the Schur complement of Q on the block matrices $Q_{\mathbf{SS}}$ and $Q_{\mathbf{S}^c\mathbf{S}^c}$, the positiveness of Q implies the positiveness of $Q_{\mathbf{S}^c\mathbf{S}^c}-Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}$. Thus,

$$\max\operatorname{diag}(Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}) \leqslant \max\operatorname{diag}(Q_{\mathbf{S}^c\mathbf{S}^c}) \leqslant \max_{j\in\mathbf{S}^c}Q_{jj}.$$

Lemma 25. *(Lemma 1 of Laurent and Massart (2000))*

Suppose that the random variable U follows χ^2 distribution with d degrees of freedom, then for any positive x,

$$\begin{aligned}
\mathbb{P}(U-d\geq 2\sqrt{dx}+2x) &\leqslant \exp(-x),\\
\mathbb{P}(d-U\geq 2\sqrt{dx}) &\leqslant \exp(-x).
\end{aligned}$$

As X follows multivariate normal, $\tilde{Q}_{jj} = \frac{nQ_{jj}}{\Theta_{jj}^2} \sim \chi_n^2$. Then by Lemma 25, we have

$$\begin{aligned}
\mathbb{P}(\max_{j\in\mathbf{S}^c}\sqrt{Q_{jj}} > 3) &\leqslant \mathbb{P}(\max_{j\in\mathbf{S}^c} Q_{jj} > 5) \leqslant \mathbb{P}(\bigcup_{j\in\mathbf{S}^c} Q_{jj} > 5) \leqslant \sum_{j\in\mathbf{S}^c}\mathbb{P}(Q_{jj} > 5) \\
&\leqslant \sum_{j\in\mathbf{S}^c}\mathbb{P}(Q_{jj} > 5\Theta_{jj}^2) = \sum_{j\in\mathbf{S}^c}\mathbb{P}(n\frac{Q_{jj}}{\Theta_{jj}^2} > 5n) \\
&\leqslant \sum_{j\in\mathbf{S}^c}\mathbb{P}(\tilde{Q}_{jj} > n + 2n + 2n) \leqslant (p - |\mathbf{S}|)\exp(-n) \\
&= \exp(-n + \log(p - |\mathbf{S}|)) \\
&\leqslant \exp(-\frac{n}{2}),
\end{aligned} \tag{74}$$

where the last inequality holds as $n > 2\log(p - |\mathbf{S}|)$. Thus,

$$\mathbb{P}(\|Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-\frac{1}{2}}\|_{2,\infty} > 3) \leqslant \mathbb{P}(\max_{j\in\mathbf{S}^c}\sqrt{Q_{jj}} > 3) \leqslant \exp(-\frac{n}{2}).$$

Similarly, let $Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}} = Q_{\mathbf{S}^c\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}$. The diagonal terms of $Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}$ is less than the diagonal terms of $Q_{\mathbf{S}^c\mathbf{S}^c}$, which implies

$$\mathbb{P}(\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\|_{2,\infty} > 3) \leqslant \mathbb{P}(\max_{j\in\mathbf{S}^c}\sqrt{Q_{jj}} > 3) \leqslant \exp(-\frac{n}{2}).$$

Condition (70)

Lemma 26. *(see Wainwright, 2009, Lemma 9)*
Suppose that $d \leqslant n$ and $X \in \mathbb{R}^{n\times d}$ have i.i.d rows $X_i \sim N(0,\Theta)$, then

$$\mathbb{P}\left(\gamma_{\max}\left(\frac{1}{n}X^\top X\right) \geqslant 9\gamma_{\max}(\Theta)\right) \leqslant 2\exp(-\frac{n}{2}),$$

$$\mathbb{P}\left(\gamma_{\max}\left((\frac{1}{n}X^\top X)^{-1}\right) \geqslant \frac{9}{\gamma_{\min}(\Theta)}\right) \leqslant 2\exp(-\frac{n}{2}).$$

As we assume that $|\mathbf{S}| \leqslant n$ and $X_{\mathbf{SS}} \sim N(0, \Theta_{\mathbf{SS}})$, then Lemma 26 implies

$$\mathbb{P}\left(\gamma_{\max}(Q_{\mathbf{SS}}) \geqslant 9\gamma_{\max}(\Theta_{\mathbf{SS}})\right) \leqslant 2\exp(-\frac{n}{2}),$$

and also

$$\mathbb{P}\left(\gamma_{\min}(\Theta_{\mathbf{SS}}) \geqslant 9\gamma_{\min}(Q_{\mathbf{SS}})\right) \leqslant 2\exp(-\frac{n}{2}).$$

Thus, by assuming that

$$\lambda_n|\mathbf{S}|^{\frac{1}{2}} \leqslant \min\left\{\frac{3\gamma_{\min}(\Theta)\beta_{\min}^*}{A_{\mathbf{S}}}, \frac{\tau\gamma_{\min}^{\frac{3}{2}}(\Theta)a_{\mathbf{S}^c}\beta_{\min}^*}{8A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\},$$

we have

$$\lambda_n |\mathbf{S}|^{\frac{1}{2}} \leqslant \min \left\{ \frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta^*_{\min}}{3A_{\mathbf{S}}}, \frac{\tau \gamma^{\frac{3}{2}}_{\min}(Q_{\mathbf{SS}}) a_{\mathbf{S}^c} \beta^*_{\min}}{72 A_{\mathbf{S}} \sum\limits_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \right\}$$

holds with high probability.

Condition (71)

For any $j \in \mathbf{S}^c$, $X_j \in \mathbb{R}^n$ is zero-mean Gaussian. Following the decomposition in Wainwright (2009), we have

$$X_j^\top = \Theta_{j\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top + E_j^\top, \tag{75}$$

where E_j are i.i.d from $N\left(0, \left[\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right]_{jj}\right)$ with $\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}} = \Theta_{\mathbf{S}^c\mathbf{S}^c} - \Theta_{\mathbf{S}^c\mathbf{S}}\left(\Theta_{\mathbf{SS}}\right)^{-1}\Theta_{\mathbf{SS}^c}$. Let $E_{\mathbf{S}^c}$ be an $|S^c| \times n$ matrix, with each row representing E_j for an element $j \in \mathbf{S}^c$, then we have

$$\begin{aligned} Q_{\mathbf{S}^c\mathbf{S}} Q_{\mathbf{SS}}^{-1} \mathbf{r}_{\mathbf{S}} &= X_{\mathbf{S}^c}^\top X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}} \\ &\overset{(75)}{=} \left(\Theta_{\mathbf{S}^c\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top + E_{\mathbf{S}^c}^\top\right) X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}} \\ &= \Theta_{\mathbf{S}^c\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \mathbf{r}_{\mathbf{S}} + E_{\mathbf{S}^c}^\top X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}} \\ &:= \Theta_{\mathbf{S}^c\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \mathbf{r}_{\mathbf{S}} + \eta. \end{aligned} \tag{76}$$

The preceding expression prompts us to establish an upper bound for the dual norm of η. To achieve this, we begin by examining the scenario in which $\underline{X_{\mathbf{S}} \text{ is fixed}}$. Our objective now is to derive the covariance matrix of η. For any $j \in \mathbf{S}^c$, we have

$$\mathbb{E}[\eta_j] = \mathbb{E}\left[E_j^\top X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}}\right] = 0.$$

For any pair of $j, k \in \mathbf{S}^c$, we have

$$\begin{aligned} \mathbb{E}[\eta_j \eta_k] &= \mathbb{E}\left[E_j^\top X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}} E_k^\top X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}}\right] \\ &= \mathbb{E}\left[\mathbf{r}_{\mathbf{S}}^\top (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} X_{\mathbf{S}}^\top E_j E_k^\top X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}}\right] \\ &= \mathbf{r}_{\mathbf{S}}^\top (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} X_{\mathbf{S}}^\top \mathbb{E}\left[E_j E_k^\top\right] X_{\mathbf{S}} (X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1} \mathbf{r}_{\mathbf{S}}, \end{aligned}$$

where

$$\begin{aligned} &\mathbb{E}\left[E_j E_k^\top\right] \overset{(75)}{=} \mathbb{E}\left[\left(X_j - X_{\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \Theta_{j\mathbf{S}}^\top\right)\left(X_k^\top - \Theta_{k\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top\right)\right] \\ &= \mathbb{E}\left[X_j X_k^\top\right] - \mathbb{E}\left[X_{\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \Theta_{j\mathbf{S}} X_k^\top\right] - \mathbb{E}\left[X_j \Theta_{k\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top \mid X_{\mathbf{S}}\right] + \mathbb{E}\left[X_{\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \Theta_{j\mathbf{S}}^\top \Theta_{k\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top\right] \\ &= \mathbb{E}\left[X_j X_k^\top\right] - X_{\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \Theta_{j\mathbf{S}} \mathbb{E}\left[X_k^\top\right] - \mathbb{E}\left[X_j\right] \Theta_{k\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top + X_{\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} \Theta_{j\mathbf{S}} \Theta_{k\mathbf{S}} \Theta_{\mathbf{SS}}^{-1} X_{\mathbf{S}}^\top \\ &= \mathbb{E}\left[X_j X_k^\top\right] - \mathbb{E}\left[X_j\right] \mathbb{E}\left[X_k^\top\right] = \operatorname{Cov}\left[X_j, X_k^\top\right] = \left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk} \mathbf{I}_{n \times n}. \end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}[\eta_j\eta_k] &= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}X_{\mathbf{S}}^\top\mathbb{E}\left[E_jE_k^\top\right]X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\\ &= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}X_{\mathbf{S}}^\top\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk}\mathbf{I}_{n\times n}X_{\mathbf{S}}(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\\ &= \mathbf{r}_{\mathbf{S}}^\top(X_{\mathbf{S}}^\top X_{\mathbf{S}})^{-1}\mathbf{r}_{\mathbf{S}}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk} = \frac{\mathbf{r}_{\mathbf{S}}^\top(Q_{\mathbf{SS}})^{-1}\mathbf{r}_{\mathbf{S}}}{n}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)_{jk}.\end{aligned}$$

And we have $\mathrm{Cov}(\eta) = \frac{\mathbf{r}_{\mathbf{S}}^\top(Q_{\mathbf{SS}})^{-1}\mathbf{r}_{\mathbf{S}}}{n}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right) := \Xi$.

Lemma 27. *(Theorem 2.26 in Wainwright (2019))*

Let $(X_1,\ldots,X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f:\mathbb{R}^n\mapsto\mathbb{R}$ be a Lipschitz function with respect to the Euclidean norm and Lipschitz constant L. Then the variable $f(X)-\mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$\mathbb{P}[|f(X)-\mathbb{E}[f(X)]|\geqslant t]\leqslant 2\exp(-\frac{t^2}{2L^2})\quad \textit{for all } t\geqslant 0.$$

To apply the concentration bound in Lemma 27, we define function $\Psi(u) = (\phi^*_{\mathbf{S}^c})\left[\Xi^{\frac{1}{2}}u\right]$. As $\eta = \Xi^{\frac{1}{2}}W$ where $W\sim N(0, I_{|\mathbf{S}^c|\times|\mathbf{S}^c|})$, $(\phi^c_{\mathbf{S}})^*(\eta)$ has the same distribution as $\Psi(W)$. We continue to show that Ψ is a Lipschitz function given fixed $X_{\mathbf{S}}$.

$$\begin{aligned}|\Psi(u)-\Psi(v)| &\leqslant \Psi(u-v) = (\phi^c_{\mathbf{S}})^*\left[\Xi^{\frac{1}{2}}(u-v)\right]\\ &\leqslant a_{\mathbf{S}}^{-1}\left\|\Xi^{\frac{1}{2}}(u-v)\right\|_\infty\\ &= a_{\mathbf{S}}^{-1}\left\|\left[\frac{\mathbf{r}_{\mathbf{S}}^\top(Q_{\mathbf{SS}})^{-1}\mathbf{r}_{\mathbf{S}}}{n}\cdot\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)\right]^{\frac{1}{2}}(u-v)\right\|_\infty\\ &\leqslant a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_2 n^{-\frac{1}{2}}\gamma_{\max}^{\frac{1}{2}}\left(Q_{\mathbf{SS}}^{-1}\right)\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)\|u-v\|_2.\end{aligned}$$

Thus, the corresponding Lipstichiz constant is

$$L_\eta = a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_2 n^{-\frac{1}{2}}\gamma_{\max}^{\frac{1}{2}}\left(Q_{\mathbf{SS}}^{-1}\right)\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right).$$

On the other hand, suppose that $\mathbb{E}\left[(\phi^c_{\mathbf{S}})^*(\eta)\right]\leqslant\frac{\tau}{4}$, since Ψ is a Lipschitiz function, by applying $t=\frac{\tau}{4}$ in concentration Lemma 27 on Lipschitz functions of multivariate standard random variables, we have

$$\begin{aligned}\mathbb{P}\left((\phi^c_{\mathbf{S}})^*[\eta]>\frac{\tau}{2}\right) &= \mathbb{P}\left(\Psi(W)>\frac{\tau}{2}\right) = \mathbb{P}\left(\Psi(W)-\frac{\tau}{4}>\frac{\tau}{4}\right)\\ &\leqslant \mathbb{P}\left(\Psi(W)-E\left[(\phi^c_{\mathbf{S}})^*(\eta)\right]>\frac{\tau}{4}\right)\\ &= \mathbb{P}\left(\Psi(W)-E\left[\Psi(W)\right]>\frac{\tau}{4}\right)\leqslant\exp\left(-\frac{\tau^2}{4L_\eta^2}\right).\end{aligned}$$

Now we further assume that $\{\gamma_{\max}(Q_{\mathbf{SS}}^{-1})\leqslant\frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\}$. Under this condition, we have

$$L_\eta = a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_2 n^{-\frac{1}{2}}\gamma_{\max}^{\frac{1}{2}}\left(Q_{\mathbf{SS}}^{-1}\right)\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)\leqslant\frac{3a_{\mathbf{S}}^{-1}\left\|\mathbf{r}_{\mathbf{S}}\right\|_2\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}}. \tag{77}$$

Lemma 28. *(Sudakov inequality, Theorem 5.27 in Wainwright (2019)) If X and Y are a.s. bounded, centered Gaussian processes on T such that*

$$\mathbb{E}\left(X_t - X_s\right)^2 \leq \mathbb{E}\left(Y_t - Y_s\right)^2$$

then

$$\mathbb{E}\sup_T X_t \leq \mathbb{E}\sup_T Y_t.$$

Lemma 29. *(Exercise 2.12 in Wainwright (2019)) Let $X_1, \ldots, X_n$ be independent σ^2-subgaussian random variables. Then*

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

On the other hand, for any u_t, u_s, we have

$$\begin{aligned}
&\mathbb{E}(u_t^\top \eta - u_s^\top \eta)^2 = \mathbb{E}(u_t^\top \Xi^{\frac{1}{2}} W - u_s^\top \Xi^{\frac{1}{2}} W)^2 = (u_t - u_s)^\top \Xi (u_t - u_s) \\
\leqslant & ||u_t - u_s||_2^2 \gamma_{\max}(\Xi) = \mathbb{E}(\gamma_{\max}^{\frac{1}{2}}(\Xi)\, u_t^\top W - \gamma_{\max}^{\frac{1}{2}}(\Xi)\, u_s^\top W)^2.
\end{aligned}$$

By using Sudakov-Fernique inequality in Lemma 28, we have

$$\mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top \Xi^{\frac{1}{2}} W\Big] \leqslant \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} \gamma_{\max}^{\frac{1}{2}}(\Xi)\, u^\top W\Big].$$

Consequently,

$$\begin{aligned}
\mathbb{E}\Big[(\phi_{\mathbf{S}}^c)^*(\eta)\Big] &= \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top \eta\Big] = \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top \Xi^{\frac{1}{2}} W\Big] \\
&\leqslant \gamma_{\max}^{\frac{1}{2}}(\Xi)\, \mathbb{E}\Big[\sup_{\phi_{\mathbf{S}}^c(u)\leqslant 1} u^\top W\Big] = \gamma_{\max}(\Xi)^{\frac{1}{2}}\, \mathbb{E}\Big[(\phi_{\mathbf{S}}^c)^*(W)\Big].
\end{aligned} \tag{78}$$

Notice that

$$\begin{aligned}
\|\mathbf{r}_{\mathbf{S}}\|_2^2 &\leqslant |\mathbf{S}| \max_{j\in\mathbf{S}} \mathbf{r}_j^2 = |\mathbf{S}| \Big(\max_{j\in\mathbf{S}}\{\beta_j^* \cdot \sum_{g\in\mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} \frac{w_g}{\|\beta^*_{G_g\cap\mathbf{S}}\|_2}\}\Big)^2 \\
&\leqslant |\mathbf{S}| \Big(\max_{j\in\mathbf{S}}\{|\beta_j^*|\} \cdot \max\{\sum_{g\in\mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} \frac{w_g}{\|\beta^*_{G_g\cap\mathbf{S}}\|_2}\}\Big)^2 \\
&\leqslant |\mathbf{S}| \Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}} \cdot \max\{\sum_{g\in\mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} \frac{w_g}{\sqrt{|G_g\cap\mathbf{S}}}\}\Big)^2 \\
&\leqslant |\mathbf{S}| \Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}} \cdot \max\{\sum_{g\in\mathsf{G}_{\mathbf{S}}^G, G_g\cap j\neq\emptyset} w_g\}\Big)^2 \\
&\leqslant |\mathbf{S}| \Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}} \cdot h_{\max}(\mathbf{G}_{\mathbf{S}}) \max_{g\in\mathsf{G}_{\mathbf{S}}^G} w_g\}\Big)^2 \\
&\leqslant \Big(\frac{\max_{j\in\mathbf{S}}\{|\beta_j^*|\}}{\beta^*_{\min}}\Big)^2 |\mathbf{S}| A_{\mathbf{S}}^2 = \max_{j\in\mathbf{S}}\{(\beta_j^*)^2\} |\mathbf{S}| \Big(\frac{A_{\mathbf{S}}}{\beta^*_{\min}}\Big)^2 \\
&\lesssim \frac{\max_{j\in\mathbf{S}}\{(\beta_j^*)^2\}}{\lambda_n^2}.
\end{aligned} \tag{79}$$

Thus, if $X_{\mathbf{S}}$ satisfies $\gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}$, we have

$$\begin{aligned}
\mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(\eta)\right] &\overset{(78)}{\leqslant} \gamma_{\max}\left(\Xi\right)^{\frac{1}{2}}\, \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\
&\leqslant \frac{\|\mathbf{r}_{\mathbf{S}}\|_2\, \gamma_{\min}^{-\frac{1}{2}}\left(Q_{\mathbf{SS}}\right) \gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{n^{\frac{1}{2}}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\
&\leqslant \frac{\|\mathbf{r}_{\mathbf{S}}\|_2\, 3\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\
&\overset{(42)}{\leqslant} \frac{\|\mathbf{r}_{\mathbf{S}}\|_2\, 3\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \mathbb{E}\left[a_{\mathbf{S}^c}^{-1}\|W\|_\infty\right] \\
&\leqslant \frac{\|\mathbf{r}_{\mathbf{S}}\|_2\, 3\gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{a_{\mathbf{S}^c}(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \mathbb{E}\left[\|W\|_\infty\right] \\
&\overset{\text{Lemma 29}}{\leqslant} \frac{6\, \|\mathbf{r}_{\mathbf{S}}\|_2\, \gamma_{\max}^{\frac{1}{2}}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{a_{\mathbf{S}^c}(n\gamma_{\min}(\Theta_{\mathbf{SS}}))^{\frac{1}{2}}} \sqrt{\log(p-|\mathbf{S}|)} \leqslant \frac{\tau}{4},
\end{aligned} \tag{80}$$

where the last inequality holds as Assumption 6 implies that

$$n \gtrsim \frac{\max\limits_{j\in\mathbf{S}}\{(\beta_j^*)^2\}\log(p-|\mathbf{S}|)}{a_{\mathbf{S}^c}^2\lambda_n^2} \overset{(79)}{\gtrsim} \frac{\|\mathbf{r}_{\mathbf{S}}\|_2^2\log(p-|\mathbf{S}|)}{a_{\mathbf{S}^c}^2} \geqslant \frac{576\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\log(p-|\mathbf{S}|)\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}{a_{\mathbf{S}^c}^2\gamma_{\min}(\Theta_{\mathbf{SS}})\tau^2}.$$

Consequently, (77) and (80) together implies

$$\begin{aligned}
&\mathbb{P}\Big(\, (\phi_{\mathbf{S}}^c)^*\left[\eta\right] > \frac{\tau}{2} \mid X_{\mathbf{S}}, \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\Big) \\
&\leqslant \exp\Big(-\frac{\tau^2}{4L_\eta^2}\Big) \leqslant \exp\Big(-\frac{\tau^2 n a_{\mathbf{S}}^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{12\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\,\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\Big).
\end{aligned} \tag{81}$$

Thus, let $\mathscr{A}$ be the event $\{X_{\mathbf{S}} \mid \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\}$. We have

$$\begin{aligned}
\mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\left[\eta\right] > \frac{\tau}{2} \mid X_{\mathbf{S}}\right) &= \mathbb{P}\Big(\, (\phi_{\mathbf{S}}^c)^*\left[\eta\right] > \frac{\tau}{2} \mid X_{\mathbf{S}}, \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) \leqslant \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\Big) \\
&\quad + \mathbb{P}\Big(\, (\phi_{\mathbf{S}}^c)^*\left[\eta\right] > \frac{\tau}{2} \mid X_{\mathbf{S}}, \gamma_{\max}(Q_{\mathbf{SS}}^{-1}) > \frac{9}{\gamma_{\min}(\Theta_{\mathbf{SS}})}\Big) \\
&\leqslant \exp\left(-\frac{\tau^2 n a_{\mathbf{S}}^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\,\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&\leqslant \exp\left(-\frac{\tau^2 n a_{\mathbf{S}}^2\gamma_{\min}(\Theta_{\mathbf{SS}})}{4\,\|\mathbf{r}_{\mathbf{S}}\|_2^2\,\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\right) + 2\exp(-\frac{n}{2}).
\end{aligned}$$

Condition (72)

Now we are going to study condition (72). Recall that $q_{\mathbf{S}^c|\mathbf{S}} = q_{\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}$ and $Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}} = Q_{\mathbf{S}^c\mathbf{S}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}$. Given X, $q_{\mathbf{S}^c|\mathbf{S}}$ is a centered Gaussian random vector with covariance matrix

$$\begin{aligned}\mathbb{E}\left[q_{\mathbf{S}^c|\mathbf{S}}q_{\mathbf{S}^c|\mathbf{S}}^\top\right] &= \mathbb{E}\left[q_{\mathbf{S}^c}q_{\mathbf{S}^c}^\top - q_{\mathbf{S}^c}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c} - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}^c}^\top + Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}\right]\\ &= \mathbb{E}\left[q_{\mathbf{S}^c}q_{\mathbf{S}^c}^\top - Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}\right]\\ &= \mathbb{E}\left[q_{\mathbf{S}^c}q_{\mathbf{S}^c}^\top\right] - \mathbb{E}\left[Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}q_{\mathbf{S}}q_{\mathbf{S}}^\top Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c}\right]\\ &= \frac{\sigma^2}{n}Q_{\mathbf{S}^c\mathbf{S}^c} - \frac{\sigma^2}{n}Q_{\mathbf{S}^c\mathbf{S}}Q_{\mathbf{SS}}^{-1}Q_{\mathbf{SS}^c} := \frac{\sigma^2}{n}Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}.\end{aligned}$$

Next, we define $\psi(u) = (\phi_{\mathbf{S}}^c)^*\left(\sigma n^{-1/2}Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}u\right)$ so that $(\phi_{\mathbf{S}^c}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right]$ has the same distribution as $\psi(W)$. Now we want to show that ψ is a Lipschitz function

$$\begin{aligned}|\psi(u)-\psi(v)| \leqslant \psi(u-v) &= (\phi_{\mathbf{S}}^c)^*\left(\sigma n^{-1/2}Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}(u-v)\right)\\ &\leqslant \sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}(u-v)\right\|_\infty\\ &\leqslant \sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}\right\|_{2,\infty}\|(u-v)\|_\infty\\ &\leqslant \sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}\right\|_{2,\infty}\|(u-v)\|_2\,.\end{aligned}$$

Suppose that $\left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\right\|_{2,\infty} \leqslant 3$, then ψ is a Lipschitz function with Lipschitz constant $3\sigma n^{-1/2}a_{\mathbf{S}^c}^{-1}$. In addition, if $\mathbb{E}[(\phi_{\mathbf{S}}^c)^*(q_{\mathbf{S}^c|\mathbf{S}})] \leqslant \frac{\lambda_n\tau}{4}$, then by Lemma 27 , we have for $t = \frac{\lambda_n\tau}{4}$,

$$\begin{aligned}\mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2}\right) &= \mathbb{P}\left(\psi(W) > \frac{\lambda_n\tau}{2}\right) = \mathbb{P}\left(\psi(W) - \frac{\lambda_n\tau}{4} > \frac{\lambda_n\tau}{4}\right)\\ &\leqslant \mathbb{P}\left(\psi(W) - \mathbb{E}[(\phi_{\mathbf{S}}^c)^*(q_{\mathbf{S}^c|\mathbf{S}})] > \frac{\lambda_n\tau}{4}\right)\\ &= \mathbb{P}\left(\psi(W) - \mathbb{E}\left[\psi(W)\right] > \frac{\lambda_n\tau}{4}\right) \leqslant \exp\left(-\frac{\tau^2\lambda_n^2 n a_{\mathbf{S}^c}^2}{144\sigma^2}\right).\end{aligned}$$

Now, we consider random X. For any u_t, u_s, we have

$$\begin{aligned}\mathbb{E}\big[(u_t-u_s)^\top q_{\mathbf{S}^c|\mathbf{S}}\big]^2 = \frac{\sigma^2}{n}(u_t-u_s)^\top Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}(u_t-u_s) &\leqslant \frac{\sigma^2}{n}\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{\frac{1}{2}}\|_2^2\|(u_t-u_s)\|_2^2\\ = \mathbb{E}\big[\sigma n^{-\frac{1}{2}}\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2^{\frac{1}{2}}(u_t-u_s)^\top W\big]^2.&\end{aligned}$$

By using Sudakov-Fernique inequality, if $\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2 \leqslant 9$, we get

$$\begin{aligned}\mathbb{E}[(\phi_{\mathbf{S}}^c)^*(q_{\mathbf{S}^c|\mathbf{S}})] &= \mathbb{E} \sup_{\phi_{\mathbf{S}}^c(u)\leq 1} u^\top q_{\mathbf{S}^c|\mathbf{S}} \\ &\leqslant \sigma n^{-1/2} \|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2^{\frac{1}{2}} \mathbb{E} \sup_{\phi_{\mathbf{S}}^c(u)\leq 1} u^\top W \\ &\leqslant \sigma n^{-\frac{1}{2}} \|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\|_2^{\frac{1}{2}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\ &\leqslant 3\sigma n^{-\frac{1}{2}} \mathbb{E}\left[(\phi_{\mathbf{S}}^c)^*(W)\right] \\ &\leqslant \frac{\lambda_n \tau}{4}.\end{aligned} \tag{82}$$

On the other hand, Assumption 1' and 6 imply that

$$\frac{9\sigma^2 \mathbb{E}^2\left[(\phi_{\mathbf{S}}^c)^*(W)\right]}{n} \leqslant \frac{9\sigma^2 \log(p-|\mathbf{S}|)}{a_{\mathbf{S}^c}^2 n} \leqslant \frac{\lambda_n^2\tau^2}{16}.$$

Therefore, we have

$$\mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X, \left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\right\|_{2,\infty} \leqslant 3\right) \leqslant \exp\left(-\frac{\tau^2 n \lambda_n^2 a_{\mathbf{S}^c}^2}{144\sigma^2}\right).$$

Let $\mathscr{B}$ be the event $\{X \mid \left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\right\|_{2,\infty} \leqslant 3\}$. We have

$$\begin{aligned}\mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X\right) &= \mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X, \left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\right\|_{2,\infty} \leqslant 3\right) \\ &+\mathbb{P}\left((\phi_{\mathbf{S}}^c)^*\left[q_{\mathbf{S}^c|\mathbf{S}}\right] \geqslant \frac{\lambda_n\tau}{2} \mid X, \left\|Q_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}^{1/2}\right\|_{2,\infty} > 3\right) \\ &\leqslant \exp\left(-\frac{\tau^2 n \lambda_n^2 a_{\mathbf{S}^c}^2}{144\sigma^2}\right) + \mathbb{P}\left(\mathscr{B}^c\right) \\ &\overset{(69)}{\leqslant} \exp\left(-\frac{\tau^2 n \lambda_n^2 a_{\mathbf{S}^c}^2}{144\sigma^2}\right) + \exp(-\frac{n}{2}).\end{aligned}$$

Condition (73)

The last condition (73) lead us to control the term $\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G)\right)$, with

$$c'(\mathbf{S}, G) = \min\left\{\frac{\gamma_{\min}\left(Q_{\mathbf{SS}}\right)\beta_{\min}^*}{3A_{\mathbf{S}}}, \frac{\tau\gamma_{\min}^{\frac{3}{2}}(Q_{\mathbf{SS}})a_{\mathbf{S}^c}\beta_{\min}^*}{72A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\}.$$

For any given X, Jenatton et al. (2011a) showed that for any $\delta > 0$,

$$\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant \delta\right) \leqslant 2|\mathbf{S}| \exp\left(-\frac{n\delta^2}{2\sigma^2}\right).$$

Recall under the event $\mathscr{A}$, we have

$$\frac{\gamma_{\min}(\Theta_{\mathbf{SS}})}{9} \leqslant \gamma_{\min}(Q_{\mathbf{SS}}).$$

Which implies that

$$
\begin{aligned}
c'(\mathbf{S}, G) &\geqslant \min\left\{\frac{\gamma_{\min}\left(\Theta_{\mathbf{SS}}\right)\beta^*_{\min}}{27A_{\mathbf{S}}}, \frac{\tau\gamma_{\min}(\Theta_{\mathbf{SS}})^{\frac{3}{2}}a_{\mathbf{S}^c}\beta^*_{\min}}{648A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\} \\
&\geqslant \min\left\{\frac{\beta^*_{\min}}{27c_1A_{\mathbf{S}}}, \frac{\tau a_{\mathbf{S}^c}\beta^*_{\min}}{648c_1^{\frac{3}{2}}A_{\mathbf{S}}\sum\limits_{g\in\mathsf{G}_{\mathbf{S}}} w_g\sqrt{|G_g\cap\mathbf{S}|}}\right\} := c(\mathbf{S}, G).
\end{aligned}
$$

Thus, consider random X, we have

$$
\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \mid \mathscr{A}\right) \leqslant \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c(\mathbf{S}, G) \mid \mathscr{A}\right) \leqslant 2|\mathbf{S}|\exp\left(-\frac{nc^2(\mathbf{S}, G)}{2\sigma^2}\right).
$$

Thus,

$$
\begin{aligned}
\mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G)\right) &= \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \cap \mathscr{A}\right) + \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \cap \mathscr{A}^c\right) \\
&\leqslant \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \cap \mathscr{A}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&= \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \mid \mathscr{A}\right)\mathbb{P}\left(\mathscr{A}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&\leqslant \mathbb{P}\left(\|q_{\mathbf{S}}\|_\infty \geqslant c'(\mathbf{S}, G) \mid \mathscr{A}\right) + \mathbb{P}\left(\mathscr{A}^c\right) \\
&\leqslant 2|\mathbf{S}|\exp\left(-\frac{nc^2(\mathbf{S}, G)}{2\sigma^2}\right) + 2\exp(-n/2).
\end{aligned}
$$

In summary, the probability of one of the conditions being violated is upper bound by

$$
8\exp(-\frac{n}{2})+\exp\left(-\frac{na_{\mathbf{S}}^2\tau^2\gamma_{\max}(\Theta_{\mathbf{SS}})}{4\left\|\mathbf{r}_{\mathbf{S}}\right\|_2^2\gamma_{\max}\left(\Theta_{\mathbf{S}^c\mathbf{S}^c|\mathbf{S}}\right)}\right)+\exp\left(-\frac{n\lambda_n^2\tau^2a_{\mathbf{S}^c}^2}{32\sigma^2c_2^4}\right)+2|\mathbf{S}|\exp\left(-\frac{nc^2(\mathbf{S}, G)}{2\sigma^2}\right).
$$

D.8.5 PART V

First, given the original group structure G and its induced counterpart $\mathscr{G}$, along with their respective weights w and $\mathscr{w}$, we consider the scenario where $\mathbf{J} = \mathbf{S}$. For all $\beta \in \mathbb{R}^p$, we have

$$
\begin{aligned}
\phi_{\mathbf{S}}^G(\beta_{\mathbf{S}}) &= \sum_{g\in\mathsf{G}_{\mathbf{S}}^G} w_g\|\beta_{\mathbf{S}\cap G_g}\|_2 \leqslant \sum_{g\in\mathsf{G}_{\mathbf{S}}} w_g\Big(\sum_{\mathscr{g}:\mathscr{g}\in F^{-1}(g),\mathscr{G}_{\mathscr{g}}\subset\mathbf{S}} \|\beta_{\mathbf{S}\cap\mathscr{G}_{\mathscr{g}}}\|_2\Big) \\
&= \sum_{\mathscr{g}:\mathscr{G}_{\mathscr{g}}\subset\mathbf{S}}\Big(\sum_{g:g\in F(\mathscr{g}),g\in\mathsf{G}_{\mathbf{S}}} w_g\Big)\|\beta_{\mathbf{S}\cap\mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g}:\mathscr{G}_{\mathscr{g}}\subset\mathbf{S}}\Big(\sum_{g:g\in F(\mathscr{g})} w_g\Big)\|\beta_{\mathbf{S}\cap\mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g}\in\mathsf{G}_{\mathbf{S}}^{\mathscr{G}}} \mathscr{w}_{\mathscr{g}}\|\beta_{\mathbf{S}\cap\mathscr{G}_{\mathscr{g}}}\|_2 = \phi_{\mathbf{S}}^{\mathscr{G}}(\beta).
\end{aligned}
\tag{83}
$$

Since $\phi_{\mathbf{S}}^G(\beta) \leqslant \phi_{\mathbf{S}}^{\mathscr{G}}(\beta)$, we can set $a_{\mathbf{S}}^{\mathscr{G}} = a_{\mathbf{S}}^G = \min\limits_{g\in\mathsf{G}_{\mathbf{S}}^G}\frac{w_g}{\sqrt{d_g}}$. Since

$$
\max_{\mathscr{g}\in\mathsf{G}_{\mathbf{S}}^{\mathscr{G}}} \mathscr{w}_{\mathscr{g}} = \max_{\mathscr{g}:\mathscr{G}_{\mathscr{g}}\cap\mathbf{S}\neq\emptyset}\sum_{g\in F(\mathscr{g})} w_g \leqslant h_{\max}(\mathbf{G}_{\mathbf{S}})\max_{g\in\mathsf{G}_{\mathbf{S}}^G} w_g,
$$

we can set $A^{\mathscr{G}}_{\mathbf{S}} = A^{G}_{\mathbf{S}}$. On the other hand, for all $\beta \in \mathbb{R}^p$, we have

$$
\begin{aligned}
(\phi^{G}_{\mathbf{S}})^c(\beta^c_{\mathbf{S}}) &= \sum_{g \in [m] \backslash \mathsf{G}^{G}_{\mathbf{S}}} w_g \|\beta_{\mathbf{S}^c \cap G_g}\|_2 \leqslant \sum_{g \in [m] \backslash \mathsf{G}_{\mathbf{S}}} w_g \big(\sum_{\mathscr{g}: \mathscr{g} \in F^{-1}(g), \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}^c} \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 \big) \\
&= \sum_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}^c} \big(\sum_{g: g \in F(\mathscr{g}), g \in [m] \backslash \mathsf{G}^{G}_{\mathbf{S}}} w_g \big) \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g}: \mathscr{G}_{\mathscr{g}} \subset \mathbf{S}^c} \big(\sum_{g: g \in F(\mathscr{g})} w_g \big) \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 \\
&= \sum_{\mathscr{g} \in [m] \backslash \mathsf{G}^{\mathscr{G}}_{\mathbf{S}}} w_{\mathscr{g}} \|\beta_{\mathbf{S}^c \cap \mathscr{G}_{\mathscr{g}}}\|_2 = (\phi^{\mathscr{G}}_{\mathbf{S}})^c(\beta).
\end{aligned}
\tag{84}
$$

Consequently, with an trivial extension, we can set $a^{\mathscr{G}}_{\mathbf{S}^c} = a^{G}_{\mathbf{S}^c} \leqslant \min\limits_{g \in \mathsf{G}^{G}_{\mathbf{S^c}}} w_g / \sqrt{d_g}$.

Based on the result of Theorem 5.1, (28) holds if

$$
\lambda_n |\mathbf{S}|^{\frac{1}{2}} \lesssim \min \Big\{ \frac{\beta^*_{\min}}{A_{\mathbf{S}}}, \frac{\beta^*_{\min} a_{\mathbf{S^c}}}{A_{\mathbf{S}} \sum\limits_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}} \Big\}.
$$

By the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
\sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|} &\leqslant \sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sum_{\mathscr{g} \in F^{-1}(g)} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \\
&= \sum_{\mathscr{g} \in F^{-1}(g), g \in \mathsf{G}_{\mathbf{S}}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \big(\sum_{g \in F(\mathscr{g})} w_g \big) \\
&= \sum_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}.
\end{aligned}
$$

If $F^{-1}(g) = O(1)$ for every $g \in \mathsf{G}_{\mathbf{S}}$, we have

$$
|G_g \cap \mathbf{S}| = \sum_{\mathscr{g} \in F^{-1}(g)} |\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}| \asymp \big(\sum_{\mathscr{g} \in F^{-1}(g)} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|} \big)^2.
$$

Consequent, we have $\sqrt{|G_g \cap \mathbf{S}|} \asymp \sum\limits_{\mathscr{g} \in F^{-1}(g)} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}$,

$$
\sum_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|} \asymp \sum_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|},
$$

and

$$
\min \Big\{ \frac{\beta^*_{\min}}{A^{G}_{\mathbf{S}}}, \frac{\beta^*_{\min} a^{G}_{\mathbf{S^c}}}{A^{G}_{\mathbf{S}} \sum\limits_{g \in \mathsf{G}_{\mathbf{S}}} w_g \sqrt{|G_g \cap \mathbf{S}|}} \Big\} \asymp \min \Big\{ \frac{\beta^*_{\min}}{A^{\mathscr{G}}_{\mathbf{S}}}, \frac{\beta^*_{\min} a^{\mathscr{G}}_{\mathbf{S^c}}}{A^{\mathscr{G}}_{\mathbf{S}} \sum\limits_{\mathscr{g} \in \mathscr{G}_{\mathbf{S}}} w_{\mathscr{g}} \sqrt{|\mathscr{G}_{\mathscr{g}} \cap \mathbf{S}|}} \Big\}.
$$