Abstract

Group lasso is a commonly used regularization method in statistical learning in which parameters are eliminated from the model according to predefined groups. However, when the groups overlap, optimizing the group lasso penalized objective can be time-consuming on large-scale problems because of the non-separability induced by the overlapping groups. This bottleneck has seriously limited the application of overlapping group lasso regularization in many modern problems, such as gene pathway selection and graphical model estimation. In this paper, we propose a separable penalty as an approximation of the overlapping group lasso penalty. Thanks to the separability, the computation of regularization based on our penalty is substantially faster than that of the overlapping group lasso, especially for large-scale and high-dimensional problems. We show that the penalty is the tightest separable relaxation of the overlapping group lasso norm within the family of $\ell_1/\ell_2$ norms. Moreover, we show that the estimator based on the proposed separable penalty is statistically equivalent to the one based on the overlapping group lasso penalty with respect to their error bounds and the rate-optimal performance under the squared loss. We demonstrate the faster computational time and statistical equivalence of our method compared with the overlapping group lasso in simulation examples and a classification problem of cancer tumors based on gene expression and multiple gene pathways.

Keywords overlapping group lasso · separable approximation · computational efficiency · statistical error bound · 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 are highly correlated within the groups. In this context, variable selection or model regularization should also consider the grouping patterns, and one may prefer to either include the whole group of variable in the selection or completely rule out the group. Group lasso (Yuan and Lin, 2006) is one popular method designed for this group selection task via adding $\ell_1/\ell_2$ regularization, as a broader class for group selection. (Zhao et al., 2009; Levina et al., 2008; Meier et al., 2008; Yan and Bien, 2017; Bach, 2008; Ravikumar et al., 2009; Xiang et al., 2015; Loh, 2014; Campbell and Allen, 2017; Tank et al., 2017; Danaher et al., 2014; Basu et al., 2015; Austin et al., 2020; Yang and Peng, 2020).

While in the original motivation of group lasso assumes disjoint groups, overlapping groups also appear frequently in many applications such as tumor metastasis analysis (Zhao et al., 2009; Jacob et al., 2009; Yuan et al., 2011; Chen et al., 2012) and structured model estimations (Cheng et al., 2017; Mohan et al., 2014; Yu and Bien, 2017; Tarzanagh and Michailidis, 2018). In tumor metastasis analysis, scientists usually aim to select a small number of tumor-related genes. Biological theory indicates that genes do not function in isolation but rather that suites of genes act in concert to perform biological functions. Hence, the gene selection is more meaningful if co-functioning groups of genes are selected together (Ma and Kosorok, 2010). In particular, pathway analysis renders mechanistic insight into the co-functioning pattern, which results in overlapping groups of genes. Applying group lasso with these overlapping groups is then a natural way to incorporate the prior group information into tumor metastasis analysis. For another example, graphical model estimation has been widely used to express 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, the groups are naturally predetermined by the parameter vector corresponding to each edge, and groups overlap because they represent edges sharing common nodes. In this scenario, the group lasso penalty provides a natural way for edge selection.
The original group lasso problem with non-overlapping groups can be solved efficiently (Qin et al., 2013; Yang and Zou, 2015; Friedman et al., 2010; Meier et al., 2008). However, solving the overlapping group lasso problems, despite its convexity, becomes nontrivial. This is because the non-separability between groups intrinsically increases the dimensionality of the problem compared with the non-overlapping situation, as revealed in the study of Yan and Bien (2017). Proposed methods 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, introduced by Yuan et al. (2011). All these exact solvers of the problem involve expensive gradient evaluations when the overlapping becomes severe, limiting the applicability of the overlapping group lasso in many large-scale applications such as genomewide association studies (Section 5) or graphical model fitting problems (Cheng et al., 2017). For instance, Cheng et al. (2017) showed that overlapping group lasso, though as a natural choice for the problem, is infeasible even for moderate-size graphs, and they used a fast lasso approach (Tibshirani, 1996) to solve the graph estimation problem without theory. 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 penalty for the overlapping group lasso penalty. The approximation is formulated as a weighted non-overlapping group lasso penalty that respects the original overlapping patterns. The proposed penalty is shown to be the tightest separable relaxation of the original overlapping group lasso penalty with the family of $\ell_1/\ell_2$ norms. The non-overlapping nature of our penalty opens the door to efficient computation. We show that our proposed estimator leads to an equivalent estimator of the original overlapping group lasso estimator in that our method allows the same or better order estimation errors and achieves a minimax rate when the original estimator is minimax. We demonstrate the value of our proposed method using simulation examples and a prediction task on a breast cancer gene data set.

The rest of the paper is structured as follows: we introduce the proposed approximation method in Section 2 and prove that it is optimality as the tightest separable relaxation of the overlapping group lasso norm. We introduce our study of statistical estimation errors in Section 3. In Section 4 and Section 5, we present several empirical evaluations on both simulation data and the real breast cancer gene expression data. Section 6 concludes the paper with further discussions.

2. Methodology

Notation and Preliminaries. Throughout this paper, for any integer $z$, we use $[z]$ to denote the index set $\{1, \cdots, z\}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ if $a_n = o(b_n)$, write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for a sufficiently large $C > 0$, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ both hold. For a set $T$, we use $|T|$ to denote the cardinal number of $T$. For a matrix $A$, we use $A_T$ to denote the sub-matrix of $A$ consisting of the columns indexed by the set $T$, and use $A_{T,T}$ to denote the sub-matrix of $A$ consisting of both rows and columns indexed by the set $T$. When $A$ is a symmetric matrix, we use $\gamma_{\text{min}}(A)$ and $\gamma_{\text{max}}(A)$ to denote the smallest and the largest eigen-values of $A$. We introduce other notations in context, and summarize all notations used in Table 4 of Appendix A.

2.1 Group Lasso

Suppose that the parameters in our model are represented by a vector $\beta \in \mathbb{R}^p$ for which $\beta_j$ denotes the $j$-th element of $\beta$. Let $G = \{G_1, \cdots, G_m\}$ be the $m$ predefined groups for the $p$ parameters, that is, each group $G_q$ is a subset of $[p]$, and $\cup_{k \in [p]} G_k = [p]$. For each group $G_q$, we use $d(G)_q = |G_q|$ to denote the group size, and abbreviate it by $d_q$ when the group structure is clear in context. Given $G_q$, we let $\beta_{G_q}$ denote the a subvector of $\beta$ indexed by $G_q$. Additionally, we could have user-defined positive weights $w = \{w_1, \cdots, w_m\}$ associated with each group. The group lasso norm (Yuan and Lin, 2006) is defined as

$$\phi_G(\beta) = \sum_{g \in [m]} w_g \|\beta_{G_q}\|_2,$$

(1)

where the weights are usually set to be 1 or $\sqrt{d_q}$ in applications. When the group $G$ is clear, we may suppress $G$ and write $\phi_G(\beta)$ as $\phi(\beta)$.

In statistical estimation problems involving group selection, the group lasso norm is combined with an empirical loss function $L_n$, and the estimator $\hat{\beta}^G$ is solved by the following M-estimation problem.

$$\hat{\beta}^G = \arg \min_{\beta \in \mathbb{R}^p} \{L_n(\beta) + \lambda_n \phi(\beta)\},$$

(2)
When groups overlap, the group lasso leads to solutions whose nonzero support is contained in the complement to a union of groups (Jenatton et al., 2011a), with a possibility that some groups may have partial zeros. Such a pattern is desired in many problems such as graph structure recovery and structure learning (Zhao et al., 2009; Jacob et al., 2009; Cheng et al., 2017; Mohan et al., 2014; Tarzanagh and Michailidis, 2018). Yet another generalization of the group lasso aims to partially eliminate parts of some groups while maintaining full nonzero in other groups; this is called the latent overlapping group lasso (Jacob et al., 2009). This method is suitable in a set of different applications (Yan and Bien, 2017; Bien, 2018) and is not our focus in this paper.

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

\[
\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} ||\mu - \beta||^2 + \lambda_n \phi_G(\beta),
\]

(3)

Jenatton et al. (2011b) developed (4) as the dual of the proximal operator (3)

\[
\begin{align*}
\minimize_{\{\xi^g \in \mathbb{R} \}_{g=1}^m} & \left( \frac{1}{2} ||\mu - \sum_{g=1}^m \xi^g||^2 \right), \\
\text{s.t.} & \ ||\xi^g||_2 \leq \lambda_n w_g \quad \text{and} \quad \xi^g_j = 0 \quad \text{if} \quad j \notin G_g,
\end{align*}
\]

(4)

and proposed to solve the proximal operator of the group lasso penalty by a block coordinate descent (BCD) algorithm (Algorithm 1). The convergence of Algorithm 1 could be guaranteed by Proposition 2.7.1 in (Bertsekas, 1997). Yuan et al. (2011) and Chen et al. (2012) further smoothed proximal gradient procedure to improve its efficiency.

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

| Input: | \( G, \{ w_g \}_{g=1}^m, u, \lambda_n, \) |
| Requirement: | \( G, \{ w_g \}_{g=1}^m > 0, \lambda_n > 0. \) |
| Initialization: | Set \( \{ \xi^g \}_{g=1}^m = 0 \in \mathbb{R}^p. \) |
| Output: | \( \beta^* \) |

1: \textbf{while} stopping criterion not reached \textbf{do}
2: \quad \textbf{for all} \( g \in \{1, \cdots, m\} \) \textbf{do}
3: \quad \quad Calculate \( r^g = \mu - \sum_{h \neq g} \xi^h. \)
4: \quad \quad \textbf{if} \ ||r^g||_2 \leq \lambda_n w_g \textbf{ then} \quad \xi^g = 0 \textbf{ if} \ j \notin G_g \qquad \ \text{if} \ j \in G_g
5: \quad \quad \textbf{else} \quad \xi^g_j = \begin{cases} 0 & \text{if} \ j \notin G_g \\ \lambda w_g r^g_j / ||r^g||_2 & \text{if} \ j \in G_g \end{cases}
6: \quad \textbf{end if}
7: \quad \textbf{end for}
8: \textbf{end while}
9: \beta^* = u - \sum_{g=1}^m \xi^g.

Even if algorithms other than proximal+BCD are used to solve (2), the formulation of (3) and (4) still delivers insights about the bottleneck caused by the overlapping groups. The overlapping groups make the penalty non-separable between variables, compared to its non-overlapping counterpart. The duality between (3) and (4) shows that the overlapping group lasso problem has an intrinsic dimension equal to a \( \sum_{g \in [m]} d_g \)-dimensional separable problem. When the groups have a nontrivial proportion of overlapping variables, the computation of the overlapping group lasso becomes substantially more difficult, eventually prohibitive on large-scale problems. This issue significantly limits the applicability of the group lasso penalty in overlapping problems. Next, we introduce our non-overlapping approximation as a solution to this rectify.

2.2 Non-overlapping Approximation of the Overlapping Group Lasso

Given that the crux lies in the non-separability of parameters, the key to improving the computation is to introduce separable operators. We first use a special interlocking group structure to illustrate our idea. In an interlocking group
structure, groups are ordered and each group overlaps with its neighbors (see Figure 1a). We assume \( w_g \equiv 1 \) in this example for simplicity.

(a) Illustration of interlocking group structure.

(b) Illustration of 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 introduce a new group partition of the variables, as in Figure 1b, which precisely separates the intersections as separable groups. We denote the new groups as \( \mathcal{G} = \{ \mathcal{G}_1, \cdots, \mathcal{G}_m \} \), and in this particular case, \( m = 2m - 1 \). By triangular inequality we have

\[
\| \beta_{G_1} \|_2 \leq \| \beta_{G_1} \|_2 + \| \beta_{G_2} \|_2.
\]

With similar inequalities applied to each item, the overlapping group lasso norm based on \( G \) can be controlled by a reweighted non-overlapping group norm based on \( \mathcal{G} \) as

\[
\sum_{g \in [m]} \| \beta_{G_g} \|_2 \leq \sum_{g \in [m]} \tilde{h}_g \| \beta_{G_g} \|_2
\]

(5)

where \( \tilde{h}_g = 1 \) for odd numbers \( g \) and 2 for even numbers \( g \). That is, controlling the right side of (5) is sufficient to control the overlapping group norm on the left side. The advantage, however, is that the right-side norm is separable and allows for efficient optimization.

This idea is indeed applicable for general overlapping group structure \( G \). The pre-defined group \( G \) can be represented by a \( m \times p \) binary matrix \( G \), where \( G_{gj} = 1 \) if and only if the \( j \)-th variable belongs to the \( g \)-th group.

Our approximation starts from finding a partition of the \( p \) variables by grouping identical columns in \( G \), and we denote this new partition as \( \mathcal{G} \). To be more specific, if \( G_{j_1} = G_{j_2} \), then the \( j_1 \)-th variable and the \( j_2 \)-th variable are partitioned into the same group in \( \mathcal{G} \). Note that the groups \( \mathcal{G} \) are disjoint by this definition. All quantities such as number of groups and group sizes can be similarly defined for \( \mathcal{G} \). To distinguish \( G \) and \( \mathcal{G} \) while maintaining the connection between the quantities, we use normal letters, such as \( \{ g, d, m, w, G \} \), to denote quantities about the original group structure, and use calligraphic letters, such as \( \{ g, d, m, w, \mathcal{G} \} \), as notations about the group structure in the induced non-overlapping groups. For example, \( m \) is the number of groups in \( \mathcal{G} \) and \( g \in [m] \) is used as the index of groups in \( \mathcal{G} \).

Each group \( \mathcal{G}_g \in \mathcal{G} \) is a subset for some groups in \( G \). Define

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

In addition, we define the number of groups that contain \( \beta_j \) as \( h_j(G) = \sum_{g \in [m]} G_{gj} \), and define \( h_{\max}(G) = \max_{j \in [p]} h_j(G) \). Notice that all variables in the same \( \mathcal{G}_g \) have the same overlapping degree by definition, denoted as \( h_g = h_{(j \in \mathcal{G}_g)} \). In the previous example of the interlocking groups, we have \( h_g \) is either 1 or 2 and \( h_{\max} = 2 \) in \( G \).

Given positive weights \( w \) of \( G \), we proposed to set the weight \( w \) of \( \mathcal{G} \) by

\[
w_g = \sum_{g \in F(g)} w_g, \ g \in [m].
\]
With the new partition $\mathcal{G}$ and the new weights $w$, we fine the following norm as the separable alternative of overlapping group lasso norm.

$$
\psi_{\mathcal{G}}(\beta) = \sum_{g=1}^{m} w_{g} \| \beta_{G_{g}} \|_{2}.
$$

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

$$
\phi_{\mathcal{G}}(\beta) = \sum_{g=1}^{m} w_{g} \| \beta_{G_{g}} \|_{2} \leq \sum_{g=1}^{m} w_{g} \| \beta_{G_{g}} \|_{2} = \psi_{\mathcal{G}}(\beta).
$$

Our proposed penalty is essentially a weighted non-overlapping group lasso on $\mathcal{G}$. For illustration, Figure 2 shows the unit ball of these two norms based on $G_{1} = \{ \beta_{1}, \beta_{2} \}$ and $G_{2} = \{ \beta_{1}, \beta_{2}, \beta_{3} \}$. All singular points of the $\phi_{\mathcal{G}}$-ball (where exactly zero happens in optimization) are also singular points of the $\psi_{\mathcal{G}}$-ball.

Readers might notice that (7) can hold for other separable norms as well. For example, we can split all $p$ variables into separate groups and use a weighted lasso norm

$$
\sum_{j=1}^{p} \left( \sum_{g|\beta_{j} \in G_{g}} w_{g} \right) |\beta_{j}|
$$

as another upper bound of $\phi_{\mathcal{G}}$, which is the strategy of Cheng et al. (2017). So what is special about the proposed norm in (6)?

Intuitively, as can be seen from our construction procedure for $\mathcal{G}$, or Figure 2, our construction only introduces additional singular points for the norm when it is necessary to achieve separability. Compared with using a lasso upper bound, our procedure does not introduce redundancy in addition to the separability. Therefore, our approximation should achieve certain tightness. Next, we rigorously verify the above intuition in a formal way. Specifically, given any group structure $G$ and weights $w$, we define the $\ell_{q_{1}}/\ell_{q_{2}}$ norm of $\beta$ for any $0 \leq q_{1}, q_{2} \leq \infty$ as

$$
\| \beta_{G, w} \|_{q_{1}, q_{2}} = \left( \sum_{g \in [m]} w_{g} \| \beta_{G_{g}} \|_{q_{1}}^{q_{1}} \right)^{1/q_{1}} = \left\{ \sum_{g \in [m]} w_{g} \left( \sum_{j \in G_{g}} |\beta_{j}|^{q_{2}} \right)^{q_{1}/q_{2}} \right\}^{1/q_{1}}.
$$

This is a general class of norms that potentially includes most of the penalties that are commonly used. In particular, it includes the weighted lasso penalty. The following theorem shows that the proposed norm is the tightest separable relaxation of the original overlapping group lasso norm among all separable $\ell_{q_{1}}/\ell_{q_{2}}$ norms.
Theorem 1. Given the original weight $w$, let $\mathbb{G}$ be the space of all possible partitions of $[p]$, then there does not exist $0 \leq q_1, q_2 \leq \infty$, $G \in \mathbb{G}$, $\bar{w} \in [0, \infty)^p$, such that
\[
\begin{aligned}
\sum_{g \in [m]} w_g \| \beta_{G_g} \|_2 &\leq \| \beta_{\{G, \bar{w}\}} \|_{q_1, q_2} \leq \sum_{g \in [m]} w_g \| \beta_{G_g} \|_2 &\text{ for all } \beta \in \mathbb{R}^p \\
\| \beta_{\{G, \bar{w}\}} \|_{q_1, q_2} &< \sum_{g \in [m]} w_g \| \beta_{G_g} \|_2 &\text{ for some } \beta \in \mathbb{R}^p.
\end{aligned}
\tag{10}
\]

3. Statistical Properties

Using our proposed norm $\psi_{\mathbb{G}}$ as an alternative to the original group lasso norm $\phi_G$, would result in a different optimization problem compared to (2). However, as we shall justify in this section, the resulting estimator enjoys a similar theoretical estimation error. Our theoretical analysis will focus on (potentially high-dimensional) linear models, which is the setting of most previous group lasso theoretical analysis (Lounici et al., 2011; Huang and Zhang, 2010; Dedieu, 2019; Negahban et al., 2012). Specifically, define the linear model as
\[
Y = X\beta^* + \epsilon,
\tag{11}
\]
where $Y \in \mathbb{R}^{n \times 1}$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the data matrix, and $\epsilon \in \mathbb{R}^{n \times 1}$ is a random noise vector. The two estimators we want to study are the original overlapping group lasso estimator, defined as
\[
\hat{\beta}^G = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \| Y - X\beta \|_2^2 + \lambda_n \phi_G(\beta)
\tag{12}
\]
and the regularized estimator by our approximation norm, defined as
\[
\hat{\beta}^\psi = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \| Y - X\beta \|_2^2 + \lambda_n \psi_{\mathbb{G}}(\beta).
\tag{13}
\]
As a remark, our target is not to argue that (13) is an approximate optimization problem of (12). Instead, we will show that the two classes of estimators defined by (12) and (13) are statistically equivalent in their estimation errors when the $\lambda_n$ values are properly chosen, which may differ for the two estimators. Therefore, from the standpoint of statistical estimation, the class of estimators introduced by our method (13) is preferable because it gives a similar estimation accuracy with significantly better computational efficiency. Our theoretical results include two parts. In the first part, we show that under reasonable assumptions, the estimation error bound of (13) is no larger than that of (12). In the second part, we give the minimax error rate of overlapping sparse group regression problem and show that both (12) and (13) are minimax optimal under additional requirements of the group structures.

3.1 Estimation Error Bounds

The theoretical analysis involves more quantities about group structures. For a group index set $I \subseteq [m]$, we define $G_I$ as $\bigcup_{g \in I} G_g$. Additionally, we define $G_I^c$ as a new group membership matrix corresponding to $G_I$.

Given a group structure $G$ and a group index set $I$, we introduce two parameter spaces $M(I) = \{ \beta \in \mathbb{R}^p | \beta_j = 0 \text{ for all } j \in (G_I)^c \}$ and $M^I(I) = \{ \beta \in \mathbb{R}^p | \beta_j = 0 \text{ for all } j \notin G_I^c \}$. We further use $\beta_{M(I)}$ to denote the projection of $\beta$ onto $M(I)$. Additionally, we introduce the following definitions.

Definition 1. Given $\beta \in \mathbb{R}^p$, the support set of $\beta$ is denoted as
\[
\text{supp}(\beta) = \{ j \in [p] | \beta_j \neq 0 \},
\]
the group support set of $\beta$ is defined as
\[
S(\beta) = \{ g \in [m] | G_g \cap \text{supp}(\beta) \neq \emptyset \},
\]
and the augmented group support set is defined as
\[
\overline{S}(\beta) = \{ g \in [m] | G_g \cap G_{S(\beta)} \neq \emptyset \}.
\]
We use $s = |\text{supp}(\beta)|$ to denote the number of non-zero elements, use $s_g = |S(\beta)|$ to denote the number of non-zero groups, and use $\overline{s}_g = |\overline{S}(\beta)|$ to denote the number of groups in the augmented group support set. For simplicity, we also abbreviate $S(\beta)$ by $S$ and $\overline{S}(\beta)$ by $\overline{S}$. We first introduce the distributional assumptions.
Assumption 1. Under model (11), we further assume

1. (Sub-Gaussian noises) The coordinates of $\varepsilon$ are i.i.d zero mean sub-Gaussian random variable denote with parameter $\sigma$. That is, there exists $\sigma > 0$ such that

$$E[|e^{t\varepsilon}|] \leq e^{\frac{t^2\sigma^2}{2}}, \quad \text{for all } t \in \mathbb{R}.$$ 

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

3. (Pattern of group structure) The pre-defined group structure $G$ satisfies $d_{\max} \leq c_2 n$ for some constant $c_2 > 0$ and $\log m \ll n$.

Theorem 2. Given $G$ and its induced $\mathcal{G}$ according to our method, define $h^g_{\min} = \min_{j \in G_g} h_j$, $d_{\max} = \max_{g \in [m]} d_g$, and $d'_{\max} = \max_{g \in [m]} d_g$. Under Assumption 1, suppose $\hat{\beta}^G$, $\hat{\beta}^\mathcal{G}$ are defined in (12) and (13). For $\delta \in (0, 1)$ which is a scalar that might depend on $n$ and some constant $c_3 > 0$, we have the following results.

1. Suppose $\beta^*$ satisfies the following group sparsity

$$\pi_g(\beta^*) \lesssim \frac{n}{\log m + d_{\max}} \min_{g \in [m]} \frac{(w^2 g h^g_{\min})}{\max_{g \in [m]} (w^2 g h^g_{\max})}. \quad (14)$$

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

$$\|\hat{\beta}^G - \beta^*\|_2 \lesssim \sigma^2 . \frac{\left(\sum_{g \in \mathcal{G}} w^2 g \right)}{\min_{g \in [m]} \left(w^2 g h^g_{\min}\right)} . \left(\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta\right). \quad (15)$$

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

2. Suppose $\beta^*$ satisfies the following group sparsity

$$\pi_g(\beta^*) \lesssim \frac{n}{\log m + d_{\max}} \min_{g \in [m]} \frac{(w^2 g)}{\max_{g \in [m]} (w^2 g)} . \quad (16)$$

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

$$\|\hat{\beta}^\mathcal{G} - \beta^*\|_2 \lesssim \sigma^2 . \frac{\sum_{g \in \mathcal{G}} w^2 g}{\min_{g \in [m]} \left(w^2 g\right)} . \left(\frac{d_{\max}}{n} + \frac{\log m}{n} + \delta\right). \quad (17)$$

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

The error bound (15) includes the non-overlapping group lasso error bound as a special case. When the groups in $G$ are disjoint, the reduced form of (15) matches the bounds studied in (Huang and Zhang, 2010; Negahban et al., 2012; Lounici et al., 2011; Wainwright, 2019). The main difference is that in the overlapping situation, we have to include the measurement of overlapping degree and the sparsity requirement is extended to the augmented groups. Moreover, the group sparsity requirements in (14) and (16) are slightly different. However, in many common situations, they are essentially the same. For example, in the previous interlocking group example, they match each other. In this case, our requirement is that the cardinality of the augmented group support is much smaller than $\frac{n}{\log m + d_{\max}}$.

Moreover, we show the relation between the two bounds in (15) and (17) under a mild assumption under which both (14) and (16) match each other.

Assumption 2. Suppose that $\max_{g \in \mathcal{G}} |F^{-1}(g)| \leq c_3$ for some constant $c_3$. Furthermore, $d_{\max} \asymp d'_{\max}$, and $m \asymp m$.
Corollary 1. Under Assumption 2, we have
\[
\sum_{g \in F^{-1}(S)} w_g g^2 \leq \left( \sum_{g \in S} w_g g^2 \right) \cdot h_{\max}(G_S) \cdot \left( \frac{d_{\max}}{n} + \frac{\log m}{n} + \delta \right).
\]
This means that the error bound of \( \hat{\beta}^G \) in (15) is also an upper bound for the error of \( \hat{\beta}^G \).

3.2 Lower Bound of Estimation Error

The previous section compares the estimation error upper bounds of the two estimators. To make the result stronger, we study the minimax rate of the estimation error in the linear regression model with overlapping group sparsity. We will focus on the following class of group-wise sparse vectors.

\[
\Omega(G, s_g) = \left\{ \beta : \sum_{G_g \in G} \mathbb{1}_{\{\|\beta_{G_g}\|_2 \neq 0\}} \leq s_g \right\}
\]

(18)

We need to further introduce additional constraints on the parameter space that are similar to the Assumptions in (Cai et al., 2022).

Assumption 3. Under model (11), further assume

1. X is a random design data matrix, and all rows of X are i.i.d centered sub-gaussian distributed with covariance matrix \( \Theta \). We further assume that \( \gamma_{\max}(\Theta) \leq c_4 \) for some constant \( c_4 \).

2. The \( m \) predefined groups of \( G \) come with equal group size \( d \), \( m \ll p, d \ll \log(p) \).

Theorem 3. (Lower bound of estimation error) Under Assumption 3, we have

\[
\inf_{\hat{\beta}} \sup_{\beta \in \Omega(G, s_g)} E\|\hat{\beta} - \beta\|^2 \geq \frac{\sigma^2 (d + \log(m/s_g))}{n}.
\]

(19)

Theorem 2 and Theorem 3 together indicate that the two estimators achieve the minimax error rate, as formulated in the following corollary. Therefore, in this case, they are precisely statistically equivalent.

Corollary 2. Under the conditions of Theorem 2 and Theorem 3, if \( m \approx m \) and \( h_{\max}(G_S) \approx 1 \), then both \( \hat{\beta}^G \) and \( \hat{\beta}^G \) achieves the minimax estimation rate specified in (19).

4. Simulation

In this section, we use synthetic data to compare overlapping group lasso and our proposed group lasso approximation. As another reference for separable but loose approximation, we include the weighted lasso (8). We use the SLEP package (Liu et al., 2009) in MATLAB for computation. Although the package may not be the most efficient for implementing non-overlapping group lasso and lasso optimization, it is the most efficient one we know for overlapping group lasso (Yuan et al., 2011), and it provides the same implementation framework for all three methods, thus ensuring a fair comparison.

For each configuration of the data-generating model, we evaluate the performance in two aspects:

- **Regularization path computing time.** We first use a line search to find approximately the smallest \( \lambda \) that selects no variables, denoted by \( \lambda_{\max} \), and also the maximum \( \lambda \) that selects the full model, denoted by \( \lambda_{\min} \). Then we pick 100 values in log-scale in \( [\lambda_{\min}, \lambda_{\max}] \). The full regularization path is computed on the 100 \( \lambda \) values and the computing time of the regularization path is recorded as a performance metric. The computing time evaluation mimics the most practical situation where the whole regularization path is solved for tuning purposes.

- **Estimation error:** From the computed regularization path, we take the smallest relative estimation error, defined as \( \|\hat{\beta} - \beta^*\|_2/\|\beta^*\|_2 \), as the estimation error for the method. This is the measure of the ideally tuned performance.

In the settings, the group weight of the overlapping group lasso is set to be \( w_g = \sqrt{d_g} \), as is usually used in practice. We generate 100 independent replicates for each configuration and report the average result.
4.1 Interlocking group structure

In the first set of experiments, we evaluate the performances based on interlocking group structure of Figure 1a. We set \( m \) interlocked groups with \( d \) variables in each group and \( 0.2d \) variables in each intersection. For example, with \( m = 5 \) and \( d = 10 \), we have \( G_1 = \{1, \ldots, 10\} \), \( G_2 = \{8, 9, \ldots, 17\} \), \( \ldots \), \( G_{10} = \{33, 34, \ldots, 42\} \).

The data matrix \( X \) is generated from a Gaussian distribution \( N(0, \Theta) \), where \( \Theta \) is specified with correlations matching the group structure. We first construct the following matrix \( \tilde{\Theta} \)

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

\( \Theta \) is then taken to be the projection of \( \tilde{\Theta} \) on the set of positive definite matrices with with minimum eigenvalue of 0.1.

We generate each coordinate of \( \beta^* \) by first sampling from \( N(10, 16) \), assigning negative and positive signs randomly, and then randomly setting \( \lfloor 0.9m \rfloor \) groups to be zero. The response variable \( Y \) is generated by \( Y = X \beta^* + \epsilon \), where \( \epsilon \sim N(0, \sigma^2) \) and \( \sigma^2 \) is set to give the signal-to-noise ratio of 3.

Figure 3 shows the average computation time (sec.) with its 95% confidence interval (CI) in logarithmic scale and average estimation error with its 95% CI under different configurations. As expected, solving the proposed approximation method is much faster than solving the original overlapping group lasso problem by a factor of 20–150, and lasso approximation is even faster than our group lasso approximation. As expected, increasing the sample size, number of variables or group size would increase the computational time. With these increase of problem scale, the computational advantage of method is more significant.

Under all configurations, the lasso approximation gives much larger estimation errors compared with the original overlapping group lasso solution. In contrast, our proposed method achieves very similar accuracy to that of the overlapping group lasso estimator. This empirically verifies our theoretical results.

In summary, the estimation from our proposed approximation gives similar statistical estimation performance to the original overlapping group lasso estimation while rendering significant improvement in computation. The weighted lasso is computationally efficient but gives very poor estimation, thus it is not a competitive approximation of the original problems.

4.2 Group structures based on real-world gene pathways

In this set of experiments, we use real-world designs to evaluate the methods. Specifically, we use the gene canonical pathways from the Molecular Signatures Database (Subramanian et al., 2005) as the group structures. Each gene pathway is a collection of genes that share common biological features. Therefore, the sets of gene pathways can be used as gene group structures. Such group structures can provide helpful and interpretable ways to improve prediction performance in practice (Yuan et al., 2011; Jacob et al., 2009; Chen et al., 2012). We use five major gene pathway sets as five different group structures in our evaluation, as summarized in Table 1.

The design matrix \( X \) in this example is the gene expression data in (Van De Vijver et al., 2002) as our design matrix. The data can be founded in R package "breastCancerNKI" (Schroeder et al., 2021). The design matrix contains 295 observations and 24481 genes. Given each gene pathway set, we filter all the genes that are undefined in the pathway sets.

The data-generating procedure for \( \beta^* \) and \( y \) remains almost the same as before, except that here we use a much sparser model because of the small sample size. Specifically, we randomly sample 0.05m active groups and set the coefficients in other groups to zero. We set the group weights in overlapping group lasso to be \( \sqrt{d_j} \) and use the proposed weight modification for our approximation. Figure 4 displays the computing time and estimation error results for the five pathway group structures. The high-level message remains the same as before: the overlapping group lasso is extremely slow in this problem. Both our proposed group lasso approximation and the lasso approximation could substantially reduce the computing time by a factor of 10–60. Meanwhile, our proposed group lasso approximation delivers very similar statistical estimation performance to the original overlapping group lasso estimator, whereas the lasso approximation fails to effectively leverage the group information, delivering worse estimation performance.
Figure 3: Regularization path computing time comparison and estimation error under different configurations. (a) Varying sample size $n$ when fixing $m = 400$ and $d = 20$; (b) Varying number of groups $m$ when fixing $n = 800$ and $d = 20$; (c) Varying group size $d$ when fixing $n = 800$ and $m = 200$.

Figure 4: Comparison of performances on different group structures.
Table 1: Summary of the collections of gene pathways, including number of groups \( (m) \), average group size \( (\bar{d}) \), average overlapping degree \( (\bar{h}) \) and the number of genes after filtering \( (p) \) are also summarized in the table.

| Pathways | Information | \( m \) | \( d \) | \( h \) | \( p \) |
|----------|-------------|--------|--------|--------|--------|
| BioCarta | BioCarta is a database of gene interaction models. The database contains high-quality images of several cellular signaling and interaction pathways, and each diagram is fully hyperlinked to products and information pages about individual genes. | 292    | 15.4   | 3.25   | 1129   |
| PID      | The Pathway Interaction Database (PID) is a freely available collection of curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes (Schafer et al., 2008) | 196    | 38.51  | 3.28   | 2297   |
| KEGG     | The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an integrated database resource for biological interpretation of genome sequences and other high-throughput data (Kanehisa et al., 2015). | 186    | 58.48  | 2.58   | 4207   |
| Wiki     | WikiPathways captures the collective knowledge represented in biological pathways. By providing a database in a curated, machine readable way, omics data analysis and visualization is enabled. WikiPathways and other pathway databases are used to analyze experimental data by research groups in many fields (Slenter et al., 2017). | 712    | 38.17  | 4.35   | 6242   |
| Reactome | The Reactome Knowledgebase, an Elixir core resource, provides manually curated molecular details across a broad range of physiological and pathological biological processes in humans, including both hereditary and acquired disease processes (Gillespie et al., 2021). | 1615   | 45.31  | 8.78   | 8331   |

5. Application to Breast Cancer Data

The gene expression data set we used in Section 4.2 also contains additional attributes for the 295 observations. In particular, the observations have labeled the status of breast cancer tumors (79 metastatic and 216 non-metastatic). Previous studies (Yuan et al., 2011; Jacob et al., 2009; Chen et al., 2012) showed that some of the canonical gene pathways might provide helpful grouping information in predicting the tumor type from gene expression data. In this example, we want to examine the pathway sets in Table 1. Specifically, we use regularized logistic regression to build a prediction model, with the overlapping group lasso penalty, our proposed group lasso approximation, and the standard lasso penalty. In this case, the lasso penalty does not take the pathway information. Hence, comparing the performance of the overlapping group lasso and the lasso approximation would provide some empirical insights into whether a specific gene pathway set contains predictive grouping information for breast cancer tumor type. Please note this example is designed to demonstrate our approximation to the overlapping group lasso. We aim to examine, rather than to claim, the effectiveness of gene pathways.

Similar to the setting in (Lee and Xing, 2014), we randomly assign 200 samples to the training data and 95 samples to the testing data. All three 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 computation time. Again, the experiment is repeated 100 times independently. Table 2 shows the average computing time and Table 3 shows the average AUC.

Table 2: Solving time (in sec) under different pathway databases

| Database     | Method            | Overlapping group lasso | Lasso      | Proposed group lasso approximation |
|--------------|-------------------|-------------------------|------------|------------------------------------|
| BioCarta     |                   | 653.94                  | 25.39      | 65.52                              |
| KEGG         |                   | 1239.71                 | 43.08      | 90.86                              |
| PID          |                   | 2219.12                 | 117.65     | 270.30                             |
| WikiPathways |                   | 4865.09                 | 196.08     | 446.41                             |
| Reactome     |                   | 10483.99                | 348.84     | 990.18                             |
Table 3: Predictive AUC results of the three methods under different pathway databases. The * indicates that the AUC is significantly different from the AUC of the overlapping group lasso by a paired t-test.

| Database    | Method                           | Overlapping group lasso | Lasso       | Proposed group lasso approximation |
|-------------|----------------------------------|-------------------------|-------------|------------------------------------|
| BioCarts    | Overlapping group lasso          | 0.7239                  | 0.6993*     | 0.7247                             |
| KEGG        | Overlapping group lasso          | 0.7071                  | 0.6836*     | 0.7098                             |
| PID         | Overlapping group lasso          | 0.7513                  | 0.7138*     | 0.7382*                            |
| WikiPathways| Overlapping group lasso          | 0.6979                  | 0.7261*     | 0.6838*                            |
| Reactome    | Overlapping group lasso          | 0.6727                  | 0.7288*     | 0.6911*                            |

From these results, we can see that the pathway sets in BioCarts, KEGG and PID databases seem predictive, while the other two are not very helpful. The proposed approximation method results in similar predictive performance with the overlapping group lasso while the computation is much faster.

6. 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 reweighting the new groups according to the original overlapping pattern. The penalty is the tightest separable relaxation of the overlapping group lasso among all $\ell_q/\ell_q$ norms. We have also shown that for linear problems, the proposed estimator is statistically equivalent to the original overlapping group lasso estimator. However, the proposed method is significantly faster in computation as compared to the original overlapping group lasso, rendering the proposed method applicable in large-scale problems.

Several interesting directions could be considered for future research. First, the theoretical study in this paper focuses on the estimator errors of the two estimators. Correct support recovery can be proved, assuming a proper group sparsity pattern. However, without assuming such a model, the output of the two estimators might not be perfectly aligned. If a group-union zero pattern is still strictly required, certain post-processing of our estimator is needed. The in-depth study of this step would be of great practical interest. Second, when a group-union nonzero pattern is desirable, the latent overlapping group lasso (Jacob et al., 2009), as a popular choice, 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.

Acknowledgments

The work is supported in part by the NSF grant DMS-2015298 and the 3-Caveliers award from the University of Virginia.

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.

J. Bien. Graph-guided banding of the covariance matrix. Journal of the American Statistical Association, 2018.
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.

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. The Annals of Applied Statistics, 6(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. Rothefs, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Raguneau, 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.

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

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.

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

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.

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.

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.

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

P. Ravikumar, J. LaFerty, 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.

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.

D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Melius, E. Cirillo, S. L. Coort, D. Digles, F. Ehrhart, P. Giesbertz, M. Kalafi, 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.

L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. Advances in Neural Information Processing 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, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. Annals of Statistics, 37(6A):3468–3497, 2009.
A. Notation summary

Table 4: Major notations used in this paper.

| Indices:                      | Parameter index set \{1, ..., p\} |
|-------------------------------|----------------------------------|
| \[m\]                         | group index set \{1, ..., m\}    |
| \[n\]                         | sample index set \{1, ..., n\}   |
| \[p\]                         | parameter index set \{1, ..., p\}|
| \(G_g\)                       | index set of \(g^{th}\) group   |
| \(G_S\)                       | collection of non-zero groups, \(\bigcup_{g \in S} G_g\)|
| \(G_{\overline{S}}\)         | \(\bigcup_{g \in \overline{S}} G_g\) |
| \(\beta_j\)                  | the \(j^{th}\) element of \(\beta\) |
| \(\beta_{G_g}\)              | subvector of \(\beta\) indexed by \(G_g\) |
| \(\beta_{M(S)}\)             | projection of \(\beta\) onto \(M(S)\) |
| \(A_{1:T}\)                  | sub-matrix consisting of the columns indexed by \(T\) |

| Parameters:                   |                                |
|-------------------------------|--------------------------------|
| \(H\)                         | a diagonal matrix, diag(\(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_p}\)) |
| \(G\)                         | group structure matrix, \(G_{gj} = 1\) iff \(\beta_j \in G_g\) |
| \(d_g\)                       | group size, \(d_g = \sum_{j \in [p]} G_{gj}\) |
| \(d_{\text{max}}\)            | maximum group size, \(d_{\text{max}} = \max_{g \in [m]} d_g\) |
| \(h_j\)                       | overlap degree, \(h_j = \sum_{g \in [m]} G_{gj}\) |
| \(h_{\text{max}}^g\)          | maximum overlap degree in \(G_g\), \(h_{\text{max}}^g = \max_{j \in G_g} h_j\) |
| \(h_{\text{min}}^g\)          | minimum overlap degree in \(G_g\), \(h_{\text{min}}^g = \min_{j \in G_g} h_j\) |
| \(h_{\text{max}}\)            | maximum overlap degree, \(h_{\text{max}} = \max_{g \in [p]} h_j\) |
| \(h_g\)                       | overlap degree of \(G_g\), \(h_{\text{max}}^g = \max_{j \in G_g} h_j\) |
| \(\sigma\)                    | parameter in the sub-Gaussian distribution |
| \(\sigma_g\)                  | number of non-zero groups \(|S|\) |
| \(\kappa\)                    | parameter controls convexity |

| Definitions:                  |                                |
| \(\phi(\beta)\)              | group lasso norm, \(\sum_{g \in [m]} w_g \|\beta_{G_g}\|_2\) |
| \(\phi^*(\beta)\)            | dual norm of \(\phi(\beta)\), \(\max_{g \in [m]} \frac{1}{w_g} \|H \beta_{G_g}\|_2\) |
| \(F(g) \subseteq [m]\)       | overlapping groups which include the variables in \(G_g\) |
| \(F^{-1}(g) \subseteq [m]\)  | non-overlapping groups that were partitioned from \(G_g\) |
| \(\|\beta_{(G,w)}\|_{2q_1,2q_2}\) | \(\ell_{2q_1,2q_2}\) norm, \(\sum_{g \in [m]} w_g \left( \sum_{j \in G_g} |\beta_j|^{2q_2} \right)^{\frac{q_1}{q_2}}\) |
| \(\text{supp}(\beta)\)       | support set, \(\{j \in [1, \cdots, p] | \beta_j \neq 0\}\) |
| \(S(\beta)\)                 | group support set, \(\{g \in [1, \cdots, m] | G_g \cap \text{supp}(\beta) \neq \emptyset\}\) |
| \(S(G)\)                     | \(\{g \in [1, \cdots, m] | G_g \cap G_{\overline{S}(\beta)} \neq \emptyset\}\) |
| \(M(S)\)                     | \(\{\beta \in \mathbb{R}^p | \beta_j = 0 \text{ for all } j \in (G_S)^c\}\) |
| \(M(\overline{S})\)          | \(\{\beta \in \mathbb{R}^p | \beta_j = 0 \text{ for all } j \in G_{\overline{S}}\}\) |

B. Additional Theoretical Results

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

**Proposition 1.** The sharp upper bound for \(\phi^*\) (the dual norm of overlapping group lasso penalty in (1)) is

\[
\max_{g \in [m]} \frac{1}{w_g} \| (H \beta)_{G_g} \|_2, \tag{20}
\]

where \(H\) is a diagonal matrix with diagonals \(\left(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_p}\right)\).

Developing the upper bound of dual norm of overlapping group lasso regularizer is an essential step as the dual norm affects the choice of \(\lambda_n\) and ultimately affects the results on the estimation error bound.
**Assumption 4.** Under model (11), we assume

1. (Sub-Gaussian noises) The coordinates of $\varepsilon$ are i.i.d zero mean sub-Gaussian random variable denote with parameter $\sigma$, which means that there exist $\sigma > 0$ such that
   \[ E[\varepsilon^2] \leq e^{\sigma^2t^2}/2, \quad \text{for all } t \in \mathbb{R}. \]

2. (Group normalization condition) $\sqrt{\gamma_{\max}}(X_{aj}X_{aj}^T/n) \leq c$ for some constant $c$.

3. (Restricted strong convexity condition) For some $\kappa > 0$,
   \[ \frac{\|X(\tilde{\beta} - \beta^*)\|^2}{n} \geq \kappa \|\tilde{\beta} - \beta^*\|^2, \quad \text{for all } \tilde{\beta} \in \left\{ \beta \mid \phi\left((\beta - \beta^*)_{M+(S)}\right) \leq 3\phi\left((\beta - \beta^*)_{M(S)}\right) \right\}. \]

**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 (Lounici et al., 2011; Huang and Zhang, 2010; Dedieu, 2019; Negahban et al., 2012; Wainwright, 2019). Additionally, the restricted curvature conditions that has been well discussed by Wainwright (2019). The curvature $\kappa$ in Assumption 4 is a parameter measure the convexity. Generally speaking, the restricted curvature conditions states 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 the estimates. Restricted curvature conditions are milder that 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 4, Theorem 4 gives $\ell_2$ norm estimation upper error bound for overlapping group lasso.

**Theorem 4.** Define $h^g_{\min} = \min_{j \in G_g} h_j$, $d_{\max} = \max_{g \in [m]} d_g$, and $d'_{\max} = \max_{g \in [m]} d_g$. Suppose Assumption 4 holds, for any $\delta \in [0, 1]$,

1. with $\lambda_n = \frac{8c\sigma}{\min_{g \in [m]} w^2_g h^g_{\min}} \sqrt{d_{\max} \log 5/n + \log m/n + \delta}$, the following bound hold for $\hat{\beta}^G$ in (12)
   \[ \left\|\hat{\beta}^G - \bar{\beta}\right\|^2 \leq \frac{\sigma^2}{\kappa^2} \frac{\sum_{g \in S} w^2_g}{\min_{g \in [m]} (w^2_g h^g_{\min})} \cdot \left(\frac{d_{\max} \log 5/n + \log m/n + \delta}{n}\right). \] \quad (21)

2. with $\lambda_n = \frac{8c\sigma}{\min_{g \in [m]} w^2_g} \sqrt{d'_{\max} \log 5/n + \log m/n + \delta}$, the following bound hold for $\hat{\beta}^g$ in (13)
   \[ \left\|\hat{\beta}^g - \bar{\beta}\right\|^2 \leq \frac{\sigma^2}{\kappa^2} \frac{\sum_{g \in \mathcal{F}^{-1}(S)} w^2_g}{\min_{g \in [m]} (w^2_g)} \cdot \left(\frac{d'_{\max} \log 5/n + \log m/n + \delta}{n}\right). \] \quad (22)

Following the framework in (Negahban et al., 2012; Wainwright, 2019), we further study the applicability of the restricted curvature conditions in terms of random design matrix. Given a group structure $G$, Theorem 4 is developed based on the assumption that the fixed design matrix $X$ satisfies the restricted curvature condition. In practice, it is difficult to verify that a given design matrix $X$ satisfies this condition. Indeed, developing methods to “certify” design matrices in this way is one line of on-going research (Wainwright, 2019). However, it is possible to give high-probability results based on following assumptions.

**Theorem 5.** Under Assumption 1, we have

1. With probability at least $1 - e^{-c' n}$, $\max_{g \in [m]} \sqrt{\gamma_{\max}(X_{aj}X_{aj}^T/n)} \leq c$ for some constants $c, c' > 0$, as long as $\log m = o(n)$.
2. The restricted strong convexity condition, which is
\[
\|X \left( \frac{\tilde{\beta} - \beta^*}{n} \right) \|_2^2 \geq \kappa \| \beta - \beta^* \|_2^2, \quad \text{for all} \quad \tilde{\beta} \in \{ \beta \mid \phi \left( (\beta - \beta^*)_{M+}^+ \right) \leq 3\phi \left( (\beta - \beta^*)_{M+}^\perp \right) \}.
\]
hold with probability at least \(1 - \frac{e^{-\frac{\kappa n}{32}}}{1 - e^{-\frac{\kappa n}{64}}}\) for some constant \(\kappa > 0\).
C. Proofs

C.1 Proof of Theorem 1

Lemma 6. If there exists a norm $||\cdot |\tilde{G}, \tilde{w}||_{q_1, q_2}$ such that (10) hold, then for any $g \in [m]$, there exists a $\tilde{g} \in [\tilde{G}]$ such that $\mathcal{G}_g \subseteq \tilde{G}_{\tilde{g}}$.

Lemma 7. If there exists a norm $||\cdot |\tilde{G}, \tilde{w}||_{q_1, q_2}$ such that (10) hold, then for any $\tilde{g} \in [\tilde{G}]$, there exists a $g \in [m]$ such that $\tilde{G}_{\tilde{g}} = \mathcal{G}_g$.

Proof

Based on Lemma 6 and Lemma 7, if there exists $||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2}$ which satisfies (10), then $\tilde{G} = \mathcal{G}$. Therefore, the difference in between $||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2}$ and our proposed norm could be only introduced by different weights or different values of either $q_1$ or $q_2$.

For any $\beta$ with non-zero elements only in $g$-th group $\mathcal{G}_g$,

$$
\sum_{g \in [m]} w_g ||\beta_{\mathcal{G}_g}||_2 = \left( \sum_{g \in [m]} w_{\tilde{g}} \right) \frac{1}{\sqrt{q_1 q_2}} \leq ||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2} \leq \sum_{g \in [m]} w_g ||\beta_{\mathcal{G}_g}||_2.
$$

(23)

It implies that $(\tilde{w}_g ||\beta_{\mathcal{G}_g}||_2) \frac{1}{\sqrt{q_1 q_2}} = w_g ||\beta_{\mathcal{G}_g}||_2$. By setting one element in $\mathcal{G}_g$ to 1, and other elements to 0, we have $\tilde{w}_g = w_g$. Because this hold for any group in $\mathcal{G}$, $\tilde{w} = w$.

From (23), we have that $(w_g ||\beta_{\mathcal{G}_g}||_2) \frac{1}{\sqrt{q_1 q_2}} = w_g ||\beta_{\mathcal{G}_g}||_2$ for any $\beta$ with non-zero elements only in $\mathcal{G}_g$. It implies that $q_1 = 1, q_2 = 2$. Therefore, the existed norm $||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2} = \sum_{g \in [m]} w_g ||\beta_{\mathcal{G}_g}||_2$, which does not meet the second condition in (10).

C.1.1 Proof of Lemma 6

Proof

Since $\tilde{G} \in \mathcal{G}$, for an arbitrary $g \in [m]$, if $\mathcal{G}_g \nsubseteq \tilde{G}_{\tilde{g}}$ for any $\tilde{g}$, then we can find the smallest set $T$, such that $\mathcal{G}_g \subseteq \bigcup_{\tilde{g} \in T} \tilde{G}_{\tilde{g}}$.

Denote $T = \{t_1, t_2, \cdots, t_{|T|}\}$, choose $\beta_j \in (\mathcal{G}_g \bigcap \tilde{G}_{t_1})$ and $\beta_k \in (\mathcal{G}_g \bigcap \tilde{G}_{t_2})$. Since $\beta_j$ and $\beta_k$ are both in $\mathcal{G}_g$, if an original group contains $\beta_j$, then this original group also contains $\beta_k$. Consider $\beta$ where $\beta_j$ and $\beta_k$ are the only non-zero elements, we have

$$
\sum_{g \in [m]} w_g ||\beta_{\mathcal{G}_g}||_2 = \left( \sum_{\{g, \beta_j \in \mathcal{G}_g\}} w_g \right) \sqrt{\beta_j^2 + \beta_k^2} \leq ||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2}
$$

$$
\leq \sum_{g \in [m]} w_g ||\beta_{\mathcal{G}_g}||_2 = \left( \sum_{\{g, \beta_j \in \mathcal{G}_g\}} w_g \right) \sqrt{\beta_j^2 + \beta_k^2},
$$

which is equivalent to

$$
||\beta_{\{\tilde{G}, \tilde{w}\}}||_{q_1, q_2} = (w_{t_1}(||\beta_j||_2 q_1 + w_{t_2}(||\beta_k||_2 q_1)) \frac{1}{\sqrt{q_1 q_2}} = \left( \sum_{\{g, \beta_j \in \mathcal{G}_g\}} w_g \right) \sqrt{\beta_j^2 + \beta_k^2},
$$

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

$$
\{ \begin{array}{ll}
\beta_j = \beta_k = 1, & \beta_{\{[j] \setminus (j, k)\}} = 0 \quad \text{if } (w_{t_1} + w_{t_2}) \frac{1}{\sqrt{q_1 q_2}} \neq \sqrt{2} \left( \sum_{\{g, \beta_j \in \mathcal{G}_g\}} w_g \right) \\
\beta_j = 2, \beta_k = 1, & \beta_{\{[j] \setminus (j, k)\}} = 0 \quad \text{if } (w_{t_1} + w_{t_2}) \frac{1}{\sqrt{q_1 q_2}} = \sqrt{2} \left( \sum_{\{g, \beta_j \in \mathcal{G}_g\}} w_g \right)
\end{array}
$$

18
we have a contradiction. Therefore, we show that if there exists a norm \( \| \cdot \|_{q_1,q_2} \), then each group in \( G \) is a union of groups in \( \mathcal{G} \).

### C.1.2 Proof of Lemma 7

**Proof** Lemma 6 implies that each group in \( G \) is a union of groups in \( \mathcal{G} \). Without loss of generality, suppose that for \( \tilde{g} \in [|G|] \), there exists \( V \subseteq [m] \), such that \( G_{\tilde{g}} = \bigcup_{g \in V} \mathcal{G}_g \) and \( |V| > 1 \). Denote \( V = \{ v_1, \cdots, v_{|V|} \} \), then one of the following cases must happen:

**Case 1:** \( \exists a \in [m] \), s.t. \( (\mathcal{G}_{v_1} \cup \mathcal{G}_{v_2}) \subseteq G_a \).

**Case 2:** \( \exists a \in [m] \), s.t. \( (\mathcal{G}_{v_1} \cup \mathcal{G}_{v_2}) \subseteq G_a \).

Under case 1, if only \( \mathcal{G}_{v_1} \) and \( \mathcal{G}_{v_2} \) have non-zero values in \( \beta \), we have

\[
\sum_{g \in [m]} \| \beta_{\mathcal{G}_g} \|_2 = \left( \sum_{g \in F(v_1)} w_g \right) \| \beta_{\mathcal{G}_{v_1}} \|_2 + \left( \sum_{g \in F(v_2)} w_g \right) \| \beta_{\mathcal{G}_{v_2}} \|_2 \leq \| \beta_{G_{\tilde{g}}} \|_{q_1,q_2}
\]

which is equivalent to

\[
w_{v_1} \sqrt{\beta_{\mathcal{G}_{v_1}}} + w_{v_2} \sqrt{\beta_{\mathcal{G}_{v_2}}} = w_{v_1}^{\frac{1}{q_1}} \left( \sum_{j \in G_{\tilde{g}}} |\beta_j|^{q_2} \right)^{\frac{1}{q_2}} = w_{v_2}^{\frac{1}{q_2}} \left( \sum_{j \in (\mathcal{G}_{v_1} \cup \mathcal{G}_{v_2})} |\beta_j|^{q_2} \right)^{\frac{1}{q_2}}
\]

This equation does not hold by picking one specific \( j \in \mathcal{G}_{v_1} \), and another specific \( k \in \mathcal{G}_{v_2} \), and setting

\[
\beta_j = \beta_k = 1, \beta_{\{p\} \setminus \{j,k\}} = 0 \quad \text{if } w_{v_1} + w_{v_2} \neq w_{v_1}^{\frac{1}{q_1}} \cdot 2^{\frac{1}{q_2}}
\]

Therefore, \( |V| > 1 \) cannot happen.

Under case 2, suppose that \( \beta \) only has one 1 at an element in \( \mathcal{G}_{v_1} \) and has 0 elsewhere, we have

\[
\sum_{g \in [m]} w_g \| \beta_{\mathcal{G}_g} \|_2 = \left( \sum_{g \in F(v_1)} w_g \right) \leq \sum_{g \in [m]} w_g \| \beta_{\mathcal{G}_g} \|_2 = w_{v_1},
\]

that is \( w_{v_1}^{\frac{1}{q_1}} = w_{v_1} \), for all \( q_2 \).

Suppose that \( \beta \) only has one 1 at an element in \( \mathcal{G}_{v_2} \) and has 0 elsewhere, we have

\[
\sum_{g \in [m]} w_g \| \beta_{\mathcal{G}_g} \|_2 = \left( \sum_{g \in F(v_2)} w_g \right) \leq \sum_{g \in [m]} w_g \| \beta_{\mathcal{G}_g} \|_2 = w_{v_2},
\]

that is \( w_{v_2}^{\frac{1}{q_2}} = w_{v_2} \), for all \( q_2 \).

Therefore, such \( w \) does not exist if \( w_{v_1} \neq w_{v_2} \). Suppose that \( w_{v_1} = w_{v_2} = w_{v_1}^{\frac{1}{q_1}} \), then for any \( \beta \) that has non-zero values only in \( \mathcal{G}_{v_1} \), we have \( w_{v_1}^{\frac{1}{q_1}} \| \beta_{\mathcal{G}_{v_1}} \|_2 = (w_{v_1} \| \beta_{\mathcal{G}_{v_1}} \|_{q_2})^{\frac{1}{q_1}} \). Therefore, if there is norm satisfies (10), it must be a \( \ell_1/\ell_2 \) norm.

Because \( \mathcal{G}_{v_1} \) and \( \mathcal{G}_{v_2} \) are two different groups after-partition, if one original group contains variables in \( \mathcal{G}_{v_1} \), then this group must contain all variables in \( \mathcal{G}_{v_1} \). Moreover, for one original group contains all variables in \( \mathcal{G}_{v_1} \), such group
either containing none variable in \( G_{v_1} \) or containing all variables in \( G_{v_2} \). Thus, we can define a partition A,B,C,D of \([m]\), where

\[
\begin{align*}
g \in A & \quad \text{if } G_{v_1} \cup G_{v_2} \subseteq G_g \\
g \in B & \quad \text{if } G_{v_1} \subseteq G_g, G_{v_2} \cap G_g = \phi \\
g \in C & \quad \text{if } G_{v_2} \subseteq G_g, G_{v_1} \cap G_g = \phi \\
g \in D & \quad \text{if } G_{v_1} \cap G_g = \phi, G_{v_2} \cap G_g = \phi
\end{align*}
\]

Since \( G_{v_1} \) and \( G_{v_2} \) are two different groups after-partition, we know that \( B \cap C \neq \phi \). Since \( w_{v_1} = \sum_{g \in A \cup B} w_g = w_{v_2} = \sum_{g \in A \cup C} w_g \), we know that both \( B \neq \phi, C \neq \phi \) and \( \sum_{g \in B} w_g = \sum_{g \in C} w_g \). Suppose that \( \beta \) has non-zero values only in \( G_{v_1} \) and \( G_{v_2} \), we have

\[
\sum_{g \in [m]} w_g \|\beta_{G_g}\|_2 = \sum_{g \in A} w_g \|\beta_{G_{v_1}} \cup \beta_{G_{v_2}}\|_2 + \sum_{g \in B} w_g \|\beta_{G_{v_1}}\|_2 + \sum_{g \in C} w_g \|\beta_{G_{v_2}}\|_2 \\
= \sum_{g \in A} w_g \|\beta_{G_{v_1}} \cup \beta_{G_{v_2}}\|_2 + \left(\|\beta_{G_{v_1}}\|_2 + \|\beta_{G_{v_2}}\|_2\right) \sum_{g \in B} w_g \\
> w_3 \|\beta_{G_{v_1}} \cup \beta_{G_{v_2}}\|_2 = \|\beta_{\{G_{v_1}, G_{v_2}\}}\|_{1,2},
\]

which is a contradiction. Therefore, in both cases, we cannot have \(|V| > 1\) which means that there exists a \( g \in [m] \) such that \( \hat{G}_g = G_g \).

C.2 Proof of Theorem 2

Proof

We first focus on the bound of \( \hat{\beta}^G \). Given a group structure \( G \) and a fixed design matrix \( X \) satisfies Assumption 4, by taking an appropriate \( \lambda_n \), Theorem 4 implies that (15) and (17) hold with probability at least \( 1 - e^{-2n \delta} \).

Under Assumption 1, Theorem 5 implies that Assumption 4 holds with probability at least \( 1 - e^{-c_2 n \delta^2} - \frac{e^{- \frac{n \delta}{\log 5}}}{1 - e^{- \frac{n \delta}{\log 5}}} \) for some positive constant \( c_2 \).

Taking account of both theorems together, we have Assumption 1, (15) and (17) hold with probability at least \( 1 - e^{-c_2 n \delta} - e^{-2n \delta} - \frac{e^{- \frac{n \delta}{\log 5}}}{1 - e^{- \frac{n \delta}{\log 5}}} \geq 1 - e^{-c' n \delta} \) for some suitable constant \( c' \).

The bound for \( \hat{\beta}^G \) is directly available by noticing that it is a group lasso estimator with group \( G \) and \( w \).

C.3 Proof of Corollary 1

Proof Recall that

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

further define

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

By assuming \( d \geq d' \) and \( m \geq m' \), we have \( \left( d_{\max} \log \frac{5}{n} + \log m + \delta\right) \geq \left( d_{\max} \log \frac{5}{n} + \log m + \delta\right) \).

With the setting \( w_g = \sum_{g \in F(g)} w_g \), by Cauchy–Schwartz inequality, we have

\[
\omega_g^2 = \left( \sum_{g \in F(g)} w_g \right)^2 \leq \sum_{g \in F(g)} w_g^2.
\]
Consequently,

$$\sum_{g \in F^{-1}(S)} w_g^2 \leq \sum_{g \in F^{-1}(S)} h_g \left( \sum_{g \in F(g)} w_g^2 \right) \leq h_{\text{max}}(G_S) \left( \sum_{g \in F^{-1}(S)} \sum_{g \in F(g)} w_g^2 \right).$$

Now we want to show that

$$\sum_{g \in F^{-1}(S)} \sum_{g \in F(g)} w_g^2 \leq \sum_{g \in F^{-1}(S)} \sum_{g \in F(g)} w_g^2 \leq h_{\text{max}}(G_S) K \sum_{g \in S} w_g^2$$

where $K = \max_{g \in S} k_g$. On the other hand, we have

$$\min_{g \in [m]} \left( w_g^2 \right) \leq \min_{g \in [m]} \left( \sum_{g \in F(g)} w_g \right)^2 \geq \min_{g \in [m]} \left( \sum_{g \in F(g)} \min_{g \in [m]} \{ w_g \} \right)^2 \geq \min_{g \in [m]} \left( \min_{g \in [m]} \{ w_g \} \right)^2 \geq \min_{g \in [m]} \left( w_g^2 h_{\text{min}}^g \right).$$

Therefore,

$$\frac{\sum_{g \in F^{-1}(S)} w_g^2}{\min_{g \in [m]} \left( w_g^2 \right)} \leq \frac{K \left( \sum_{g \in S} w_g^2 \right)}{\min_{g \in [m]} \left( w_g^2 h_{\text{min}}^g \right)} \cdot h_{\text{max}}(G_S).$$

Consequently,

$$\frac{\sum_{g \in F^{-1}(S)} w_g^2}{\min_{g \in [m]} \left( w_g^2 \right)} \cdot \left( d_{\text{max}} \log \frac{5}{n} + \log m \frac{n}{n} + \delta \right) \leq \frac{\left( \sum_{g \in S} w_g^2 \right)}{\min_{g \in [m]} \left( w_g^2 h_{\text{min}}^g \right)} \cdot h_{\text{max}}(G_S) \left( d_{\text{max}} \log \frac{5}{n} + \log m \frac{n}{n} + \delta \right).$$
C.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

\[
\phi^* (v) = \sup_{\phi(u) \leq 1} \left\{ u^T v \right\} = \sup_{\phi(u) \leq 1} \left\{ u_1 v_1 + u_2 v_2 + \cdots + u_p v_p \right\}
\]

\[
= \sup_{\phi(u) \leq 1} \left\{ \frac{v_1}{h_1} \cdot u_1 + \cdots + \frac{v_p}{h_p} \cdot u_p \right\}
\]

\[
= \sup_{\phi(u) \leq 1} \left\{ \sum_{g=1}^{m} (H_{G_g} v_{G_g})^T u_{G_g} \right\} = \sup_{\phi(u) \leq 1} \left\{ \sum_{g=1}^{m} \left( \frac{(H v)_{G_g}}{w_g} \right) : w_g \cdot u_{G_g} \right\}
\]

\[
\leq \sup_{\phi(u) \leq 1} \left\{ \sum_{g=1}^{m} \left\| (H v)_{G_g} \right\|_2 \cdot \left\| w_g u_{G_g} \right\|_2 \right\} \leq \left( \max_{g \in [m]} \frac{1}{w_g} : \left\| (H v)_{G_g} \right\|_2 \right) \cdot \phi(u)
\]

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

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

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

\[
u_j = \begin{cases} \frac{1}{w_{g_0}} \cdot \frac{1}{h_j} & \text{for } j \notin G_{g_0} \\ \frac{1}{w_{g_0}} \cdot \frac{1}{h_j} \cdot \frac{1}{\left\| (H v)_{G_{g_0}} \right\|_2} & \text{for } j \in G_{g_0}, \end{cases}
\]

then we have

\[
\phi(u) = \sum_{g=1}^{m} w_g \left\| u_{G_g} \right\|_2 = \frac{1}{w_{g_0}} \cdot \frac{1}{\left\| (H v)_{G_{g_0}} \right\|_2} \cdot \left( \sum_{j \in G_{g_0}} \frac{v_j^2}{h_j^2} \right)
\]

\[
= \frac{1}{\left\| (H v)_{G_{g_0}} \right\|_2} \cdot \left( \sum_{j \in G_{g_0}} \frac{v_j^2}{h_j^2} \right) = 1,
\]

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

\[
u^T v = \frac{1}{w_{g_0}} \cdot \frac{1}{\left\| (H v)_{G_{g_0}} \right\|_2} \cdot \sum_{j \in G_{g_0}} \frac{v_j^2}{h_j^2} = \frac{1}{w_{g_0}} \cdot \frac{1}{\left\| (H v)_{G_{g_0}} \right\|_2} \cdot \left\| (H v)_{G_{g_0}} \right\|_2^2
\]

\[
= \frac{1}{w_{g_0}} \cdot \left\| (H v)_{G_{g_0}} \right\|_2 = \max_{g \in [m]} \frac{1}{w_g} \cdot \left\| (H v)_{G_g} \right\|_2 = \phi^*(v).
\]

Therefore, this is a sharp bound.
C.5 Proof of Theorem 4

Proof By default, we take \( S = S(\beta^*) \) and \( \overline{S} = \overline{S}(\beta^*) \) in all settings. From the optimality of \( \hat{\beta}^G \), we have

\[
0 \geq \frac{1}{n} \| Y - X \hat{\beta} \|_2^2 - \frac{1}{n} \| Y - X \hat{\beta}^* \|_2^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
= \frac{1}{n} \left( Y^T Y - 2 Y^T X \hat{\beta} + \hat{\beta}^T X^T X \hat{\beta} - Y^T Y + 2 Y^T X \hat{\beta}^* - \hat{\beta}^T X^T X \hat{\beta}^* \right) + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
= \frac{1}{n} \left( 2 X^T X \hat{\beta}^* - 2 X^T Y (\hat{\beta} - \beta^*) + (\hat{\beta} - \beta^*)^T X^T X (\hat{\beta} - \beta^*) \right) + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
= \left\langle \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right\rangle + \frac{\| X (\hat{\beta} - \beta^*) \|_2^2}{n} + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
\geq \left\langle \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right\rangle + \kappa \left\| (\hat{\beta} - \beta^*) \right\|_2^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
\geq - \left\langle \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right\rangle + \kappa \left\| (\hat{\beta} - \beta^*) \right\|_2^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right),
\]

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

By applying Hölder’s inequality with the regularizer \( \phi \) and its dual norm \( \phi^* \), we have

\[
\left\| \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right\| \leq \phi^* \left( \nabla \frac{\| Y - X \beta^* \|_2^2}{n} \right) \phi \left( \hat{\beta} - \beta^* \right). \tag{25}
\]

Next, we have

\[
\phi(\hat{\beta}) = \phi \left( \beta^* + (\hat{\beta} - \beta^*) \right) = \phi \left( \beta^*_{M(S)} + \beta^*_{M(\overline{S})} + (\hat{\beta} - \beta^*)_{M(\overline{S})} + (\hat{\beta} - \beta^*)_{M(S)} \right) \\
\geq \phi \left( \beta^*_{M(S)} + (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi(\beta_{M(S)}) - \phi((\hat{\beta} - \beta^*)_{M(\overline{S}))} - \phi((\hat{\beta} - \beta^*)_{M(S)}) \\
= \phi(\beta^*_{M(S)}) + \phi \left( (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi(\beta^*_{M(\overline{S}))} - \phi((\hat{\beta} - \beta^*)_{M(S)}) \\
\]

The inequality hold by applying the triangle inequality on \( \phi(\hat{\beta}) \), and the last step hold by applying Lemma 9. Consequently, we have

\[
\phi(\hat{\beta}) - \phi(\beta^*) \geq \phi \left( (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) - 2 \phi(\beta^*_{M(S)}) \\
= \phi \left( (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right), \tag{26}
\]

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

Based on equation (25) and equation (26), we have

\[
\frac{1}{n} \| Y - X \hat{\beta} \|_2^2 - \frac{1}{n} \| Y - X \hat{\beta}^* \|_2^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
\geq - \left\langle \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right\rangle + \kappa \left\| (\hat{\beta} - \beta^*) \right\|_2^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \\
\geq \kappa \left\| (\hat{\beta} - \beta^*) \right\|_2^2 + \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) \right) - \left\langle \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right\rangle \\
\geq \kappa \left\| (\hat{\beta} - \beta^*) \right\|_2^2 + \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) \right) - \phi^* \left( \nabla \frac{\| Y - X \beta^* \|_2^2}{n}, (\hat{\beta} - \beta^*) \right) \\
\geq \kappa \left\| (\hat{\beta} - \beta^*) \right\|_2^2 + \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_{M(\overline{S})} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) \right) - \frac{\lambda_n}{2} \phi \left( (\hat{\beta} - \beta^*) \right),
\]
where the last step is valid because Lemma 8 implies that we can guarantee \( \lambda_n \geq 2\phi^* \left( \sqrt{\frac{\|Y-X\beta^*\|^2}{n}} \right) \) with high probability by taking appropriate \( \lambda_n \). Moreover, Lemma 10 implies that

\[
\hat{\beta} \in \left\{ \beta \in \mathbb{R}^p \mid \phi \left( (\beta - \beta^*_M)^T S \right) \leq 3\phi \left( (\beta - \beta^*_M)^T S \right) \right\}.
\]

By the triangle inequality, we have

\[
\phi(\hat{\beta} - \beta^*) = \phi \left( (\hat{\beta} - \beta^*)_M + (\hat{\beta} - \beta^*)_M \right) \leq \phi \left( (\hat{\beta} - \beta^*)_M \right) + \phi \left( (\hat{\beta} - \beta^*)_M \right),
\]

and hence we have

\[
\frac{1}{n} \|Y - X\hat{\beta}\|^2 - \frac{1}{n} \|Y - X\beta^*\|^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right)
\]

\[
\geq \kappa \left( \|\hat{\beta} - \beta^*\|^2 \right) + \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_M \right) - \phi \left( (\beta - \beta^*)_M \right) \right) - \frac{\lambda_n}{2} \phi \left( (\beta - \beta^*)_M \right)
\]

\[
\geq \kappa \left( \|\hat{\beta} - \beta^*\|^2 \right) + \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_M \right) - 3\phi(\hat{\beta} - \beta^*)_M - \phi \left( (\beta - \beta^*)_M \right) \right)
\]

\[
\geq \kappa \left( \|\hat{\beta} - \beta^*\|^2 \right) - \frac{3\lambda_n}{2} \phi \left( (\beta - \beta^*)_M \right).
\]

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

\[
\sum_{g \in S} w_g \left\| (\hat{\beta} - \beta^*)_{G_g} \right\|_2 \leq \sqrt{\sum_{g \in S} w_g^2 \cdot \sqrt{h_{\max}(G_S) \cdot \max_{g \in S} \left\| (\hat{\beta} - \beta^*)_{G_g} \right\|_2^2}}
\]

\[
\leq \sqrt{\sum_{g \in S} w_g^2 \cdot \sqrt{h_{\max}(G_S) \cdot \left\| (\hat{\beta} - \beta^*) \right\|_2^2}}
\]

\[
= \sqrt{\sum_{g \in S} w_g^2 \cdot \sqrt{h_{\max}(G_S) \cdot \left\| (\hat{\beta} - \beta^*) \right\|_2}}.
\]

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

\[
\left\| \hat{\beta} - \beta^* \right\|_2^2 \leq \frac{9\lambda_n^2}{4\kappa^2} \sum_{g \in S} w_g^2 \cdot h_{\max}(G_S)
\]

\[
\leq \frac{9}{4\kappa^2} \cdot \frac{64c^2\sigma^2}{\min_{g \in [m]} \left( w_g^2 h_g^2 \right)} \cdot \left( d_{\max} \log 5 \frac{\log m}{n} + \frac{\log m}{n} + \delta \right)
\]

\[
\leq \frac{144c^2\sigma^2}{\kappa^2} \cdot \frac{\sum_{g \in S} w_g^2 \cdot h_{\max}(G_S)}{\min_{g \in [m]} \left( w_g^2 h_g \right)} \cdot \left( d_{\max} \log 5 \frac{\log m}{n} + \frac{\log m}{n} + \delta \right).
\]
C.6 Lemmas for the proof of Theorem 4

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

**Lemma 8.** Under the Assumption 4 and (2), taking
\[
\lambda_n = \frac{\log 5}{\log n} E_{(\mu_u, \mu_{\epsilon u})} \left( \frac{d_{\max} \log 5}{n} \right) + \frac{\log m}{n} + \delta \quad \text{for some } \delta \in [0, 1],
\]
then $P \left( \lambda_n \geq 2 \phi^* \left( \frac{X^T \epsilon}{n} \right) \right) \geq 1 - e^{-2n\delta}$.

**Proof [Proof of Lemma 8]**

Let $V_{i,g} = -\epsilon_i t^2 2n \left( \sum_{i=1}^{d_g} u_{i,g} X_{i,g} \right)^2 \left( \sum_{i=1}^{d_g} \sum_{j=1}^{n} u_{j,g} X_{i,g} \right)^2 \left( \sum_{i=1}^{d_g} \sum_{j=1}^{n} u_{j,g} X_{i,g} \right)^2$.

By Assumption 4, we have $\gamma_{\max} \left( \frac{X^T G^T X G_u}{n} \right) \leq \frac{c^2}{2}$. Combining this with the previous proof, we have
\[
\frac{1}{n} \log E \left( e^{t n \sum_{i=1}^{n} V_{i,g}} \right) \leq \frac{c^2 \sigma^2}{2w_2^2 (h_{\min})^2}. \]

Therefore, the random variable $\left( u, \sum_{i=1}^{n} V_{i,g} \right)$ is the sub-Gaussian with the parameter at most $\frac{c^2 \sigma^2}{2w_2^2 (h_{\min})^2}$, and by properties of sub-Gaussian variables, we have
\[
\log P \left( \left( u, \sum_{i=1}^{n} V_{i,g} \right) \geq \lambda_n \right) \leq -\frac{\lambda_n^2}{8C^2 \sigma^2}.
\]
We can find a $\frac{1}{n}$ covering of $S^{d_g-1}$ in Euclidean norm: $\{u^1, u^2, \ldots, u^N\}$ with $N \leq 5d_g$, recall that $\frac{1}{n} \| \sum_{i=1}^{n} V_{i,g} \|_2 = \frac{1}{n} \sup_{u \in S^{d_g-1}} \left\langle u, \frac{1}{n} \sum_{i=1}^{n} V_{i,g} \right\rangle$, 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 $\|u^{q(u)} - u\|_2 \leq \frac{1}{2}$, and

$$\frac{1}{n} \sup_{u \in S^{d_g-1}} \left\langle u, \frac{1}{n} \sum_{i=1}^{n} V_{i,g} \right\rangle = \frac{1}{n} \sup_{u \in S^{d_g-1}} \left( \left\langle u - u^{q(u)}, \frac{1}{n} \sum_{i=1}^{n} V_{i,g} \right\rangle + \left\langle u^{q(u)}, \frac{1}{n} \sum_{i=1}^{n} V_{i,g} \right\rangle \right) \leq \frac{1}{n} \sup_{u \in S^{d_g-1}} \left\langle u - u^{q(u)}, \sum_{i=1}^{n} V_{i,g} \right\rangle + \frac{1}{n} \max_{q \in [N]} \left(u^{q}, V_{i,g}\right).$$

By Cauchy inequality, $\frac{1}{n} \sup_{u \in S^{d_g-1}} \left\langle u - u^{q(u)}, \frac{1}{n} \sum_{i=1}^{n} V_{i,g} \right\rangle \leq \frac{1}{n} \sup_{u \in S^{d_g-1}} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2 \leq \frac{1}{2n} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2$, hence we have $\frac{1}{n} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2 \leq \frac{1}{2n} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2 + \frac{1}{n} \max_{q \in [N]} \left(u^{q}, \sum_{i=1}^{n} V_{i,g}\right)$, which indicates that $\frac{1}{n} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2 \leq 2 \max_{q \in [N]} \left(u^{q}, \frac{1}{n} \sum_{i=1}^{n} V_{i,g}\right)$. Consequently,

$$P \left( \frac{1}{n} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2 \geq \frac{\lambda_n}{2} \right) \leq P \left( \max_{q \in [N]} \left(u^{q}, \frac{1}{n} \sum_{i=1}^{n} V_{i,g}\right) \geq \frac{\lambda_n}{4} \right) \leq P \left( \bigcup_{q=1}^{N} \left(u^{q}, \frac{1}{n} \sum_{i=1}^{n} V_{i,g}\right) \geq \frac{\lambda_n}{4} \right) \leq \frac{N}{\lambda_n} \leq \exp \left(-\frac{n\lambda_n^2 w_{g,2h_{\text{min}}}}{32\sigma^2} \right),$$

and by setting $\lambda_n = \frac{2\sigma\log(5n) - \log m}{d_{\text{max}} \log 5 + \log m}$, we have

$$P \left( \frac{\lambda_n}{2} \right) \leq \exp \left(-\frac{n\lambda_n^2 w_{g,2h_{\text{min}}}}{32\sigma^2} \right) \leq \exp \left(-2n\delta\right),$$

which equivalent to $P \left( \lambda_n \geq 2 \max_{g \in [m]} \frac{1}{n} \left\| \sum_{i=1}^{n} V_{i,g} \right\|_2 \right) = 1 - \exp \left(-2n\delta\right)$. From Proposition 1, we have

$$\phi^{*} \left( \frac{X^T \varepsilon}{n} \right) \leq \max_{g \in [m]} \frac{1}{w_{g}} \left\| \frac{HX^T \varepsilon}{n} \right\|_{G_g} = \max_{g \in [m]} \frac{1}{w_{g}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{i} \left( \frac{X_{ig_1}}{h_{g_1}}, \ldots, \frac{X_{ig_{\lambda_n}}}{h_{g_{\lambda_n}}} \right) \right\|_{2} \leq \max_{g \in [m]} \frac{1}{w_{g}} \left\| \frac{1}{n} \sum_{i=1}^{n} V_{i,g} \right\|_{2},$$

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

\begin{lemma}
The group lasso regularizer (1) is decomposable with respect to the pair $\{M (S), M^{\perp} (S)\}$. That is, $\phi(a + b) = \phi(a) + \phi(b)$, for all $a \in M (S)$ and for all $b \in M^{\perp} (S)$.
\end{lemma}
Therefore, from which the claim follows.

**Proof** [Proof of Lemma 9]

\[
\phi(a + b) = \sum_{g=1}^{m} w_g \|(a + b)_{G_g}\|_2 = \sum_{g \in M(S)} w_g \|(a + b)_{G_g}\|_2 + \sum_{g \notin M(S)} w_g \|(a + b)_{G_g}\|_2 \\
= \sum_{g \in M(S)} w_g \|a_{G_g}\|_2 + \sum_{g \notin M(S)} w_g \|b_{G_g}\|_2 = \sum_{g \in M(S)} w_g \|a_{G_g}\|_2 + \sum_{g \notin M(S)} w_g \|b_{G_g}\|_2 \\
= \phi(a) + \phi(b)
\]

\[
\square
\]

**Lemma 10.** If \( \lambda_n \geq 2\phi^* \left( \frac{X^T x}{n} \right) \), then \( \phi \left( (\hat{\beta} - \hat{\beta}^*)_{M^c(S)} \right) \leq 3\phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) \).

**Proof** [Proof of Lemma 10 (also see proposition 9.13 in (Wainwright, 2019))] From equation (26), we have

\[
\phi(\hat{\beta}) - \phi(\hat{\beta}^*) \geq \phi \left( (\hat{\beta} - \hat{\beta}^*)_{M^c(S)} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right),
\]

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

\[
\frac{1}{n} \|Y - X\hat{\beta}\|_2^2 - \frac{1}{n} \|Y - X\beta^*\|_2^2 \geq \left\langle \nabla \frac{1}{n} \|Y - X\beta^*\|_2^2, (\hat{\beta} - \beta^*) \right\rangle \geq -\left\langle \nabla \frac{1}{n} \|Y - X\beta^*\|_2^2, (\hat{\beta} - \beta^*) \right\rangle.
\]

By applying Holder’s inequality with the regularizer \( \phi \) and its dual norm \( \phi^* \), we have

\[
\left\langle \nabla \frac{1}{n} \|Y - X\beta^*\|_2^2, (\hat{\beta} - \beta^*) \right\rangle \leq \phi^* \left( \nabla \frac{1}{n} \|Y - X\beta^*\|_2^2 \right) \phi \left( \hat{\beta} - \beta^* \right).
\]

Therefore,

\[
\frac{1}{n} \|Y - X\hat{\beta}\|_2^2 - \frac{1}{n} \|Y - X\beta^*\|_2^2 \geq -\phi^* \left( \nabla \frac{1}{n} \|Y - X\beta^*\|_2^2 \right) \phi \left( \hat{\beta} - \beta^* \right) \geq -\frac{\lambda_n}{2} \phi \left( \hat{\beta} - \beta^* \right) \geq -\frac{\lambda_n}{2} \left( \phi(\hat{\beta} - \beta^*)_{M(S)} + \phi(\hat{\beta} - \beta^*)_{M^c(S)} \right),
\]

and

\[
0 = \frac{1}{n} \|Y - X\hat{\beta}\|_2^2 - \frac{1}{n} \|Y - X\beta^*\|_2^2 + \lambda_n \left( \phi(\hat{\beta}) - \phi(\beta^*) \right) \geq \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_{M^c(S)} \right) - \phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) - 2\phi(\beta^*)_{M^c(S)} \right) - \frac{\lambda_n}{2} \left( \phi(\hat{\beta} - \beta^*)_{M(S)} + \phi(\hat{\beta} - \beta^*)_{M^c(S)} \right) \]

\[
= \lambda_n \left( \phi \left( (\hat{\beta} - \beta^*)_{M^c(S)} \right) - 3\phi \left( (\hat{\beta} - \beta^*)_{M(S)} \right) \right),
\]

from which the claim follows.

\[
\square
\]

**C.7 Proof of Theorem 5**

**Lemma 11.** (Theorem 6.5 in (Wainwright, 2019))

Let \( |||\cdot|||_2 \) 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

\[
E \left( e^{t|||\hat{\Theta} - \Theta|||_2} \right) \leq e^{c_3 \frac{t^2 p^2}{n} + 4p} \quad \text{for all} \quad |t| < \frac{n}{64e^2 |||\Theta|||_2}.
\]

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

\[
P \left( \frac{|||\hat{\Theta} - \Theta|||_2}{|||\Theta|||_2} \leq c_5 \left( \sqrt{\frac{p}{n} + \frac{p}{n}} + \delta \right) \right) > 1 - c_4 e^{-c_2 n \delta^2}
\]

(28)
Lemma 12. Under Assumption 1, and use $\rho(\Theta)$ to denote the maximum diagonal of a covariance matrix $\Theta$. For any vector $\beta \in \mathbb{R}^p$ and a given group structure with $m$ groups, we have

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

with probability at least $1 - e^{-\frac{c \beta^2 n}{2}}$.

Proof [Proof of Theorem 5 Part 1]

By Lemma 11, we have

$$P\left( \frac{\|XG^{T}XG\|_2}{\|\Theta_{G,G} \|_2} \leq c_5 (\sqrt{d_g/n} + \delta) \right) > 1 - c_4 e^{-c_2 n \delta^2}$$

By triangle inequality, since $XG^{T}XG$ is a positive semi-definite, we have

$$\gamma_{\max}(\frac{XG^{T}XG}{n}) = \|\frac{XG^{T}XG}{n}\|_2 = \|\frac{XG^{T}XG}{n} - \Theta_{G,G} \|_2 + \|\Theta_{G,G} \|_2$$

$$\leq (1 + c_5 (\sqrt{d_g/n} + \delta)) \|\Theta_{G,G} \|_2,$$

with probability at least $1 - c_4 e^{-c_2 n \delta^2}$. Because $\|\Theta_{G,G} \|_2 \leq \|\Theta \|_2 \leq c_1$ for some constant $c_1$ and $d_g \leq n$, we have $\gamma_{\max}(\frac{XG^{T}XG}{n}) \leq c + \delta$ for some constant $c$, with probability at least $1 - e^{-c_2 n \delta^2}$. Taking the union probability for all $m$ groups, we have

$$\max_{g \in [m]} \gamma_{\max}(\frac{XG^{T}XG}{n}) \leq c + \delta$$

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 $\delta$ as a constant as the result.

Proof [Proof of Theorem 5 Part 2] First note that we must have $\rho(\Theta) \leq \gamma_{\max}(\Theta) \leq c_1$ by Assumption 1. By applying Minkowski inequality, we have

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

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

$$\phi(\beta) = \phi(\beta_{M(\tilde{S})}) + \phi(\beta_{M-(\tilde{S})}) \leq 4 \phi(\beta_{M(\tilde{S})}) \leq 4 \sqrt{8} \sqrt{\max_{g \in \tilde{S}} w_g^2 h^g_{\max} \|\beta_{M(\tilde{S})} \|_2^2}$$

From Lemma 12, we have

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

$$\geq \frac{1}{4 \sqrt{c_1}} \|\beta\|_2 - 32 \rho(\Theta) \max_{g \in [m]} \frac{1}{w_g \sqrt{h^g_{\min}}} \sqrt{2(\log m + d_{\max} \log 5)} \sqrt{\max_{g \in S} w_g^2 h^g_{\max} \|\beta \|_2^2} \frac{\phi(\beta)}{n}$$

$$\geq \frac{1}{64 \sqrt{c_1}} \|\beta\|_2,$$
where the last step is valid due to Assumption 1.

C.8 Lemmas for the proof of Theorem 5

Proof [Proof of Lemma 12] To begin with, for a vector $\beta \in \mathbb{R}^p$ with a fixed group structure, we define the set $S^{p-1}(\Theta) = \left\{ \beta \in \mathbb{R}^p \left| \left\| \Theta \frac{1}{2} \beta \right\|_2 = 1 \right. \right\}$, the function

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

and the event

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

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

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

and the event:

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

Based on lemma 12.1 and lemma 12.2, we have

$$P(X \in \mathcal{E}) \leq P(\mathcal{A}(0, v)) + \sum_{\ell=1}^{\infty} P(\mathcal{A}(2^{\ell-1}v, 2^\ell v)) \leq e^{-\frac{v}{2}} \left\{ \sum_{\ell=0}^{\infty} e^{-\frac{v}{2} 2^{2\ell}} \right\}. $$

Since $v = \frac{1}{4}$ and $2^\ell \geq 2\ell$, we have $P(X \in \mathcal{E}) \leq e^{-\frac{v}{2}} \sum_{\ell=0}^{\infty} e^{-\frac{v}{2} 2^{2\ell}} \leq e^{-\frac{v}{2}} \sum_{\ell=0}^{\infty} e^{-\frac{v}{2} \ell^2} \leq e^{-\frac{v}{2}} \frac{1}{1 - e^{-\frac{v}{2}}}.$

We just get upper bound of $P(X \in \mathcal{E})$. We next show that the bound in (29) always hold on the complementary set $\mathcal{E}^c$.

If $X \notin \mathcal{E}$, based on the definition of $\mathcal{E}$, we have $\inf_{\beta \in S^{p-1}(\Theta)} \frac{\left\| X \beta \right\|_2}{\sqrt{n}} \geq \frac{1}{4} - 2g(\phi(\beta))$. That is $\forall \beta \in S^{p-1}(\Theta)$.

Therefore, for any $\beta' \in \left\{ \beta' \in \mathbb{R}^p \left| \frac{\beta'}{\left\| \Theta \frac{1}{2} \beta' \right\|_2} \in S^{p-1}(\Theta) \right. \right\}$, we have

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

By substituting the definition of $g(\phi(\beta))$, we finish the proof.

Lemma 12.1 For $v = \frac{1}{4}$, we have $\mathcal{E} \subseteq \mathcal{A}(0, v) \cup \left( \bigcup_{\ell=1}^{\infty} \mathcal{A}(2^{\ell-1}v, 2^\ell v) \right)$.

Lemma 12.2 For any pair $(r_\ell, r_u)$, where $0 \leq r_\ell \leq r_u$, we have $P(\mathcal{A}(r_\ell, r_u)) \leq e^{-\frac{v}{2}} e^{-\frac{v}{2} r_u^2}$.

Proof [Proof of Lemma 12.1] By definition, $\mathcal{K}(0, v) \cup \left( \bigcup_{\ell=1}^{\infty} \mathcal{K}(2^{\ell-1}v, 2^\ell v) \right)$ is a cover of $S^{p-1}(\Theta)$. Therefore, for any $\beta$, it either belongs to $\mathcal{K}(0, v)$ or $\mathcal{K}(2^{\ell-1}v, 2^\ell v)$.

Case 1 If $\beta \in \mathcal{K}(0, v)$, by definition, we have $g(\phi(\beta)) \in [0, v]$ and

$$\frac{\left\| X \beta \right\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\phi(\beta)) \leq \frac{1}{4} = \frac{1}{2} - v.$$
Therefore, the event $\mathcal{A}(0, v)$ must happen in this case.

**Case 2:** If $\beta \notin \mathbb{K}(0, v)$, we must have $\beta \in \mathbb{K}(2^{f-1}v, 2^{f}v)$ for some $f = 1, 2, \ldots$, and moreover

$$\frac{\|X\beta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\phi(\beta)) \leq \frac{1}{4} - 2\cdot (2^{f-1}v) \leq \frac{1}{2} - (2\cdot 2^{f-1}v) \leq \frac{1}{2} - 2^{f}v.$$ 

So that the event $\mathcal{A}(2^{f-1}v, 2^{f}v)$ must happen. Therefore, $\mathcal{B} \subseteq \mathcal{A}(0, v) \cup \bigcup_{f=1}^{\infty} \mathcal{A}(2^{f-1}v, 2^{f}v)$.

**Proof** [Proof of Lemma 12.2] To prove Lemma 12.2, we define and bound the random variable $T(r_\ell, r_u) = - \inf_{\beta \in \mathbb{K}(r_\ell, r_u)} \frac{\|X\beta\|_2}{\sqrt{n}}$. Let $S^{n-1}$ be a unit ball on $\mathbb{R}^n$, by the variational representation of the $\ell_2$-norm, we have

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

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

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

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

Define $Z_{u,v} = (u, Wv)$, 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:

$$Y_{u,v} = \frac{\langle \zeta, u \rangle}{\sqrt{n}} + \frac{\langle \xi, v \rangle}{\sqrt{n}}$$

where $\zeta \in \mathbb{R}^n, \xi \in \mathbb{R}^p$, have i.i.d $N(0,1)$ entries.

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

$$E(Z_{u,v} - Z_{u',v'})^2 = E \left( \frac{(u', Wv') - (u, Wv)}{\sqrt{n}} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (u_i v_j - u'_i v'_j)^2 \quad \text{(30)}$$

On one hand, since $\|v\|^2_2 \leq 1$, $\|u\|^2_2 \leq 1$, (7) \leq \frac{1}{n} \left( \|u - u'\|^2_2 + \|v - v'\|^2_2 \right) .

One the other hand, we have

$$E(Y_{u,v} - Y_{u',v'})^2 = E \left( \frac{\langle \zeta, u - u' \rangle}{\sqrt{n}} + \frac{\langle \xi, v - v' \rangle}{\sqrt{n}} \right)^2 = \frac{1}{n} \left( \|u - u'\|^2_2 + \|v - v'\|^2_2 \right) \quad \text{(31)}$$

Taking equation (30) and (31) together, we have

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

If $V = V'$, then $nE \left( Z_{u,v} - Z_{u',v'} \right)^2 = \|u - u'\|^2_2 = nE \left( Y_{u,v} - Y_{u',v'} \right)^2 .$
By applying Gordon’s inequality, we have

\[
E \left( \sup_{v \in \mathbb{R}^2} \inf_{u \in S^{n-1}} Z_{u,v} \right) \leq E \left( \sup_{v \in \mathbb{R}^2} \inf_{u \in S^{n-1}} Y_{u,v} \right).
\]

Therefore,

\[
E(T(r_t, r_u)) = E \left( \sup_{v \in \mathbb{R}^2} \inf_{u \in S^{n-1}} \frac{\langle u, Wv \rangle}{\sqrt{n}} \right) \leq E \left( \sup_{v \in \mathbb{R}^2} \inf_{u \in S^{n-1}} \left( \frac{\langle \xi, v \rangle}{\sqrt{n}} + \frac{\langle \xi, u \rangle}{\sqrt{n}} \right) \right)
\]

\[
= E \left( \sup_{\beta \in \mathbb{K}(r_t, r_u)} \left( \frac{\Sigma^{1/2} \xi, \beta}{\sqrt{n}} \right) \right) - E \left( \frac{\|\xi\|_2^2}{\sqrt{n}} \right).
\]

Next, we bound these two terms. For the second term, we have

\[
E \left( \frac{\|\xi\|_2^2}{\sqrt{n}} \right) = E \left( \frac{\xi_1^2 + \cdots + \xi_n^2}{n} \right) = E \left( \frac{\|\xi\|_1^2}{n} \right) = \sqrt{\frac{2}{\pi}}.
\]

For the first term, we have

\[
E \left( \sup_{\beta \in \mathbb{K}(r_t, r_u)} \frac{\langle \theta^T \xi, \beta \rangle}{\sqrt{n}} \right) \leq E \left( \sup_{\beta \in \mathbb{K}(r_t, r_u)} \frac{\phi(\beta) \phi^*(\theta^T \xi)}{\sqrt{n}} \right),
\]

where \(\phi^*(\theta^T \xi)\) is the dual norm defined before. Since \(\beta \in \mathbb{K}(r_t, r_u), g(\phi(\beta)) \leq r_u\), by the definition of \(g(t)\), we have

\[
\phi(\beta) \leq \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}}
\]

Let \(\eta_{G_g} = (\Theta^T \xi) G_g\), to bound \(E \left( \max_g \left\| (\Theta^T \xi) G_g \right\|_2 \right) = E \left( \max_g \left\| \eta_{G_g} \right\|_2 \right)\). Since \(\Theta^T \xi \sim N(0, \Theta)\), by the properties of normal distribution, its corresponding marginal distribution of \(j\) - th variable \((\Theta^T \xi)_j\) also follows zero mean normal distribution with covariance matrix \(\Theta_{jj}\), which is the \(j\) - th diagonal elements of \(\Theta\). Therefore, any subset of \(\Theta^T \xi\) is a zero-mean sub-Gaussian random sequence with parameters at most \(\rho(\Theta)\). By equation (32) and Lemma 12.2.3, we have

\[
E \left( \sup_{\beta \in \mathbb{K}(r_t, r_u)} \frac{\phi(\beta) \phi^*(\theta^T \xi)}{\sqrt{n}} \right) \leq E \left( \sup_{\beta \in \mathbb{K}(r_t, r_u)} \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}} \right)
\]

\[
= \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}} E \left( \frac{\phi^*(\theta^T \xi)}{\sqrt{n}} \right)
\]

\[
\leq \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}} E \left( \max_{g \in [m]} \frac{1}{\sqrt{n}w_g} \left\| H \left( \Theta^T \xi \right) G_g \right\|_2 \right)
\]

\[
\leq \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}} E \left( \max_{g \in [m]} \frac{1}{\sqrt{n}w_g h_{\min}^2} \left\| \left( \Theta^T \xi \right) G_g \right\|_2 \right)
\]

\[
\leq \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}} E \left( \left\| \max_{g \in [m]} \left( \Theta^T \xi \right) G_g \right\|_2 \right)
\]

\[
\leq \frac{r_u}{4\rho(\Theta) \max_{g \in [m]} \frac{1}{w_g h_{\min}^2} \sqrt{\frac{2(\log m + d_{\max} \log 5)}{n}}} \left( 2\rho(\Theta) \sqrt{(\log m + d_{\max} \log 5) 2\sigma^2} \right) \leq \frac{r_u}{2}
\]

Therefore, \(E[T(r_t, r_u)] \leq -\sqrt{r_t^2 + r_u^2}\). Next we want to bound \(P(T(r_t, r_u) \geq -\frac{1}{2} + r_u)\) based on the bound of this expectation. To apply Lemma 12.2.4, we first show that, the \(f = T(r_t, r_u)\), 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

$$|T(W_1) - T(W_2)| = \left| \sup_{v \in \mathcal{R}(r_u, r_u)} \inf_{u \in \mathbb{S}^{n-1}} \frac{\langle u, W_1 v \rangle}{\sqrt{n}} - \sup_{v \in \mathcal{R}(r_u, r_u)} \inf_{u \in \mathbb{S}^{n-1}} \frac{\langle u, W_2 v \rangle}{\sqrt{n}} \right|$$

$$= \left| \sup_{v \in \mathcal{R}(r_u, r_u)} \left( -\frac{\|W_1 v\|_2}{\sqrt{n}} \right) - \sup_{v \in \mathcal{R}(r_u, r_u)} \left( -\frac{\|W_2 v\|_2}{\sqrt{n}} \right) \right|$$

$$= \left| -\inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}} - \left( -\inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}} \right) \right|$$

Suppose that $\frac{\|W_1 v\|_2}{\sqrt{n}} = \inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}}$ and $\frac{\|W_2 v\|_2}{\sqrt{n}} = \inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}}$.

**Case 1** If $\|W_1 v\|_2 > \|W_2 v\|_2$, then we have

$$|T(W_1) - T(W_2)| = \left| \inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}} - \inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}} \right|$$

$$= \left| \left( \|W_1 v\|_2 - \|W_2 v\|_2 \right) \frac{\sqrt{n}}{\sqrt{n}} \right|$$

$$\leq \left| \|W_1 v\|_2 - \|W_2 v\|_2 \right| \frac{\sqrt{n}}{\sqrt{n}}$$

Case 2 If $\|W_1 v\|_2 \leq \|W_2 v\|_2$, then we have

$$|T(W_1) - T(W_2)| = \left| \inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_2 v\|_2}{\sqrt{n}} - \inf_{v \in \mathcal{R}(r_u, r_u)} \frac{\|W_1 v\|_2}{\sqrt{n}} \right|$$

$$= \left| \left( \|W_2 v\|_2 - \|W_1 v\|_2 \right) \frac{\sqrt{n}}{\sqrt{n}} \right|$$

$$\leq \left| \|W_1 v\|_2 - \|W_2 v\|_2 \right| \frac{\sqrt{n}}{\sqrt{n}}$$

where $\|\cdot\|_F$ represent the Frobenius norm of a matrix. Thus under the Euclidean norm, $T(W)$ is a $\frac{1}{\sqrt{n}}$-Lipschitz function. Therefore, by lemma 12.2.3, we have

$$P(T(r_t, r_u) - E(T(r_t, r_u)) \geq t) \leq e^{-nt^2/2}, \forall t \geq 0$$

Set $t = \sqrt{\frac{2}{\pi} - \frac{1}{2} + \frac{\epsilon}{2}} \geq \frac{1}{2} + \frac{\epsilon}{2}$, we have, $E(T(r_t, r_u)) + t \leq -\frac{1}{2} + r_u$ and $P \left[ T(r_t, r_u) \geq -\frac{1}{2} + r_u \right] \leq e^{-nt} e^{-\frac{\epsilon}{2} r_u^2}$, which is actually the Lemma 12.2.1

**Lemma 12.2.1 (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 processes indexed by a non-empty index set $I = U \times V$. If

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

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

32
Lemma 12.2.2 Suppose that $\alpha = (\alpha_1, \ldots, \alpha_d)$, where each $\alpha_i, i \in [d]$ is a zero-mean sub-Gaussian random variable with parameter at most $\sigma^2$, then for any $t \in \mathbb{R}$, we have $E \left( \exp (t \| \alpha \|_2) \right) \leq \exp \left( 2t^2 \sigma^2 \right)$.

Lemma 12.2.3 Suppose that $\alpha = (\alpha_1, \ldots, \alpha_d)$, where each $\alpha_i, i \in [d]$ is a zero-mean sub-Gaussian random variable with parameter at most $\sigma^2$, and for a given group structure $G$, let $\| \alpha_{G_s} \|$ be the corresponding group norm, $m$ be the number of groups and $d_{max}$ be the maximum group size, then

$$E \left( \max_g \| \alpha_{G_s} \| \right) \leq 2\sqrt{2\sigma^2 \left( \log m + d_{max} \log 5 \right)}$$

Lemma 12.2.4 (Theorem 2.26 in (Wainwright, 2019)): Let $x = (x_1, \ldots, x_n)$ 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 $f(x) - Ef(x)$ is sub-Gaussian with parameter at most $L$, and hence $P \left( \left( f(x) - Ef(x) \right) \geq t \right) \leq e^{-\frac{t^2}{2L^2}}$, $\forall t \geq 0$.

C.8.1 Proof of Lemma 12.2.2

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 $N \leq 5^d$, say there exists $u^q(u) \in \{ u, \ldots, u^N \}$, such that $\| u^q(u) - u \|_2 \leq \frac{1}{2}$.

By the variational representation of the $\ell_2$ norm, we have $\| \alpha \|_2 = \max_{u \in S^{d-1}} \langle u, \alpha \rangle \leq \max_{q(u) \in [N]} \langle u^q(u), \alpha \rangle + \frac{1}{2} \| \alpha \|_2$. Therefore, $\| \alpha \|_2 \leq 2 \max_{q(u) \in [N]} \langle u^q(u), \alpha \rangle$. Consequently,

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

C.8.2 Proof of Lemma 12.2.3

For any $t > 0$, by Jensen’s inequality, we have $\exp \left( tE \left( \max_g \| \alpha_{G_s} \| \right) \right) \leq E \left( \exp \left( t \max_g \| \alpha_{G_s} \| \right) \right)$

$$= E \left( \max_j \left( \exp \left( t \| \alpha_{G_j} \| \right) \right) \right) \leq \sum_{j=1}^{m} E \left( \exp \left( t \| \alpha_{G_j} \| \right) \right) \leq \sum_{j=1}^{m} 5^{d_j} \exp \left( 2t^2 \sigma^2 \right) \leq m \cdot 5^{d_{max}} \cdot \exp (2t^2 \sigma^2).$$

By taking log at both sides, we have $tE \left( \max_g \| \alpha_{G_s} \| \right) \leq \log m + d_{max} \log 5 + 2t^2 \sigma^2$. That is $E \left( \max_g \| \alpha_{G_s} \| \right) \leq \frac{\log m + d_{max} \log 5 + 2t^2 \sigma^2}{t}$.

Let $t = \sqrt{\frac{\log m + d_{max} \log 5}{2\sigma^2}}$, we have $E \left( \max_g \| \alpha_{G_s} \| \right) \leq 2 \sqrt{\left( \log m + d_{max} \log 5 \right) 2\sigma^2}$.

C.9 Proof of Theorem 3

Lemma 13. Given $m, s_g \in \mathbb{R}$, define the set $A = \left\{ \alpha \in \{ 0, 1 \}^m \mid \sum_{j=1}^{m} a_j \leq s_g \right\}$, then the $\sqrt{s_g}$-packing number of set $A \geq \left( \frac{m}{s_g} \right)^{\frac{m}{s_g}} 2^{\frac{s_g}{s_g}}$, and log $\left( \frac{m}{s_g} \right)^{\frac{m}{s_g}} 2^{\frac{s_g}{s_g}} \approx s_g \log \left( \frac{m}{s_g} \right)$.  

33
Lemma 14. The $\sqrt{2ds_g}$-packing number of $\Omega(G, s_g) \gtrsim \left( \frac{m}{s_g} \right)^{-\frac{3}{2}} \cdot (\sqrt{2})^{ds_g}$, and

$$
\log \left( \frac{m}{s_g} \right)^{-\frac{3}{2}} \cdot (\sqrt{2})^{ds_g} \gtrsim s_g(d + \log \left( \frac{m}{s_g} \right)).
$$

Proof [Proof of Theorem 3]

First, take $N$ points $\omega^{(1)}, \ldots, \omega^{(N)}$ from $\Omega(G, s_g)$ such that $\|\omega^{(i)} - \omega^{(j)}\| > \sqrt{2ds_g}$ for all distinct $i, j$. Clearly we have $\|\omega^{(i)} - \omega^{(j)}\| \leq 4s_g d$.

Now denote $\beta^{(i)} = r\omega^{(i)}$, we have

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

Next, denote $y^{(i)} = X\beta^{(i)} + \varepsilon$ for $1 \leq i \leq N$, we consider the Kullback-Leibler divergence between different distribution pairs:

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

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

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

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

$$
D_{KL} \left( (y^{(i)}, X), (y^{(j)}, X) \right) = E_X \frac{\|X(\beta^{(i)} - \beta^{(j)}))\|^2}{2\sigma^2} = \frac{n(\beta^{(i)} - \beta^{(j)}))^\top \Sigma (\beta^{(i)} - \beta^{(j)})}{2\sigma^2}
$$

From Lemma 14, we have $\log N \gtrsim s_g \left( d + \log \left( \frac{m}{s_g} \right) \right)$, by setting $\frac{n dr^2 s_g + \log 2}{\sigma^2} = \frac{1}{2}$, we have $r \gtrsim \sqrt{\frac{d + \log \left( \frac{m}{s_g} \right)}{3nd}}$.

By generalized Fano’s Lemma, we have $\inf \sup_{\beta} E\|\hat{\beta} - \beta\| \geq \frac{\sqrt{2rs_g dr^2 s_g + \log 2}}{\sigma^2} \left( 1 - \frac{n dr^2 s_g + \log 2}{\sigma^2} \right)$, consequently

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

C.9.1 Proof of Lemma 13

Proof

Notice that the cardinality of $A$ is $\left( \frac{m}{s_g} \right)$. 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 \( a \in A \),
\[
\left\{ b \in A \mid h(a, b) \leq \frac{s_g}{2} \right\} = \left( m | \frac{s_g}{2} \right) \cdot 2|a|^2.
\]

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 \( \left| \frac{s_g}{2} \right| \), 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 \( T = \left( m | \frac{s_g}{2} \right) - 2 \), then we have
\[
\left| \{ b \in A \mid \text{there exist } a \in A_s \text{ with } h(a, b) \leq \frac{s_g}{2} \} \right| \leq (|A_s|) \cdot \left( m | \frac{s_g}{2} \right) \cdot 2|a|^2 < |A|.
\]

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 exists at least \( T \) points in \( A \) such that the distance between any two points is greater than \( \sqrt{\frac{s_g}{2}} \).

Moreover, since
\[
\left( \frac{m}{s_g} \right) = \left( \frac{m}{s_g} \right) \frac{(m-s_g+1) \cdots (m-s_g)}{(s_g-1)!} = \prod_{i=1}^{s_g} \frac{m-s_g+i}{s_g} \leq \left( \frac{m}{s_g} \right)^{s_g} \leq \left( \frac{m}{s_g} \right)^{s_g} \frac{s_g}{2|a|^2},
\]

and therefore we can find \( C_1, C_2 \), such that \( C_1 s_g \log(\frac{m}{s_g}) \leq \log T \leq C_2 s_g \log(\frac{m}{s_g}) \), so that
\[
\log \left( \frac{m}{s_g} \right) - 2 \approx s_g \log(\frac{m}{s_g}).
\]

---

**C.9.2 Proof of Lemma 14**

**Proof** Given a group support \( a \in A \), define \( k_a = |i \mid i \in \bigcup_{g(a_i=0)} G_g \bigg)^C \}, \) and the set
\[
\Omega^{(a)} = \left\{ \omega \in \mathbb{R}^p \mid \omega_i = 0 \text{ if } i \in \bigcup_{g(a_i=0)} G_g, \omega_i \in \{-1, 1\} \text{ if } i \in \bigcup_{g(a_i=0)} G_g \bigg)^C \right\}
\]

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 \( x \in \Omega^{(a)} \),
\[
|\{ y \in \Omega^{(a)} , h(x, y) \leq \frac{k_a}{10} \} | = \sum_{j=0}^{k_a} \left( \frac{k_a}{j} \right)
\]
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

$$\left| \{ y \in \Omega^{(a)} \mid \text{there exist } x \in \Omega_s^{(a)} \text{ with } h(x, y) \leq \frac{k_a}{10} \} \right| \leq \left( \Omega_s^{(a)} \right)^{\frac{2^{k_a} - 2}{\sum_{j=0}^{\lfloor \frac{k_a}{10} \rfloor} \binom{k_a}{j}}} < |\Omega^{(a)}|.$$  

On the other hand, $h(x, y) > \frac{k_a}{10}$ implies $\|x - y\| \geq \sqrt{\frac{2k_a}{5}}$. Therefore, there exists at least $N^{(a)}$ points in $\Omega^{(a)}$ such that the distance between any two points is greater than $\sqrt{\frac{2k_a}{5}}$.

As stated in Chapter 9 in (Graham et al., 1994),

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

we have $N^{(a)} = \frac{2^{k_a} - 2}{\sum_{j \leq \lfloor \frac{k_a}{10} \rfloor} \binom{k_a}{j}} > \frac{8}{9} 2^{\frac{k_a}{2}} \geq (\sqrt{2})^{k_a}$.

Notice that $k_a$ depends on the predefined groups and the group support $a$, the range is $0 \leq k_a \leq s_g d$. Since Lemma 14 aims to find a lower bound for all possible overlapping patterns, we need to consider the upper bound of $k_a$.

On the other hand, based on Lemma 13, we can find at least $T$ points in $A$ such that the distance between any two points is greater than $\sqrt{\frac{s_g}{2}}$. 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 $\sqrt{\frac{2s_g d}{5}}$, then Lemma 13 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 $\|a - b\| > \sqrt{\frac{s_g}{2}}, \|x - y\| > \sqrt{\frac{d s_g}{2}} > \sqrt{\frac{2d s_g}{5}}$ for any $x \in \Omega^{(a)}$ and $y \in \Omega^{(b)}$. Thus, considering all possible overlapping pattern, we can find at least $\frac{m}{\lfloor \frac{2s_g}{2} \rfloor} \cdot \frac{8}{9} (\sqrt{2})^{d s_g}$ point in $\Omega(G, s_g)$, such that the distance between every pair of points is greater than $\sqrt{\frac{2d s_g}{5}}$. 

\[\blacksquare\]