Electronic Journal of Statistics
Vol. 15 (2021) 1993–2028
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1831

Graphical-model based high dimensional generalized linear models*

Yaguang Li

Department of Mathematics and Statistics
York University
Dalla Lana School of Public Health
University of Toronto
Toronto, Canada
e-mail: liygcr7@gmail.com

Wei Xu†

Dalla Lana School of Public Health
University of Toronto
Toronto, Canada
e-mail: wxu@uhnres.utoronto.ca

Xin Gao†

Department of Mathematics and Statistics
York University
Toronto, Canada
e-mail: xingao@mathstat.yorku.ca

Abstract: We consider the problem of both prediction and model selection in high dimensional generalized linear models. Predictive performance can be improved by leveraging structure information among predictors. In this paper, a graphic model-based doubly sparse regularized estimator is discussed under the high dimensional generalized linear models, that utilizes the graph structure among the predictors. The graphic information among predictors is incorporated node-by-node using a decomposed representation and the sparsity is encouraged both within and between the decomposed components. We propose an efficient iterative proximal algorithm to solve the optimization problem. Statistical convergence rates and selection consistency for the doubly sparse regularized estimator are established in the ultra-high dimensional setting. Specifically, we allow the dimensionality grows exponentially with the sample size. We compare the estimator with existing methods through numerical analysis on both simulation study and a microbiome data analysis.

MSC2020 subject classifications: Primary 62J07, 62J12; secondary 62F12.
Keywords and phrases: Undirected graph, regularization, statistical consistency, model selection, random design, proximal operator.

Received September 2020.

*The work of Wei Xu was funded by Natural Sciences and Engineering Research Council of Canada (NSERC Grant RGPIN-2017-06672), Crohn's and Colitis Canada (CCC Grant CCC-GEMIII), and Helmsley Charitable Trust. The work of Xin Gao was supported by the Natural Sciences and Engineering Research Council of Canada.
†Co-corresponding authors.

Contents

1. Introduction

We consider a regularization method of high-dimensional generalized linear model (GLM) [23] where the dimensionality greatly surpasses the sample size. GLM is one of the most commonly used statistical methods for modeling, estimation, prediction and classification. It has been widely used in high dimensional data analysis. Many traditional statistical tools are not well-suited for the ultra-high dimensional data. Regularization methods have been widely used in the literature. Hoerl and Kennard [12] proposed the ridge regression which uses a ridge penalty to improve the estimation efficiency through a bias-variance trade-off. Tibshirani [39] proposed the Lasso regression which includes the ℓ_1 penalty for both shrinkage and variable selection. Many theoretical properties of ℓ_1 penalty for the high-dimensional GLM have been established, ranging from estimation consistency [26], selection consistency [4], persistence property for prediction [11] and risk consistency [41]. Other methods penalize the likelihood function with folded nonconvex penalty functions including the smoothly clipped absolute deviation (SCAD) [7, 8], the adaptive Lasso penalty [52] and the minimum convex penalty (MCP) [49]. The elastic net method was proposed by [53] to perform variable selection where the variables could be highly correlated. A more generalized Lasso penalty was proposed by [40] for some prespecified modifying variables.

If the true sparsity structure comprises clusters or groups of predictors, one can use group Lasso to select the coefficients [48, 27, 14, 10]. [44] discussed the hierarchical sparse modeling which utilized the group Lasso and the latent overlapping group Lasso penalty. Different from these works, we consider an undirected graph structure among the predictors [46, 35, 51]. As predictors in a neighborhood are connected, they are simultaneously effective or not effective

for predicting the response. As an example, in the diagnosis of the metastatic melanoma using commensal microbial composition information, the patients' microbiomes are naturally correlated. Relevant microbiomes in the same neighborhood of the underlying graph often either influence or not influence the clinical response together, see [22]. Incorporating the structure information among the microbiomes can lead to the construction of a better classifier.

In literature, many methods have utilized the edge-by-edge information in the graph to solve the regression problem. For example, the method OSCAR in [2] used the ℓ_∞ penalty for every pairs of predictors and the method GRACE in [16] used the network-constrained penalty on the pairwise differences of the connected predictors. Yu and Liu [46] proposed the sparse regression method incorporating graph structure (SRIG) with a node-wise neighborhood based penalty where the penalty term is distributed over all nodes instead of all edges. In addition, they proposed an efficient computational method to solve the node-wise penalty. Liu et al. [18] proposed a graph-based high dimensional sparse linear discrimination analysis method which is specific for classification. Recently, Zhou et al. (2019) extended the sparse regression leveraging graphical structure to generalized linear models and establish the estimation error and prediction error of the penalized estimator. Yu and Liu [46] and Zhou et al. [51] assumed that all the components within each decomposition would be shrunk to zero simultaneously. This assumption is restrictive and can be further relaxed. It is known that while the predictors are correlated with each other in a neighborhood, they may not be all important predictors to the response variable. To overcome this, Stephenson et al. [36] proposed a doubly sparse method (DSRIG) which encourages sparsity both within and among the decomposition under linear regression model with fixed design matrix. They also applied the doubly sparse method to the logistic regression [37] using the predictor duplication (PD) algorithm [27]. But the PD method is not very efficient and requires large computing memory for models with high dimensional predictors or predictor graphs having large number of edges. Thus it is necessary to develop new efficient optimization algorithm for doubly sparse graphic model-based GLMs. Furthermore, no work has been done on finite sample bounds of the estimation error and the model selection consistency in graphic model-based doubly sparse generalized linear models. Accordingly, there is a great need to investigate these statistical properties.

In this paper, theoretical investigation is presented for the doubly sparse high-dimensional GLM estimators incorporating the graph structure through a node-wise penalty. In general, the graphic structure can be either given or estimated from the study samples. We generalize the sparse least squares estimator of [46] to allow for a wider class of loss functions as well as a more general structured regularization. In terms of optimization, the constraint can be expressed as a latent sparse group Lasso over neighborhood sets. Besides using the predictor duplication method, we combine the fast iterative shrinkage thresholding algorithm (FISTA) [1] with the proximal splitting methods [47, 46] for solving the proximal operator. On the theoretical side, we establish the finite sample bounds of the optimal estimation, in addition with the prediction error

bound for random design with some regularity conditions. The classical result of the estimation error bound of the penalized GLMs can be recovered if there is no edge in the predictor graph. Moreover, the model selection consistency is established for the ultra-high dimensional graphic model-based doubly sparse GLMs. In the simulation studies and the real data application, we show that the method can improve the performance in the aspects of estimation, prediction, and model selection compared with the regularization methods without using the predictors' graphic structure.

The paper is organized as follows. In Section 2, we set up the basic notation and introduce the penalization method. In Section 3, an efficient algorithm for the optimization problem is presented. In Section 4, the main theoretical results are provided. In Section 5 and 6, numerical simulations and an application on a human microbiome dataset demonstrate the competitive performance of the method. We provide some discussions in Section 7. Technical proofs are contained in Appendix.

Notation Let $\mathbb{R}^p$ denote the p-dimensional real Euclidean space. Let $f(n) \precsim g(n)$ indicate $f(n) \leq cg(n)$ for some positive constant; let $f(n) \succsim g(n)$ indicate $f(n) \geq c'g(n)$ for some positive constant c'; $f(n) \asymp g(n)$ means that $f(n) \precsim g(n)$ and $f(n) \succsim g(n)$ both hold true. Let $\mathbf{v}_S \in \mathbb{R}^S$ denote the vector $\mathbf{v} \in \mathbb{R}^p$ restricted to a subset $S \subseteq \{1, \ldots, p\}$. For any vector $\mathbf{x}$, let $\|\mathbf{x}\|_q = (\sum_j |x_j|^q)^{1/q}$ denote the L_q-norm of $\mathbf{x}$ with $1 \leq q \leq \infty$. For a matrix M, let $|\!|\!|M|\!|\!|_2$ denote the spectral norms, and let $|\!|\!|M|\!|\!|_{\max} := \max_{i,j} |m_{ij}|$ denote the elementwise ℓ_∞-norm of matrix M. Let ∇h denote a gradient or subgradient for any function $h : \mathbb{R}^p \to \mathbb{R}$. Let $\text{supp}(V)$ indicate the support of the vector V, and let $|\mathcal{N}|$ indicate the cardinality of the set $\mathcal{N}$. Let $\mathbb{B}_q(r)$ represent the centered ball of radius r in ℓ_q norm for $q, r > 0$. Let $\text{sign}(\cdot)$ represent the sign function.

2. Methodology

We consider a common setting of GMLs. Let $\{(X_i, y_i); i = 1, \ldots, n\}$ denote independent and identically distributed (i.i.d) samples, where y_i is a response variable and $X_i = (x_{i,1}, \ldots, x_{i,p})^{\mathrm{T}}$ is a p-dimensional covariate vector. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$. Throughout the paper, the dimensionality p is allowed to grow with the sample size n. It is assumed that the conditional density of y_i given X_i is from the exponential family,

$$f(y_i|X_i, \boldsymbol{\beta}, \phi) = \exp\left(\frac{1}{\phi}(y_i\theta_i - b(\theta_i)) + c(y_i, \phi)\right)$$

with the functions $b(\cdot)$ and $c(\cdot, \cdot)$, the nuisance parameter ϕ, and the canonical parameter $\theta_i = X_i^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the p-dimensional unknown regression coefficient. By standard properties of exponential families [23], we have $E(\boldsymbol{y}|X_i) = b'(\theta_i) = \mu_i$. The canonical link function $g(\mu_i) = \theta_i$ is used.

In our analysis, it is assumed that $b''(\cdot) \leq c$ for some constant $c > 0$. This boundedness condition implies that y_i has a bounded conditional variance. This

condition is required to establish the estimation error bound. Similar condition was assumed in [30], which is satisfied in many models including linear regression, logistic regression, and multinomial regression. Even though it does not hold for Poisson regression, a truncated Poisson regression model will satisfy this boundedness condition [45].

The population loss based on the negative log likelihood is formulated as

$$\mathcal{L}(\boldsymbol{\beta}) = -\mathbb{E}\log(\mathbb{P}(X_i, y_i)) = -\mathbb{E}(\log(\mathbb{P}(X_i))) - \frac{1}{\phi}\mathbb{E}[y_i X_i^{\mathrm{T}}\boldsymbol{\beta} - b(X_i^{\mathrm{T}}\boldsymbol{\beta})].$$

The empirical loss function takes the form $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{\phi}\cdot\frac{1}{n}\sum_{i=1}^n[b(X_i^{\mathrm{T}}\boldsymbol{\beta}) - y_i X_i^{\mathrm{T}}\boldsymbol{\beta}]$, and the population-level and empirical gradients are given by

$$\nabla\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbb{E}[(b'(X_i^{\mathrm{T}}\boldsymbol{\beta}) - y_i)X_i], \text{ and } \nabla\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{\phi}\cdot\frac{1}{n}\sum_{i=1}^n[(b'(X_i^{\mathrm{T}}\boldsymbol{\beta}) - y_i)X_i].$$

Without loss of generality, we assume the nuisance parameter $\phi = 1$ for the rest of the paper. It can be verified that $\nabla\mathcal{L}(\boldsymbol{\beta}^0) = 0$ for the true parameter $\boldsymbol{\beta}^0$ of the GLMs. We assume

$$\nabla^2\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n b''(X_i^{\mathrm{T}}\boldsymbol{\beta})X_i X_i^{\mathrm{T}} \succeq 0,$$

so $\mathcal{L}_n$ is convex.

Assume that the i.i.d. p-dimensional random variables $X_1, \ldots, X_p$ follow some multivariate distribution with a zero mean and a covariance matrix Σ. Denote the precision matrix $\Omega = \Sigma^{-1}$, which contains elements ω_{jk} and assumed to be sparse. Let G be an undirected predictor graph over the set of nodes $\mathcal{J} = \{1, 2, \ldots, p\}$. The edge set E in the graphical model represents the conditional dependence structure of the observed variables. When two predictors are conditionally independent, they are not connected by an edge in the graph, i.e., $e_{jk} \notin E \iff X_j \perp\!\!\!\perp X_k \mid \{X_l : l \neq j, k\}$. Specifically, in a Gaussian graphical model, there will be edges between any pair of nodes $(j, k), j \neq k$, if $\omega_{jk} \neq 0$. In discrete graphical models which can be represented by a minimal exponential family, Loh and Wainwright [20] investigated the connection between the support of a generalized inverse covariance matrix and the conditional independence structure of the graph. In particular, they showed that for binary variables, the inverse of the usual covariance matrix corresponds exactly to the edge structure of the tree. For continuous but non-Gaussian distribution, Spantini et al. [34] showed that if $X_1, \ldots, X_p$ have a smooth and strictly positive density $\pi(\mathbf{x})$, the pairwise conditional independence of the random variables X_j and X_k can be assessed by $X_j \perp\!\!\!\perp X_k \mid \{X_l : l \neq j, k\} \iff \partial^2_{j,k}\log\pi(\mathbf{x}) = 0$, where Ω^* denotes the generalized precision matrix with elements $\Omega^*_{jk} = \mathbb{E}_\pi[|\partial^2_{j,k}\log\pi(\mathbf{x})|]$. Then if $\Omega^*_{jk} = 0, j \neq k$, nodes j and k are conditionally independent, hence there's no corresponding edge (j, k) in G. Define $\mathcal{N}_j = \{k : e_{jk} \in E\}$ to be a neighborhood

set of the node j. Define the degree of node j be the size of its neighbourhood $d_j = |\mathcal{N}_j|$. For any node $k \notin \mathcal{N}_j$, if $e_{jk} \notin E$ and we assume that the node j will not contribute to the decomposition of β_k.

Given the predictor graph G, the neighbourhoods, $\mathcal{N}_j, j \in \mathcal{J}$, represent a set of groups which are possibly overlapping. As was discussed in [51], under the GLM settings, based on the inverse regression method from Theorem 2.1 of [6] and Condition 3.1 of [17] we have

$$\begin{aligned}\mathbb{E}(\mathbf{X} \mid Y = \boldsymbol{y}) &= \mathbb{E}\left[\mathbb{E}\left(\mathbf{X} \mid \boldsymbol{\beta}^0\mathbf{X}\right) \mid Y = \boldsymbol{y}\right] \\ &= \mathbb{E}\left[\left\{\boldsymbol{\mu} + \frac{\Sigma\boldsymbol{\beta}^0(\boldsymbol{\beta}^0)^T(\mathbf{X}-\boldsymbol{\mu})}{(\boldsymbol{\beta}^0)^T\Sigma\boldsymbol{\beta}^0}\right\} \mid Y = \boldsymbol{y}\right] \\ &= \boldsymbol{\mu} + \frac{\Sigma\boldsymbol{\beta}^0\mathbb{E}\left[(\boldsymbol{\beta}^0)^T(\mathbf{X}-\boldsymbol{\mu}) \mid Y = \boldsymbol{y}\right]}{(\boldsymbol{\beta}^0)^T\Sigma\boldsymbol{\beta}^0} \\ &= \boldsymbol{\mu} + \Sigma\boldsymbol{\beta}^0 k(\boldsymbol{y}),\end{aligned}$$

where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}), \Sigma = \mathrm{Var}(\mathbf{X})$ and $k(\boldsymbol{y}) = \frac{\mathbb{E}\left[(\boldsymbol{\beta}^0)^T(\mathbf{X}-\boldsymbol{\mu})|Y=\boldsymbol{y}\right]}{(\boldsymbol{\beta}^0)^T\Sigma\boldsymbol{\beta}^0}$. Let $\eta(\boldsymbol{y}) = \boldsymbol{\mu} + \Sigma\boldsymbol{\beta}^0 k(\boldsymbol{y})$, then

$$\boldsymbol{\beta}^0 \propto \Sigma^{-1}(\eta(\boldsymbol{y}) - \boldsymbol{\mu}) = \Omega(\eta(\boldsymbol{y}) - \boldsymbol{\mu}), \tag{2.1}$$

where Ω is the precision matrix. Let the predictor graph G, be denoted by a $p \times p$ adjacency matrix E, where $E_{jk} = 1$ for connected predictors j and k and $E_{jk} = 0$ otherwise. We always set $E_{jj} = 1$ for each j. Then, we have $N_j = \{k : E_{jk} = 1\}$. Base on the connection between the conditional independence and the graphical structure, by (2.1), $\boldsymbol{\beta}^0$ can be decomposed into

$$\begin{aligned}\beta_1^0 &= \Theta_1^{(1)}E_{11} + \Theta_1^{(2)}E_{12} + \cdots + \Theta_1^{(j)}E_{1j} + \cdots + \Theta_1^{(p)}E_{1p}, \\ \beta_2^0 &= \Theta_2^{(1)}E_{21} + \Theta_2^{(2)}E_{22} + \cdots + \Theta_2^{(j)}E_{2j} + \cdots + \Theta_2^{(p)}E_{2p}, \\ &\vdots \\ \beta_p^0 &= \Theta_p^{(1)}E_{p1} + \Theta_p^{(2)}E_{p2} + \cdots + \Theta_p^{(j)}E_{pj} + \cdots + \Theta_p^{(p)}E_{pp},\end{aligned} \tag{2.2}$$

where the term $\{\Theta_k^{(j)} : k \in \mathcal{N}_j\}$ arises from the marginal correlation between the predictor j and the response. Let $V_k^{(j)} = \Theta_k^{(j)}E_{kj}$ for $k \in \mathcal{N}_j$, then the decomposition of the true coefficients $\boldsymbol{\beta}^0$ can be expressed as:

$$\begin{aligned}\beta_1^0 &= V_1^{(1)} + V_1^{(2)} + \cdots + V_1^{(j)} + \cdots + V_1^{(p)}, \\ \beta_2^0 &= V_2^{(1)} + V_2^{(2)} + \cdots + V_2^{(j)} + \cdots + V_2^{(p)}, \\ &\vdots \\ \beta_p^0 &= V_p^{(1)} + V_p^{(2)} + \cdots + V_p^{(j)} + \cdots + V_p^{(p)},\end{aligned} \tag{2.3}$$

where $V^{(j)} = (V_1^{(j)}, \ldots, V_p^{(j)})^{\mathrm{T}}$ depends on the interaction among predictors (e.g., graphical structure) and the marginal correlation between the predictors

and the response. The term $\{V_k^{(j)} : k \in \mathcal{N}_j\}$ contains the contribution of the predictor k to the response through the predictor j. Such contribution depends on the strength of the conditional dependency relationship between j and k and the marginal correlation between the predictor j and the response variable. From the derivation in (2.3), $V_k^{(j)} = \Theta_k^{(j)} E_{kj}$. If $V_k^{(j)} = 0$, then either there's no edge between the predictor j and k (e.g., $E_{kj} = 0$), or even though there's an edge between the predictor j and k (e.g., $E_{kj} = 1$) for each $k \in \mathcal{N}_j$, but the predictor j and the response variable are uncorrelated (e.g., $\Theta_k^{(j)} = 0$).

Therefore, we assume a doubly sparse decomposition to help mitigate estimation bias, which means there are only a small number of vectors $V^{(j)}$s that are nonzero, and even for the predictors with nonzero vectors $V^{(j)}$, there are a small number of nonzero $V_k^{(j)}$s within the vector $V^{(j)}$s. The underlying graph G is assumed to be known or estimated from data. Therefore, under the GLM setting and given the predictor graph G with neighborhoods $\mathcal{N}_1, \ldots, \mathcal{N}_p$, we optimize the following penalized maximum likelihood objective function (2.4) which induce sparsity both between and within $V^{(j)}$, $j = 1, \ldots, p$.

$$\min_{\boldsymbol{\beta}, V^{(1)}, \ldots, V^{(p)}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda \left(\sum_{j=1}^{p} \left[\tau_j \|V^{(j)}\|_2 + \xi \|V^{(j)}\|_1 \right] \right),$$
$$\text{subject to } \boldsymbol{\beta} = \sum_{j=1}^{p} V^{(j)}, \quad \text{supp}(V^{(j)}) \subseteq \mathcal{N}_j, \tag{2.4}$$

where $\lambda \geq 0$ is the tuning parameter, τ_j is the positive group-specific weight, and ξ is the mixing parameter which can be viewed as a trade-off weight that balances the contributions of the ℓ_1 and ℓ_2 norms. We assume τ_j and ξ are bounded. Different values of ξ correspond to different shapes of the constraints [28]. We use the weight of the form $\tau_j \propto d_j^{\gamma}$ with $0 < \gamma < 1/2$ and $d_j = |\mathcal{N}_j|$, the size of the neighborhood. Here, we set $\gamma < 1/2$ for the possible overlapping groups. This was also suggested in [27] and [51]. The penalty function $\mathcal{R}(\boldsymbol{\beta}) := \sum_{j=1}^{p} [\tau_j \|V^{(j)}\|_2 + \xi \|V^{(j)}\|_1]$ can be viewed as the sparse group Lasso penalty with additional constraints. The choice of tuning parameters will be discussed in Section 5. The L_1 component of (2.4) will control the sparsity within $V^{(j)}$ while the L_2 component of (2.4) will control the sparsity among the neighbourhoods. It will be reduced to a sparse GLM incorporating graphic structure discussed in [51] when $\xi = 0$. When the graph G has no edges, then the objective function (2.4) will reduce to that of the GLM Lasso [41, 25]. Under the GLM settings, $\mathcal{L}_n(\boldsymbol{\beta})$ is smooth, while $\mathcal{R}(\boldsymbol{\beta})$ is not smooth. Note that the penalty function proposed in (2.4) is similar to a sparse group Lasso [32], but it shrinks the decomposition $V^{(j)}$s rather than directly penalizes the regression coefficients $\boldsymbol{\beta}$.

3. Computation

Typically, the problem (2.4) can be transformed into a sparse group Lasso problem [32] and the predictor duplication (PD) method [27, 46] can be used to solve

the problem. However, the PD method is not efficient because of large memory requirement.

Therefore, we combine the FISTA with a proximal splitting method to solve the problem which is stable and efficient. Given the positive weights τ_j and ξ, for $\boldsymbol{\beta} \in \mathbb{R}^p$, denote

$$\|\boldsymbol{\beta}\|_{G,\tau} = \min_{\boldsymbol{\beta}=\sum_{j=1}^p V^{(j)}} \sum_{j=1}^p \tau_j \|V^{(j)}\|_2 \text{ and } \|\boldsymbol{\beta}\|_{S,\xi} = \min_{\boldsymbol{\beta}=\sum_{j=1}^p V^{(j)}} \sum_{j=1}^p \xi \|V^{(j)}\|_1,$$

where $\text{supp}(V^{(j)}) \subseteq \mathcal{N}_j$. Then, $\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{G,\tau} + \|\boldsymbol{\beta}\|_{S,\xi}$ is a norm and therefore convex. The optimization problem (2.4) is equivalently to the following problem

$$\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda(\|\boldsymbol{\beta}\|_{G,\tau} + \|\boldsymbol{\beta}\|_{S,\xi}). \tag{3.1}$$

For the exponential family of distributions, the loss function $\mathcal{L}_n(\cdot)$ is a continuously differentiable function, then the gradient satisfies the Lipschitz property with constant L, i.e.,

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}_1) - \nabla \mathcal{L}_n(\boldsymbol{\beta}_2)\|_2 \leq L\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

As $\mathcal{L}_n$ is twice continuously differentiable, any bound on the operator norm of $\nabla^2 \mathcal{L}_n$ is a Lipschitz constant for $\nabla \mathcal{L}_n$. Then, a simple tight global upper bound for $\mathcal{L}_n$ is

$$\mathcal{L}_n(\boldsymbol{\beta}) \leq \mathcal{L}_n(\boldsymbol{\beta}^0) + (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^{\mathrm{T}} \nabla \mathcal{L}_n(\boldsymbol{\beta}^0) + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2^2.$$

Consider that $\mathcal{R}$ contains a non-smooth L_1 norm, a simple iterative strategy is to minimize the upper bound for $\mathcal{L}_n$ at each iteration, without modifying $\mathcal{R}$. At iteration $t+1$, the updated $\boldsymbol{\beta}^{(t+1)}$ is given by

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^{\mathrm{T}} \nabla \mathcal{L}_n(\boldsymbol{\beta}^{(t)}) + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2 + \lambda \mathcal{R}(\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta} - (\boldsymbol{\beta}^{(t)} - \frac{1}{L}\nabla \mathcal{L}_n(\boldsymbol{\beta}^{(t)}))\|_2^2 + \frac{\lambda}{L}\mathcal{R}(\boldsymbol{\beta}) \\
&= \text{prox}_{\frac{\lambda \mathcal{R}}{L}}\left(\boldsymbol{\beta}^{(t)} - \frac{1}{L}\nabla \mathcal{L}_n(\boldsymbol{\beta}^{(t)}\right),
\end{aligned}$$

where the associated proximal operator is defined as

$$\text{prox}_{\frac{\lambda \mathcal{R}}{L}}(h) = \arg\min_{\boldsymbol{\beta}} \frac{\|h - \boldsymbol{\beta}\|_2^2}{2} + \frac{\lambda}{L}\mathcal{R}(\boldsymbol{\beta}). \tag{3.2}$$

Thus to solve the optimization problem (3.1), we use the algorithm summarized in the following Algorithm 1.

Algorithm 1 FISTA for Regularized GLMs

Input: Initialize $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ and L is the largest eigenvalue of $\nabla^2 \mathcal{L}_n(\boldsymbol{\beta})$.

1: Take $Z^{(1)} = \boldsymbol{\beta}^{(0)}$ and $\alpha_1 = 1$, $t = 1$.
2: **repeat**
3: $\quad h^{(t)} = Z^{(t)} - \frac{1}{L}\nabla \mathcal{L}_n(Z^{(t)})$;
4: $\quad \boldsymbol{\beta}^{(t)} = \text{prox}_{\frac{\lambda}{L}(\|\boldsymbol{\beta}\|_{G,\tau} + \|\boldsymbol{\beta}\|_{S,\xi})}(h^{(t)})$;
5: $\quad \alpha_{t+1} = \frac{1+\sqrt{1+4\alpha_t^2}}{2}$;
6: $\quad Z^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{\alpha_t - 1}{\alpha_{t+1}}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)})$;
7: **until** convergence.

It is shown in [1] that the convergence rate to the optimal solution is of $O(1/t^2)$. With high dimensional p, the associated proximal step in Algorithm 1 is the most time consuming step. Following [47], we can solve the proximal step 4 by finding the projection via the Parallel Dykstra-like Proximal Method [5] and converting the minimization problem (3.2) into a series of convex problems which have closed form solutions. The details about Dykstra-like Proximal Splitting Method is shown in the supplementary material. Based on this method, we only need to know the graphical structure information about the neighbours $\mathcal{N}_j$ for $j = 1, \ldots, p$ to identify the overlap groups, which is pre-determined and unique. It is stable and very efficient for high-dimensional data, especially when the predictor graph are not very dense and can be decomposed into several disconnected components, see Example 1 in the simulations.

4. Theoretical properties

Denote $\mathcal{J}_0 = \{j : \beta_j^0 \neq 0\}$, $\mathcal{J}_0^c = \{j : \beta_j^0 = 0\}$, and $s_0 = |\mathcal{J}_0|$. As $\mathcal{R}(\boldsymbol{\beta})$ is convex and coercive, Rao et al. [28] proved that an optimal decomposition of $\boldsymbol{\beta}$ minimizing $\mathcal{R}(\boldsymbol{\beta})$ always exists but may not be unique. For each $\beta \in \mathbb{R}^p$, we denote $\mathfrak{V}(\beta)$ as the set of all optimal decompositions of $\boldsymbol{\beta}$. Define $\mathcal{K}(\beta) = \min_{(V^{(1)}, V^{(2)}, \cdots, V^{(p)}) \in \mathfrak{V}(\beta)} \left|\left\{j : \left\|V^{(j)}\right\|_2 \neq 0\right\}\right|$, which denotes the minimal number of nonzero $V^{(j)}$ among all the optimal decompositions of β. Denote $\mathcal{K} = \sup_{\text{supp }(\beta) \subset \mathcal{J}_0} \mathcal{K}(\beta)$ the maximum nonzero group among all decompositions. There are at most $d_{\max} = \max_{j=1,\ldots,p}\{d_j\}$ non-zero elements in all nonzero $V^{(j)}$s where $d_j = |\mathcal{N}_j|$. Note that the optimal decomposition of $\boldsymbol{\beta}$ is the smallest decomposition for the associated penalty $\mathcal{R}(\boldsymbol{\beta})$. In order to prove the main theorem, we first need to show the following property with regard to the subgradient conditions for the optimization problem (2.4). This property can be obtained directly by the Karush-Kuhn-Tucker (KKT) conditions.

Proposition 4.1. *The necessary and sufficient condition for $\widehat{\boldsymbol{\beta}}$ being the solution of (2.4) is that $\widehat{\boldsymbol{\beta}}$ can be decomposed as $\widehat{\boldsymbol{\beta}} = \sum_{j=1}^p V^{(j)}$ where $V^{(j)}$ satisfies that, for all $1 \leq j \leq p$, (i) $V^{(j)}_{\mathcal{N}_j^c} = 0$; (ii) $V^{(j)}_{\mathcal{N}_j} = 0$ and $\|\nabla_{\mathcal{N}_j} \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda h_{\mathcal{N}_j}\|_2 \leq \lambda\tau_j$ with $|h_k| \leq \xi$ for any $k \in \mathcal{N}_j$; (iii) $V^{(j)}_{\mathcal{N}_j} \neq 0$ and $\nabla_{\mathcal{N}_j} \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda\tau_j \frac{V^{(j)}_{\mathcal{N}_j}}{\|V^{(j)}_{\mathcal{N}_j}\|_2} +$*

$\lambda h_{\mathcal{N}_j} = 0$, *where for any* $k \in \mathcal{N}_j$ *either* $V_k^{(j)} = 0$ *and then* $|\nabla_k \mathcal{L}_n(\widehat{\boldsymbol{\beta}})| \leq \lambda\xi$, *or* $V_k^{(j)} \neq 0$ *and then* $h_k = \xi \operatorname{sign}(V_k^{(j)})$.

These subgradient conditions are similar to those of [32] which were established for a least squares loss function with group Lasso.

Assumption 1. *The neighborhood* $\mathcal{N}_j \subseteq \mathcal{J}_0$ *for each* $j \in \mathcal{J}_0$.

Assumption 2. *The covariates* $X_1, \ldots, X_n$ *are i.i.d. samples from a zero mean sub-Gaussian distribution with covariance matrix* $\operatorname{Cov}(X_i) = \Sigma$ *such that* $0 < \eta_l \leq \Lambda_{\min}(\Sigma)$. *Assume the Hessian of the cumulant function is uniformly bounded that* $\|b''\|_\infty \leq c$ *for some constant* $c > 0$.

Remark 4.1. *Assumption 1 implies that neighbors of important predictors are also important predictors to the modelling of the response. This condition is also assumed in [46]. Assumption 2 of sub-Gaussianality and low bounded eigenvalues is often used in high dimensional setting. This assumption was also used to derive a lower bound of the Taylor-series error of the GLM log-likelihoods with sub-Gaussian covariates in [26] and [43]. The boundedness assumption on cumulant function* $b(\cdot)$ *is also assumed in [26]. From Assumption 2, we can extend the restricted strong convexity (RSC) condition for GLMs [26, 43] to the current setting. That is, for any subset* $J \subset \{1, 2, \ldots, p\}$ *with* $|J| \leq s_0$, *and all the optimal decompositions* $(V^{(1)}, \ldots, V^{(p)})$ *of any vector* Δ *that* $\|\Delta\|_2 \leq 1$, *we have for some* $\kappa_l > 0$ *and* $\kappa_2 \geq 0$,

$$\mathcal{L}_n(\boldsymbol{\beta}^0 + \Delta) - \mathcal{L}_n(\boldsymbol{\beta}^0) - \nabla\mathcal{L}_n(\boldsymbol{\beta}^0)^T\Delta \geq \kappa_l \sum_{j \in J} (\tau_j + \sqrt{d_j}\xi)^2 \|V^{(j)}\|_2^2 - \kappa_2 \frac{\log p}{n} \mathcal{R}^2(\Delta),$$

with probability tending to 1.

In the following, we establish the error bounds of the doubly sparse GLM estimator.

Theorem 4.1. *Under the Assumptions 1-2, let* $\tau_{\min} = \min_{1 \leq j \leq p}\{\tau_j\}$. *If we choose*

$$\lambda(\tau_{\min} + \xi) \geq c_0 \sqrt{\frac{d_{\max} \log p}{n}}$$

with a positive constant c_0 *and* $\mathcal{R}(\boldsymbol{\beta}^0) \leq \rho$ *with* $\rho \leq c' \frac{\sqrt{d_{max}}}{\tau_{\min} + \xi} \sqrt{\frac{n}{\log p}}$ *for some positive constant* c', *then with sample size* $n \gtrsim s_0 \log p$, *any optimal solution* $\widehat{\boldsymbol{\beta}}$ *of problem (2.4) satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \leq \frac{\lambda\mathcal{K}}{4\kappa_l(\tau_{\min} \wedge \xi)}, \qquad \mathcal{R}(\widehat{\Delta}) \leq \frac{2\lambda\mathcal{K}}{\kappa_l},$$

with probability greater than or equal to $1 - c_1 \exp(-c_2 \log p)$ *for some* $c_1, c_2 > 0$. *Here,* κ_l *is a positive constant that depends on* $\|\boldsymbol{\beta}^0\|_2$, $b(\cdot)$, $\Lambda_{\min}(\Sigma)$, *and the*

sub-Gaussian parameters of the $X_i, i = 1, \ldots, n$.

From the results above, we can see that our theoretical results do not depend on any uniqueness assumption on the decomposition of $\boldsymbol{\beta}$ that minimizes $\|\boldsymbol{\beta}\|_{G,\tau} + \|\boldsymbol{\beta}\|_{S,\xi}$. In contrast to [27], our result depends only on $\mathcal{K}$ which represents the maximal structured sparsity of such decompositions. We consider the constraint $\mathcal{R}(\boldsymbol{\beta}^0) \leq \rho$ to ensure the existence of local/global optima which was also discussed in Loh and Wainwright [21]. Note that the setting $\rho \leq c' \frac{\sqrt{d_{max}}}{\tau_{\min}+\xi} \sqrt{\frac{n}{\log p}}$ is feasible based on the assumption of sample size $n \gtrsim s_0 \log p$. The following corollary from Theorem 4.1 provides the prediction error bound, which is defined as $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \langle \nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \rangle$. Note that under our GLM model setting, this error measure $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is equivalent to the symmetrized Bregman divergence defined by the cumulant function $b(\cdot)$ [21].

Corollary 4.1. *Under the same assumptions as Theorem 4.1, the prediction error is bounded by*

$$\langle \nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \rangle \leq \frac{3\lambda^2 \mathcal{K}}{\kappa_l}.$$

Remark 4.2. *Note that when there is no edge in the predictor graph* G *and* $\xi = \tau_i = 1$ *for each* i, *we have* $\mathcal{K} = s_0$ *and* $\mathcal{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1$. *If we choose* $\lambda \asymp \sqrt{\log p/n}$, *then* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \precsim s_0 \sqrt{\log p/n}$ *with probability at least* $1 - c_1 \exp(-c_2 \log p)$ *for some* $c_1, c_2 > 0$. *In this case, we recover the rate in [21] and [19] for high dimensional regularized* M*-estimators. However, from Theorem 4.1, the error bounds of our method depend on the minimal number of nonzero optimal decomposition* $\mathcal{K}$ *rather than the true model size* s_0. *If the true graph* G *consists of disconnected complete sub-graphs and* $\mathcal{J}_0$ *is the union of* K_0 *node sets of those disconnected subgraphs, see Example 1 in Section 5, then* $\mathcal{K} = K_0$. *Our method gives better results compared to the Lasso if* K_0 *is much smaller than* s_0. *This demonstrates our method's advantage over standard Lasso procedure when the structure sparsity of the predictor graph is incorporated. From the simulations, we find that our method indeed outperforms the GLM Lasso in all the simulation settings. Moreover, in Corollary 4.1, considering the fixed design linear regression, the expression* $\langle \nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \rangle$ *corresponds to the commonly used error measure* $\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n$.

Next, we establish the model selection consistency of the doubly sparse high-dimensional GLM estimation.

Assumption 3. *The number of nonzero coefficients* $s_0 = O(n^{\delta_1})$ *and* $\log p = O(n^{\delta_2})$, *and* $0 < 3\delta_1 + \delta_2 < 1$ *for some constant* δ_1 *and* δ_2.

Assumption 4. *The subset of the Fisher information matrix corresponding to the relevant covariates has bounded eigenvalues, i.e.* $\Lambda_{\min}(\mathcal{Q}_{\mathcal{J}_0 \mathcal{J}_0}) \geq C_{\min}$ *for some constant* $C_{\min} > 0$.

Assumption 5. *There exists a constant* $\zeta \in (0,1)$ *such that* $\|\mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0}(\mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_\infty \leq 1-\zeta$.

Note that Assumption 3 allows for graphs and sample sizes in the "large p, small n" regime, as long as the degrees are bounded, or grow at a sufficiently slow rate. Assumption 4 ensures that the Fisher information matrix of the relevant covariates is not singular. Assumption 5 refers to the incoherence condition which requires that the large number of irrelevant covariates cannot be strongly correlated with the subset of relevant covariates. The incoherence condition is also assumed in [25] to obtain the ℓ_2-norm consistency of Lasso for fixed designs. Analogous conditions are required for the success of the Lasso in the case of high dimensional graphical model [24, 50, 29]. Under these assumptions, the proposed method is model selection consistent for high dimensional setting.

Theorem 4.2. *Under Assumptions 1-5, suppose the weight* $\tau_j = \sqrt{d_j}m_j$ *for each* j, *where* $\sqrt{s_0}(\sqrt{s_0}\max_{j\in\mathcal{J}_0} m_j \vee \xi) = o(\min_{\mathcal{J}_0^c} m_j)$, *the tuning parameters* λ *and the minimum absolute nonzero coefficient* $\beta^0_{\min} = \min_{j\in\mathcal{J}_0}|\beta^0_j|$ *satisfy that,*

$$\beta^0_{\min} > \frac{\sqrt{s_0}}{C_{\min}}\left(\sqrt{\frac{\log p}{n}} + \lambda(\max_{j\in\mathcal{J}_0}\tau_j + \xi)\right),$$

and

$$\frac{s_0^{1/2}\lambda\left(\mathcal{K}/(\tau_{\min}\wedge\xi)\right)^2 + \xi}{\min_{j\in\mathcal{J}_0^c} m_j} \to 0,$$

Then as $n\to\infty$ *and* $n \gtrsim s_0^3\log p$, *there exists a solution* $\widehat{\boldsymbol{\beta}}$ *of problem (2.4) such that* $\operatorname{sign}(\widehat{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta}^0)$ *with probability tending to* 1.

Remark 4.3. *Note that the quantities* τ_j *and* d_j *depend on* n. *If we set the tuning parameter* $\lambda \asymp \sqrt{s_0\log p/n}$, *then* $\beta^0_{\min}$ *has to be greater than* $C_1 n^{(3c_1+c_2-1)/2}$ *for some constants* $C_1 > 0$ *and* $0 < 3c_1 + c_2 < 1$ *to ensure the model selection consistency. Compare to the estimation consistency in Theorem 4.1, we need the assumption* $n \gtrsim s_0^3\log p$. *The extra factor of* s_0^2 *is to ensure the consistency of the sample Fisher information matrix. Note that the theoretical results in this section requires the assumption that the true graph* G *is known. If the graph* G *is unknown, then a data splitting procedure can be used to first estimate the graph* G *and then the graph can be used to estimate the coefficient parameters.*

5. Simulation study

In this section, we conduct simulations to compare the performance of the graphic model-based doubly sparse generalized linear model (GDSGLM) with other existing penalized methods, such as the Lasso, the ridge regression, the elastic net (Enet), and the recently proposed method sGLMg [51]. Throughout the simulation studies, we assume that the covariates are normally distributed

with zero mean. Different covariance structure are explored and the graph G is given by the estimated precision matrix. We denote GDSGLM-O and sGLMg-O as the GDSGLM and sGLMg methods with the knowledge of the true graphs. We also present the so-called oracle estimation GLM-O by the maximum likelihood method based on the true subset of the predictors but not utilizing the graph structure.

In general, the tuning parameter λ, the positive group-specific weight γ, and the mixing parameter ξ in (2.4) can be chosen by a cross-validation (CV) procedure, like [15], but this may be time consuming for high dimensional settings. Therefore, throughout the simulations, we simplify the cross-validation to select one tuning parameter λ, while setting $\tau_j = d_j^\gamma$ with $\gamma = \log(2)/\{2\log(3)\}$, which is suggested in [27, 51], and $\xi = \max(\tau_j)$.

In the simulations, we generate the data from the logistic regression model, that is $\boldsymbol{y}|\mathbf{X} \sim$ Bernoulli $\{p(\mathbf{X}\boldsymbol{\beta})\}$ and

$$p(\mathbf{X}\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}.$$

For each example, our simulated data are divided into three sets such that a training set, an independent validation set, and an independent test set. All model fitting is based on training set. As we assume the predictors follow the multivariate normal distribution, we estimate the graph structure by the graphical Lasso method [9] using the training set, where the tuning parameter for the graphical lasso is selected by `huge.select` function in R software. We use the validation set to select the tuning parameters and use the independent test set to compare different methods. For each example, we set the dimension $p = 100$. We consider two cases of the sample size n: (I) 80/80/500, (II) 120/120/500, where ./././ denote the sample sizes of the training sets, the validation sets, and the independent test sets, respectively. To make comparisons, we consider the following three different predictor structures in [46] and [51].

Example 1 (block diagonal Ω)**.** *Let $p = 100$, $s_0 = 15$, and $\boldsymbol{\beta}^0 = (1, 1, \ldots, 1, 0, 0, \ldots, 0)^T$. We generate the predictors as $X_j = W_i + 0.4\epsilon_j, W_i \sim N(0,1), i = [j/5], 1 \le j \le 15$, where $[.]$ is the ceiling function, and $X_j \sim i.i.d\, N(0,1), 16 \le j \le 100$, and $\epsilon_j \sim i.i.d\, N(0,1)$, $j = 1, 2, \ldots, 15$.*

Example 2 (banded Ω)**.** *Let $p = 100$, and $\boldsymbol{\beta}^0$ is the same as in Example 1. The predictors have a multivariate normal distribution with zero mean and covariance Σ with $\Sigma_{ij} = 0.5^{|i-j|}$, $\omega_{ii} = 1.333$, $\omega_{ij} = -0.667$ for $|i-j| = 1$ and ω_{ij} is zero otherwise.*

Example 3 (sparse Ω)**.** *Let $p = 100$, and The predictors have a multivariate normal distribution with zero mean and covariance $\Sigma = \Omega^{-1}$, where $\Omega = B + \delta I$, $B_{ii} = 0$, $B_{ij} \sim 0.5 * Binom(1, 0.05)$, $i \ne j$. To ensure that Ω's conditional number equals p, we choose appropriate δ. We standardize Ω so that $\Omega_{ii} = 1$. We set $\boldsymbol{\beta}^0 = \Omega\boldsymbol{\eta}$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^{\mathrm{T}}$ with $\eta_i = 5, i = 1, 2, 3, 4$, for the top four largest degrees of predictors graph and $\eta_i = 0$ otherwise.*

As we assume X_is are sub-Gaussian vectors, we consider another simulation setting with non-Gaussian X_is following a uniform distribution.

Example 4 (block diagonal Ω)**.** *Let $p = 100$, and $\boldsymbol{\beta}^0$ is the same as Example 1. We generate the predictors as $X_j = W_i + 0.75\epsilon_j, W_i \sim$ Uniform$[-1, 1], i = [j/5], 1 \le j \le 15$, where $[.]$ is the ceiling function, and $X_j \sim i.i.d$ Uniform$[-1, 1]$, $16 \le j \le 100$, and $\epsilon_j \sim i.i.d$ Uniform$[-1, 1]$, $j = 1, 2, \ldots, 15$.*

We adopt the following measures to compare the performance of the different approaches: the estimation consistency by ℓ_2 distance $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$; the prediction misclassification error ($\mathcal{P}$), which is based on a test set; and false positive rate (FPR) and false negative rate (FNR) for variable selection accuracy. All the true predictor graphs (defined by Ω) in the above three examples are displayed in Figure 1. We simulate 50 data sets for each example.

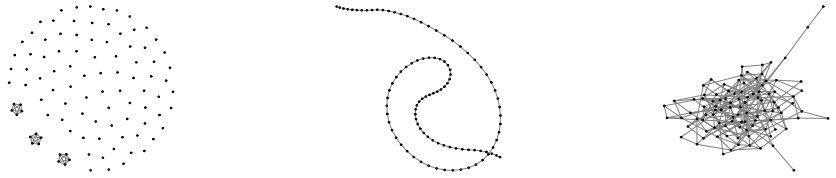

FIG 1. *True predictor graphs of three simulation examples.*

Tables 1 and Table 2 demonstrate the simulation results for the first example. The results shows that the GDSGLM method has the best performance of the estimation consistency, the misclassification error, the false positives rates and the false negatives rates among all the competing methods. The performances of sGLMg-O and GDSGLM-O are close to those of sGLMg and GDSGLM because the estimated graph is very accurate. The performance of GLM-O method is not very good due to the high correlation among predictors.

TABLE 1
Performance comparison of estimation and prediction for Example 1.

	L_2 distance		Error	
Methods	(I)	(II)	(I)	(II)
GLM-O	2.650 (0.104)	2.543 (0.103)	0.091 (0.004)	0.076 (0.003)
LASSO	3.125 (0.025)	2.970 (0.041)	0.128 (0.005)	0.113 (0.005)
Ridge	2.962 (0.034)	2.802 (0.030)	0.152 (0.005)	0.137 (0.004)
Enet	2.803 (0.047)	2.670 (0.042)	0.124 (0.004)	0.107 (0.004)
sGLMg	2.639 (0.051)	2.443 (0.049)	0.115 (0.004)	0.094 (0.004)
sGLMg-O	2.630 (0.051)	2.437 (0.049)	0.114 (0.004)	0.094 (0.004)
GDSGLM	2.069 (0.051)	2.010 (0.049)	0.075 (0.004)	0.069 (0.003)
GDSGLM-O	2.067 (0.051)	2.003 (0.049)	0.074 (0.004)	0.069 (0.003)

Tables 3 and Table 4 show the simulation results for the second example. The GDSGLM method and the sGLMg method outperform the other non-graph based methods. Moreover, by adding a ℓ_1 penalty, the proposed GDS-

TABLE 2
Performance comparison of model selection for Example 1.

	FPR		FNR	
Methods	(I)	(II)	(I)	(II)
GLM-O	-	-	-	-
LASSO	0.144 (0.011)	0.215 (0.015)	0.416 (0.028)	0.310 (0.018)
Ridge	0.990 (0.002)	0.990 (0.002)	0.000 (0.000)	0.000 (0.000)
Enet	0.271 (0.017)	0.337 (0.017)	0.083 (0.014)	0.040 (0.012)
sGLMg	0.181 (0.016)	0.117 (0.008)	0.000 (0.000)	0.000 (0.000)
sGLMg-O	0.180 (0.016)	0.115 (0.008)	0.000 (0.000)	0.000 (0.000)
GDSGLM	0.007 (0.003)	0.009 (0.003)	0.000 (0.000)	0.000 (0.000)
GDSGLM-O	0.004 (0.002)	0.003 (0.002)	0.000 (0.000)	0.000 (0.000)

TABLE 3
Performance comparison of estimation and prediction for Example 2.

	L_2 distance		Error	
Methods	(I)	(II)	(I)	(II)
GLM-O	2.064 (0.087)	1.719 (0.085)	0.126 (0.005)	0.117 (0.004)
LASSO	2.971 (0.035)	2.556 (0.044)	0.189 (0.005)	0.155 (0.004)
Ridge	2.963 (0.032)	2.736 (0.031)	0.201 (0.006)	0.181 (0.006)
Enet	2.841 (0.041)	2.544 (0.044)	0.178 (0.007)	0.153 (0.005)
sGLMg	2.655 (0.037)	2.374 (0.030)	0.165 (0.007)	0.139 (0.005)
sGLMg-O	2.571 (0.034)	2.306 (0.031)	0.152 (0.006)	0.133 (0.005)
GDSGLM	2.649 (0.036)	2.296 (0.054)	0.155 (0.006)	0.137 (0.004)
GDSGLM-O	2.559 (0.033)	2.202 (0.067)	0.133 (0.005)	0.117 (0.004)

TABLE 4
Performance comparison of model selection for Example 2.

	FPR		FNR	
Methods	(I)	(II)	(I)	(II)
GLM-O	-	-	-	-
LASSO	0.153 (0.006)	0.192 (0.011)	0.243 (0.012)	0.115 (0.016)
Ridge	0.994 (0.002)	0.992 (0.002)	0.000 (0.000)	0.000 (0.000)
Enet	0.350 (0.015)	0.316 (0.012)	0.073 (0.017)	0.048 (0.011)
sGLMg	0.544 (0.017)	0.501 (0.015)	0.033 (0.006)	0.008 (0.004)
sGLMg-O	0.333 (0.013)	0.389 (0.014)	0.046 (0.010)	0.017 (0.005)
GDSGLM	0.538 (0.015)	0.487 (0.018)	0.033 (0.006)	0.012 (0.005)
GDSGLM-O	0.324 (0.016)	0.307 (0.044)	0.032 (0.009)	0.000 (0.000)

GLM method is better than the sGLMg method and it achieves the smallest misclassification error and the lowest FPR and FNR.

Tables 5 and 6 display the comparative results for the third example. The results are consistent with the previous two examples. Compared with the sGLMg (sGLMg-O) method, the GDSGLM (GDSGLM-O) method has smaller estimation errors of the $\boldsymbol{\beta}$ and smaller misclassification errors for the prediction due to the additional ℓ_1 penalty. Particularly, in the estimated predictor graph setting, GDSGLM method even offers improved performance over sGLMg-O. We notice that the GDSGLM method has a much smaller FPR with slightly higher FNR compared to sGLMg. Tables 7 and 8 display the comparative results for the non-Gaussian example. It is demonstrated that our method still outperforms all the other methods for this non-Gaussian setting.

TABLE 5
Performance comparison of estimation and prediction for Example 3.

Methods	L_2 distance		Error	
	(I)	(II)	(I)	(II)
GLM-O	2.622 (0.106)	2.038 (0.113)	0.148 (0.005)	0.135 (0.004)
LASSO	5.623 (0.079)	4.769 (0.116)	0.378 (0.014)	0.266 (0.009)
Ridge	6.048 (0.004)	6.017 (0.005)	0.464 (0.004)	0.462 (0.005)
Enet	5.954 (0.024)	5.677 (0.050)	0.433 (0.010)	0.356 (0.008)
sGLMg	4.296 (0.116)	3.780 (0.129)	0.258 (0.007)	0.189 (0.005)
sGLMg-O	3.987 (0.094)	3.123 (0.116)	0.214 (0.005)	0.149 (0.005)
GDSGLM	3.422 (0.172)	3.108 (0.238)	0.210 (0.005)	0.178 (0.006)
GDSGLM-O	2.735 (0.224)	2.552 (0.296)	0.169 (0.006)	0.152 (0.005)

TABLE 6
Performance comparison of model selection for Example 3.

Methods	FPR		FNR	
	(I)	(II)	(I)	(II)
GLM-O	-	-	-	-
LASSO	0.197 (0.030)	0.344 (0.017)	0.654 (0.052)	0.254 (0.039)
Ridge	0.978 (0.004)	0.976 (0.005)	0.032 (0.012)	0.016 (0.009)
Enet	0.175 (0.034)	0.336 (0.036)	0.818 (0.031)	0.472 (0.058)
sGLMg	0.407 (0.024)	0.409 (0.025)	0.005 (0.004)	0.001 (0.001)
sGLMg-O	0.246 (0.035)	0.302 (0.034)	0.000 (0.000)	0.000 (0.000)
GDSGLM	0.291 (0.027)	0.320 (0.034)	0.082 (0.021)	0.048 (0.014)
GDSGLM-O	0.290 (0.024)	0.224 (0.025)	0.009 (0.006)	0.004 (0.004)

TABLE 7
Performance comparison of estimation and prediction for Example 4.

Methods	L_2 distance		Error	
	(I)	(II)	(I)	(II)
GLM-O	3.007 (0.096)	2.595 (0.097)	0.129 (0.003)	0.115 (0.002)
LASSO	2.998 (0.046)	2.661 (0.033)	0.171 (0.004)	0.139 (0.003)
Ridge	3.597 (0.003)	3.586 (0.002)	0.158 (0.003)	0.158 (0.002)
Enet	2.898 (0.036)	2.540 (0.032)	0.169 (0.004)	0.137 (0.003)
sGLMg	2.452 (0.032)	2.150 (0.039)	0.154 (0.004)	0.123 (0.003)
sGLMg-O	2.422 (0.034)	2.153 (0.039)	0.154 (0.004)	0.123 (0.003)
GDSGLM	2.191 (0.038)	1.995 (0.036)	0.127 (0.004)	0.114 (0.003)
GDSGLM-O	2.188 (0.039)	1.990 (0.037)	0.125 (0.003)	0.114 (0.003)

6. Real data example

Human microbiome has received great interest in medical research. The microbiome composition has been found to link to many aspects of human health. In order to understand why only a subset of patients benefit from the immunotherapy in cancer treatment, Matson et al. [22] conducted studies to find whether or not some microbiome species are predictors which can be used to classify metastatic melanoma patients' response to the immunotherapy. The microbiome sequencing data (16s sequencing) includes 38 cancer patients with 153 operational taxonomic units (OTUs). As the dataset is zero-inflated, we first filter out the variables with more than 50 percent zero samples and combine the same species. After pre-processing, we obtain 33 microbiome OTUs as the predictors. The goal of the study is to use the commensal microbial composition

TABLE 8
Performance comparison of model selection for Example 4.

Methods	FPR		FNR	
	(I)	(II)	(I)	(II)
GLM-O	-	-	-	-
LASSO	0.161 (0.007)	0.185 (0.010)	0.315 (0.012)	0.200 (0.012)
Ridge	1.000 (0.000)	1.000 (0.002)	0.000 (0.000)	0.000 (0.000)
Enet	0.202 (0.009)	0.233 (0.009)	0.225 (0.013)	0.137 (0.011)
sGLMg	0.212 (0.010)	0.177 (0.010)	0.004 (0.002)	0.000 (0.000)
sGLMg-O	0.215 (0.009)	0.173 (0.010)	0.000 (0.000)	0.000 (0.000)
GDSGLM	0.029 (0.005)	0.023 (0.005)	0.001 (0.001)	0.000 (0.000)
GDSGLM-O	0.024 (0.005)	0.016 (0.004)	0.001 (0.001)	0.000 (0.000)

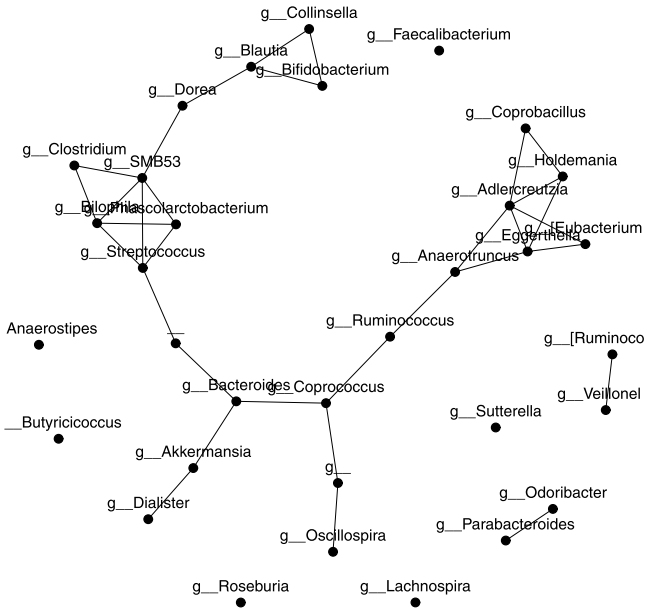

FIG 2. *Estimated graph of* 33 *OTUs features.*

to predict the clinical response regarding whether or not the patients will benefit from the immunotherapy. A doubly sparse logistic model incorporating the graphical structure (2.4) is used to analyze the dataset.

We compare the performance of the GDSGLM method with sGLMg [51], Lasso, ridge regression, Adaptive Lasso and Elastic net. We normalize the data set and split the data sets equally into training data and test data. The cross validation (CV) is used to compare different approaches. We use the graphical Lasso [9] to estimate the predictor (OTUs) graph G only based on the training data. Figure 2 shows the estimated graph structure of OTUs feature based on all the data. The training set is used to build the models and the testing set is used to measure all the error measures including the percentage of correct classifications (PCC), the specificity and sensitivity. An inner five-fold CV procedure is

TABLE 9
Comparison for various methods on the microbiome sequencing dataset.

Methods	Specificity	Sensitivity	PCC
LASSO	0.940 (0.122)	0.037 (0.110)	0.606 (0.112)
ALasso	0.881 (0.180)	0.053 (0.120)	0.576 (0.140)
Ridge	0.973 (0.066)	0.033 (0.095)	0.626 (0.075)
Enet	0.956 (0.093)	0.023 (0.090)	0.605 (0.103)
sGLMg	0.947 (0.083)	0.611 (0.326)	0.828 (0.120)
GDSGLM	0.924 (0.089)	0.775 (0.192)	0.913 (0.087)

used to select the tuning parameters [15]. We conduct the equal-splitting cross validation process 50 times. Table 9 shows the average misclassification error of all methods. The GDSGLM method provides the best classification with the highest PCC and a much higher sensitivity but a relatively lower specificity.

Based on 50 times of two-fold cross validation, we obtain 100 models for each method. The GDSGLM method selects about 16 OTUs on average. There are five OTUs selected with more than 75 times by the GDSGLM method. In details, the feature names are *g_Adlercreutzia* [33], *g_Collinsella* [38], *g_SMB53* [13], *g_Odoribacter* [31], and *g_Sutterella* [3]. All these five bacteria species have been shown to influence patients response to immunotherapy in literature. Further biological experiments are required to check specifically whether these OTUs are closely related to anti-PD-1 efficacy in patients with metastatic melanoma.

7. Discussion

In this paper, we investigate the doubly sparse penalized estimator for high dimensional GLMs using the graphical structure of the predictors. We establish the tight finite sample bounds for both estimation and prediction. We also establish the model selection consistency under the ultra-high dimensional setting. Relevant directions for future work include a generalization of the statistical consistency results to nonconvex regularized M-estimators. Specifically, it would be interesting to expand the theory to the minimization of nonsmooth hinge loss function for classification.

In this paper, Assumption 1 assumes that the neighbors of true predictors are true predictors as well. This assumption is somehow restrictive and it excludes some common graphs, for example, the chain graphs. A sensitivity study is discussed in [46], and it is shown that if Assumption 1 is not violated seriously, the proposed method still has good performance. Here, we propose some future directions on how to relax this assumption. Let $|\mathcal{N}_j \cap \mathcal{N}_k| = p^*$ for $j \in \mathcal{J}_0$ and $k \in \mathcal{J}_0^c$. For example, we have $p^* \leq 2$ for the chain structure. The proof of Theorem 4.1 needs to be modified by adding a small term $\delta(p^*)$ in the following equation:

$$\sum_{j\in\mathcal{J}_0^c} (\tau_j(\|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c}\|_2 + \|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_2) + \xi(\|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c}\|_1 + \|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_1)$$
$$\leq 3\sum_{j\in\mathcal{J}_0} (\tau_j(\|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c}\|_2 + \|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_2) + \xi(\|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c}\|_1 + \|T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_1) + 4\delta(p^*),$$

where

$$\delta(p^*) = \| \sum_{j\in\mathcal{J}_0} T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c} + \sum_{j\in\mathcal{J}_0^c} T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_{G,\tau} + \| \sum_{j\in\mathcal{J}_0} T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c} + \sum_{j\in\mathcal{J}_0^c} T^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_{S,\xi}$$
$$+ \| \sum_{j\in\mathcal{J}_0} S^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c} + \sum_{j\in\mathcal{J}_0^c} S^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_{G,\tau} + \| \sum_{j\in\mathcal{J}_0} S^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0^c} + \sum_{j\in\mathcal{J}_0^c} S^{(j)}_{\mathcal{N}_j\cap\mathcal{J}_0}\|_{S,\xi}.$$

To extend the results in Theorem 4.1 to more general graphs, one has to control the additional term $\delta(p^*)$, which warrants future investigation.

Appendix A: Proofs of main theorems

In this section, we provide the proofs of the main theorems stated in the paper. The proof of Theorem 4.1 begins with the proofs of some technical lemmas. The following Lemma is quoted from the Lemma 2 in [46].

Lemma A.1. *For any predictor graph G and positive weights $\tau_1, \tau_2, \ldots, \tau_p$ and ξ, suppose $V^{(1)}, V^{(2)}, \ldots, V^{(p)}$ is an optimal decomposition of $\boldsymbol{\beta} \in \mathbb{R}^p$, then for any $S \subset \{1, 2, \ldots, p\}$, $\{V^{(j)} : j \in S\}$ is also an optimal decomposition of $\sum_{j\in S} V^{(j)}$.*

A.1. Proof of Theorem 4.1

Proof. We first show that $\mathcal{R}(\boldsymbol{\beta})$ is a norm and decomposable with respect to a pair of subspaces $(\mathcal{M}, \mathcal{M}^{\perp})$. By Lemma 3.2 in [36], we note that $\mathcal{R}(\boldsymbol{\beta}) \geq 0$ with equality only when $\boldsymbol{\beta} = 0$. Then for $u \in \mathbb{R}\backslash\{0\}$, we have the positive homogeneity, i.e., $\mathcal{R}(u\boldsymbol{\beta}) = |u|\mathcal{R}(\boldsymbol{\beta})$. To demonstrate the triangle inequality, let the set of vectors $V^{(j)}$, $j = 1, \ldots, p$, be a decomposition of the regression parameter vector $\boldsymbol{\beta}$ and let $W^{(j)}$, $j = 1, \ldots, p$, be a decomposition of another regression parameter vector $\boldsymbol{\theta}$. Then,

$$\mathcal{R}(\boldsymbol{\beta} + \boldsymbol{\theta}) = \sum_{j=1}^{p} [\tau_j \|V^{(j)} + W^{(j)}\|_2 + \xi\|V^{(j)} + W^{(j)}\|_1]$$
$$\leq \mathcal{R}(\boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\theta}).$$

Therefore, $\mathcal{R}(\boldsymbol{\beta})$ is a norm. If we have vectors $\boldsymbol{\beta} \in \mathcal{M}$ and $\boldsymbol{\theta} \in \mathcal{M}^{\perp}$, then $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ will have supports that do not overlap. By Assumption 1, it follows that any optimal decomposition of $\boldsymbol{\beta}$, say $V^{(j)}, j = 1, \ldots, p$, and any optimal decomposition of $\boldsymbol{\theta}$, say $W^{(j)}, j = 1, \ldots, p$, will also have supports that do not overlap. As we have mutually exclusive sets, the triangle inequality will hold with equality and $\mathcal{R}(\boldsymbol{\beta})$ is decomposable with respect to the subspaces $\mathcal{M}$ and $\mathcal{M}^{\perp}$.

Next, we show that the event $\{\lambda \geq 4\mathcal{R}^*(\nabla\mathcal{L}_n(\boldsymbol{\beta}))\}$ holds with a high probability. Define $\mathbf{u}$ to be a $p \times 1$ vector and let $\mathbf{u}_{N_j}$ be constrained to have support

matching $V^{(j)}$, i.e., $\mathbf{u}_{\mathcal{N}_j}$ has non-zero elements $k \in \mathcal{N}_j$. By the proof of Lemma 3 in [27], its dual norm satisfies

$$\mathcal{R}^*(\mathbf{u}) \leq \max_{j=1,\ldots,p} \frac{1}{\xi + \tau_{\min}} \|\mathbf{u}_{\mathcal{N}_j}\|_2 \leq \frac{1}{\xi + \tau_{\min}} \max_{j=1,\ldots,p} \sqrt{d_j} \|\mathbf{u}_{\mathcal{N}_j}\|_\infty = \frac{\sqrt{d_{\max}}}{\xi + \tau_{\min}} \|\mathbf{u}\|_\infty.$$

In the following, we show that there are universal constants (c, c_1, c_2) such that

$$P\left(\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty \geq c\sqrt{\frac{\log(p)}{n}} \right) \leq c_1 \exp(-c_2 \log(p)).$$

For each $1 \leq i \leq n$ and $1 \leq j \leq p$, define the random variable $\mathcal{Z}_{ij} := (b'(X_i^{\mathrm{T}}\boldsymbol{\beta}^0) - y_i)X_{ij}$. Our goal is to bound $\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{ij}|$. Note that

$$P\left(\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{ij}| \geq \delta \right) \leq P(\mathcal{A}^c) + P\left(\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{ij}| \geq \delta | \mathcal{A} \right), \quad \text{(A.1)}$$

where

$$\mathcal{A} = \left\{ \max_{j=1,\ldots,p} \left\{ \frac{1}{n}\sum_{i=1}^n X_{ij}^2 \right\} \leq 2\mathbb{E}[X_{ij}^2] \right\}$$

and $\mathbb{E}[X_{ij}^2] \leq \kappa_u^2$. As $n \gtrsim \log p$, $\{X_{ij}, j = 1, \ldots, p\}$ are sub-Gaussian, and the squared X_{ij}s are sub-exponential, there exist universal constants (c_1, c_2) such that $P(\mathcal{A}^c) \leq c_1 \exp(-c_2 n)$. We focus on the second term on the right side of (A.1). For any $t \in \mathbb{R}$, we have

$$\begin{aligned} \log \mathbb{E}[\exp(t\mathcal{Z}_{ij})|X_{ij}] &= \log(\exp(tX_{ij}b'(X_i^{\mathrm{T}}\boldsymbol{\beta}^0))) \cdot \mathbb{E}[\exp(-tX_{ij}y_i)] \\ &= tX_{ij}b'(X_i^{\mathrm{T}}\boldsymbol{\beta}^0) + (b(-tX_{ij} + X_i^{\mathrm{T}}\boldsymbol{\beta}^0) - b(X_i^{\mathrm{T}}\boldsymbol{\beta}^0)), \end{aligned}$$

where $b(\cdot)$ is the cumulant generating function for the underlying exponential family. Thus, by a Taylor series expansion, there exists some $\nu_i \in [0, 1]$ such that

$$\log E[\exp(t\mathcal{Z}_{ij})|X_{ij}] = \frac{t^2 X_{ij}^2}{2} b''(X_i^{\mathrm{T}}\boldsymbol{\beta}^0 - \nu_i t X_{ij}) \leq \frac{Ct^2 X_{ij}^2}{2}, \quad \text{(A.2)}$$

where the inequality is based the boundedness of $b''(\cdot)$. Consequently, conditioned on the event $\mathcal{A}$, the variable $\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{ij}$ is a sub-Gaussian random variable with the parameter less than or equal to $\kappa = C \max_{j=1,\ldots,p} \mathbb{E}[X_{ij}^2]$ for each $j = 1, \ldots, p$. By a union bound, we have

$$P\left(\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{ij}| \geq \delta | \mathcal{A} \right) \leq p \exp\left(-\frac{n\delta^2}{2\kappa^2} \right).$$

Then $\nabla \mathcal{L}_n(\boldsymbol{\beta}^0) \in \mathbb{R}^p$ is zero-mean sub-Gaussian random vector. By the assumption $\lambda(\tau_{\min} + \xi) \geq c\sqrt{d_{\max} \log p / n}$ for some constant c, therefore, we

have $\lambda \geq 4\mathcal{R}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta}))$.

Next, there is an arbitrary optimal decomposition of $\boldsymbol{\beta}^0$ as $S^{(1)}, S^{(2)}, \ldots, S^{(p)}$, and an arbitrary optimal decomposition of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ as $T^{(1)}, T^{(2)}, \ldots, T^{(p)}$. Then, by Assumption 1,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_{G,\tau} + \|\boldsymbol{\beta}^0\|_{G,\tau} - \|\widehat{\boldsymbol{\beta}}\|_{G,\tau} \\
&\qquad = \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0^c} T^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0} S^{(j)}\|_{G,\tau} - \|\widehat{\boldsymbol{\beta}}\|_{G,\tau}, \\
&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_{S,\xi} + \|\boldsymbol{\beta}^0\|_{S,\xi} - \|\widehat{\boldsymbol{\beta}}\|_{S,\xi} \\
&\qquad = \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi} + \|\sum_{j\in\mathcal{J}_0^c} T^{(j)}\|_{S,\xi} + \|\sum_{j\in\mathcal{J}_0} S^{(j)}\|_{S,\xi} - \|\widehat{\boldsymbol{\beta}}\|_{S,\xi}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}\|_{G,\tau} &= \|\sum_{j\in\mathcal{J}_0} T^{(j)} + \sum_{j\in\mathcal{J}_0^c} T^{(j)} + \sum_{j\in\mathcal{J}_0} S^{(j)}\|_{G,\tau}, \\
&\geq \|\sum_{j\in\mathcal{J}_0} S^{(j)} + \sum_{j\in\mathcal{J}_0^c} T^{(j)}\|_{G,\tau} - \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau}, \\
&= \|\sum_{j\in\mathcal{J}_0} S^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0^c} T^{(j)}\|_{G,\tau} - \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau},
\end{aligned}$$

and similarly,

$$\|\widehat{\boldsymbol{\beta}}\|_{S,\xi} \geq \|\sum_{j\in\mathcal{J}_0} S^{(j)}\|_{S,\xi} + \|\sum_{j\in\mathcal{J}_0^c} T^{(j)}\|_{S,\xi} - \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi}.$$

Hence,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_{G,\tau} + \|\boldsymbol{\beta}^0\|_{G,\tau} - \|\widehat{\boldsymbol{\beta}}\|_{G,\tau} &\leq 2\|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau}, \\
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_{S,\xi} + \|\boldsymbol{\beta}^0\|_{S,\xi} - \|\widehat{\boldsymbol{\beta}}\|_{S,\xi} &\leq 2\|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi},
\end{aligned}$$

and

$$\mathcal{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \mathcal{R}(\boldsymbol{\beta}^0) - \mathcal{R}(\widehat{\boldsymbol{\beta}}) \leq 2\Big(\|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi}\Big). \tag{A.3}$$

Note that $\mathcal{J}_0 = \{j : \beta_j^0 \neq 0\}$ be the true support of $\boldsymbol{\beta}^0$, and the local optimal error vector $\widehat{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ where $\widehat{\boldsymbol{\beta}}$ is an arbitrary local optimum of (3.1). Let

$$\mathcal{F}(\widehat{\Delta}) := \mathcal{L}_n(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^0) + \lambda\{\mathcal{R}(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{R}(\boldsymbol{\beta}^0)\},$$

then the optimal error $\widehat{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ must satisfy $\mathcal{F}(\widehat{\Delta}) \leq 0$. As the distribution belongs to exponential family, $\mathcal{L}_n(\boldsymbol{\beta})$ is a convex and differentiable loss function.

Using the mean value theorem, we have

$$\begin{aligned}\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\Delta} \rangle &= \frac{1}{n}\sum_{i=1}^{n}(b'(\langle X_i, \boldsymbol{\beta}^0 + \widehat{\Delta}\rangle) - b'(\langle X_i, \boldsymbol{\beta}^0\rangle))X_i^{\mathrm{T}}\widetilde{\Delta} \\ &= \frac{1}{n}\sum_{i=1}^{n} b''(\langle X_i, \boldsymbol{\beta}^0\rangle + t_i\langle X_i, \widehat{\Delta}\rangle)(\langle X_i, \widehat{\Delta}\rangle)^2,\end{aligned}$$

where $t_i \in [0,1]$. From the proof of Proposition 2 in [26] and Theorem 9.36 in [43], there exist positive constants κ_1 and κ_2 such that, for all $\Delta \in \mathbb{R}^p$ with $\|\Delta\| \le 1$,

$$\mathcal{L}_n(\boldsymbol{\beta}^0 + \Delta) - \mathcal{L}_n(\boldsymbol{\beta}^0) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0)^{\mathrm{T}}\Delta \ge \kappa_1\|\Delta\|_2^2 - \kappa_2\frac{\log p}{n}\mathcal{R}^2(\Delta) \tag{A.4}$$

with probability at least $1 - c_1\exp(-c_2 n)$ for some $c_1, c_2 > 0$. Then

$$\begin{aligned}\mathcal{F}(\widehat{\Delta}) &= \mathcal{L}_n(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^0) + \lambda\{\mathcal{R}(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{R}(\boldsymbol{\beta}^0)\} \\ &\ge \langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\Delta}\rangle + \kappa_1\|\widehat{\Delta}\|_2^2 - \kappa_2\frac{\log p}{n}\mathcal{R}^2(\widehat{\Delta}) + \lambda\{\mathcal{R}(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{R}(\boldsymbol{\beta}^0)\} \\ &\ge \langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\Delta}\rangle + \kappa_1\|\widehat{\Delta}\|_2^2 - \kappa_2\frac{\log p}{n}\rho\mathcal{R}(\widehat{\Delta}) + \lambda\{\mathcal{R}(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{R}(\boldsymbol{\beta}^0)\}\end{aligned}$$

Applyig the Cauchy-Schwarz inequality to the regularizer $\mathcal{R}$ and its dual $\mathcal{R}^*$ gives the result that $|\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\Delta}\rangle| \le \mathcal{R}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta}^0))\mathcal{R}(\widehat{\Delta})$. According to the assumption $\rho \le c'\frac{\sqrt{d_{max}}}{\tau_{\min}+\xi}\sqrt{\frac{n}{\log p}}$ for some positive c', and $\lambda \ge 4\mathcal{R}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta}))$ yields $|\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^0), \widehat{\Delta}\rangle| \le (\lambda/4)\mathcal{R}(\widehat{\Delta})$, we have

$$0 \ge \mathcal{F}(\widehat{\Delta}) \ge \kappa_1\|\widehat{\Delta}\|_2^2 - \frac{\lambda}{2}\mathcal{R}(\widehat{\Delta}) + \lambda\{\mathcal{R}(\boldsymbol{\beta}^0 + \widehat{\Delta}) - \mathcal{R}(\boldsymbol{\beta}^0)\}, \tag{A.5}$$

and

$$\mathcal{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + 2(\mathcal{R}(\boldsymbol{\beta}^0) - \mathcal{R}(\widehat{\boldsymbol{\beta}})) \ge 0.$$

Then, we have

$$\begin{aligned}\mathcal{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) &\le 2\{\mathcal{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \mathcal{R}(\boldsymbol{\beta}^0) - \mathcal{R}(\widehat{\boldsymbol{\beta}})\} \\ &\le 4\Big(\|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi}\Big).\end{aligned}$$

By Lemma A.1, we have

$$\|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi} = \sum_{j\in\mathcal{J}_0}\tau_j\|T^{(j)}\|_2 + \sum_{j\in\mathcal{J}_0}\xi\|T^{(j)}\|_1.$$

Then, by definition,

$$\sum_{j\in\mathcal{J}_0^c}(\tau_j\|T^{(j)}\|_2 + \xi\|T^{(j)}\|_1) \le 3\sum_{j\in\mathcal{J}_0}(\tau_j\|T^{(j)}\|_2 + \xi\|T^{(j)}\|_1).$$

Note that the optimal decomposition of Δ must be a decomposition minimizing the norm $\mathcal{R}(\cdot)$. For the vector $\Delta \in \mathbb{R}^p$, there exist a decomposition $U^{(j)}$'s such that the supports of $U^{(j)}$ do not overlap. Then, for any subset $J \subset \{1, 2, \ldots, p\}$ with $|J| \le s_0$, and all the optimal decompositions $(V^{(1)}, \ldots, V^{(p)})$ of any vector Δ, we have

$$\begin{aligned}
\left(\sum_{j\in J}(\tau_j + \sqrt{d_j}\xi)^2\|V^{(j)}\|_2^2\right)^{1/2} &\le \left(\sum_{j=1}^{p}(\tau_j + \sqrt{d_j}\xi)^2\|V^{(j)}\|_2^2\right)^{1/2} \\
&\le \sum_{j=1}^{p}(\tau_j + \sqrt{d_j}\xi)\|V^{(j)}\|_2 \\
&\le \sum_{j=1}^{p}\tau_j\|V^{(j)}\|_2 + \sqrt{d_j}\xi\|V^{(j)}\|_1 \\
&\le \sum_{j=1}^{p}\tau_j\|U^{(j)}\|_2 + \sqrt{d_j}\xi\|U^{(j)}\|_1 \\
&\le \sum_{j=1}^{p}(\tau_j + d_j\xi)\|U^{(j)}\|_2 \\
&\le \sqrt{\mathcal{K}}(\tau_{\max} + d_{\max}\xi)\left(\sum_{j=1}^{p}\|U^{(j)}\|_2\right)^{1/2} \\
&= \sqrt{\mathcal{K}}(\tau_{\max} + d_{\max}\xi)\|\Delta\|_2.
\end{aligned}$$

From (A.4), we have with probability at least $1 - c_1\exp(-c_2 n)$ for some $c_1, c_2 > 0$,

$$\begin{aligned}
&\mathcal{L}_n(\boldsymbol{\beta}^0 + \Delta) - \mathcal{L}_n(\boldsymbol{\beta}^0) - \nabla\mathcal{L}_n(\boldsymbol{\beta}^0)^T\Delta \\
&\qquad \ge \kappa_l\sum_{j\in J}(\tau_j + \sqrt{d_j}\xi)^2\|V^{(j)}\|_2^2 - \kappa_2\frac{\log p}{n}\mathcal{R}^2(\Delta), \quad \text{(A.6)}
\end{aligned}$$

for $\kappa_l = \frac{\kappa_1}{\mathcal{K}(\tau_{\max}+d_{\max}\xi)^2}$ and $\kappa_2 \ge 0$.

From (A.5) and (A.6),

$$\mathcal{F}(\widehat{\Delta}) \ge \kappa_l\sum_{j\in\mathcal{J}_0}(\tau_j + \xi)^2\|T^{(j)}\|_2^2 - \frac{\lambda}{2}\mathcal{R}(\widehat{\Delta}) + \lambda\{\mathcal{R}(\widehat{\boldsymbol{\beta}}) - \mathcal{R}(\boldsymbol{\beta}^0)\}.$$

By $\mathcal{F}(\widehat{\Delta}) \le 0$, and adding $\lambda\mathcal{R}(\widehat{\Delta})$ to both sides of the resulting inequality, we have

$$\begin{aligned}
&\lambda\mathcal{R}(\widehat{\Delta}) + 2\kappa_l\sum_{j\in\mathcal{J}_0}(\tau_j + \sqrt{d_j}\xi)^2\|T^{(j)}\|_2^2 \\
&\qquad \le 2\lambda(\mathcal{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \mathcal{R}(\boldsymbol{\beta}^0) - \mathcal{R}(\widehat{\boldsymbol{\beta}}))
\end{aligned}$$

$$
\begin{aligned}
&\leq 4\lambda\Big(\|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\in\mathcal{J}_0} T^{(j)}\|_{S,\xi}\Big)\\
&= 4\lambda\Big(\sum_{j\in\mathcal{J}_0}\tau_j\|T^{(j)}\|_2 + \sum_{j\in\mathcal{J}_0}\xi\|T^{(j)}\|_1\Big)\\
&\leq 4\lambda\sum_{j\in\mathcal{J}_0}(\tau_j+\sqrt{d_j}\xi)\|T^{(j)}\|_2\\
&\leq 4\lambda\mathcal{K}^{1/2}\sqrt{\sum_{j\in\mathcal{J}_0}(\tau_j+\sqrt{d_j}\xi)^2\left\|T^{(j)}\right\|_2^2},
\end{aligned}
$$

where the last inequality follows from for each $j\in\mathcal{J}_0$, there is at most $\mathcal{K}$ nonzero $T^{(j)}$. Note that $2xy\leq tx^2+y^2/t$ for all $t>0$. Then, we have

$$
\begin{aligned}
&\lambda\mathcal{R}(\widehat{\Delta}) + 2\kappa_l\sum_{j\in\mathcal{J}_0}(\tau_j+\sqrt{d_j}\xi)^2\|T^{(j)}\|_2^2\\
&\qquad\leq 4\lambda\mathcal{K}^{1/2}\sqrt{\sum_{j\in\mathcal{J}_0}(\tau_j+\sqrt{d_j}\xi)^2\left\|T^{(j)}\right\|_2^2}\\
&\qquad\leq 4t\lambda^2\mathcal{K} + \frac{1}{t}\sum_{j\in\mathcal{J}_0}(\tau_j+\sqrt{d_j}\xi)^2\left\|T^{(j)}\right\|_2^2.
\end{aligned}
$$

By choosing $t=\frac{1}{2\kappa_{l'}}$, we obtain

$$
\mathcal{R}(\widehat{\Delta})\leq\frac{2\lambda\mathcal{K}}{\kappa_l}.
$$

Besides, we have

$$
\begin{aligned}
2\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2 &= 2\|\sum_{j=1}^{p}T^{(j)}\|_2\\
&\leq \|\sum_{j=1}^{p}T^{(j)}\|_2 + \|\sum_{j=1}^{p}T^{(j)}\|_1\\
&\leq \|\sum_{j=1}^{p}\tau_j T^{(j)}\frac{1}{\tau_j}\|_2 + \|\sum_{j=1}^{p}\xi T^{(j)}\frac{1}{\xi}\|_1\\
&\leq \frac{\mathcal{R}(\widehat{\Delta})}{\tau_{\min}\wedge\xi}\leq\frac{2\lambda\mathcal{K}}{\kappa_l(\tau_{\min}\wedge\xi)}
\end{aligned}
$$

Then,

$$
\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2\leq\frac{\lambda\mathcal{K}}{\kappa_l(\tau_{\min}\wedge\xi)}.\square
$$

A.2. Proof of Corollary 4.1

Proof. According to the first order stationary condition in [21],

$$\langle \nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \partial \lambda \mathcal{R}(\widehat{\boldsymbol{\beta}}),\, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle \geq 0, \quad \text{for all feasible } \boldsymbol{\beta} \in \mathbb{R}^p.$$

Let $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ in the formula above, then

$$\begin{aligned}\langle \nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0),\, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \rangle &\leq \langle -\partial \lambda \mathcal{R}(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0),\, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \rangle \\ &\leq \lambda(\mathcal{R}(\boldsymbol{\beta}^0) - \mathcal{R}(\widehat{\boldsymbol{\beta}})) + \mathcal{R}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta}^0)) \mathcal{R}(\widehat{\Delta}).\end{aligned}$$

Note that $\mathcal{J}_0 = \{j : \beta_j^0 \neq 0\}$ be the true support of $\boldsymbol{\beta}^0$. By the condition that $\lambda \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta}^0))$, we obtain

$$\begin{aligned}\langle \nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^0),\, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \rangle &\leq \lambda(\mathcal{R}(\widehat{\Delta}_{\mathcal{J}_0}) - \mathcal{R}(\widehat{\Delta}_{\mathcal{J}_0^c})) + \frac{\lambda}{2}\mathcal{R}(\widehat{\Delta}) \\ &\leq \frac{3\lambda}{2}\mathcal{R}(\widehat{\Delta}_{\mathcal{J}_0}) - \frac{\lambda}{2}\mathcal{R}(\widehat{\Delta}_{\mathcal{J}_0^c}) \\ &\leq \frac{3\lambda}{2}\mathcal{R}(\widehat{\Delta}).\end{aligned}$$

Then, substituting the ℓ-2 bound with the result from Theorem 4.1 yields the desired result. □

First we provide a lemma of the bound of covariance matrices from sub-Gaussian ensembles quoted from [43].

Lemma A.2. *There are universal constants* $\{c_j\}_{j=1}^3$ *such that, for any row-wise sub-Gaussian random matrix* $\mathbf{X} \in \mathbb{R}^{n \times p}$*, the sample covariance* $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^{\mathrm{T}}$ *satisfies the bounds*

$$P\left(\frac{\|\widehat{\Sigma} - \Sigma\|_2}{\sigma_x^2} > c_1\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right) + \delta\right) \leq c_2 \exp(-c_3 n \min\{\delta, \delta^2\})$$

for all $\delta > 0$.

Proof. The proof can be obtained via a discretization argument in [43]. □

A.3. Proof of Theorem 4.2

Proof. By Proposition 4.1, it is known that $\widehat{\boldsymbol{\beta}}$ is a solution if and only if the estimator $\widehat{\boldsymbol{\beta}}$ can be decomposed as $\widehat{\boldsymbol{\beta}} = \sum_{j=1}^p V^{(j)}$, where $V^{(j)}$'s satisfy the following conditions for all $1 \leq j \leq p$: (i) $V^{(j)}_{\mathcal{N}_j^c} = 0$; (ii) $V^{(j)}_{\mathcal{N}_j} = 0$ and $\|\nabla_{\mathcal{N}_j}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda h_{\mathcal{N}_j}\|_2 \leq \lambda\tau_j$ with $|h_{\mathcal{N}_j}| \leq \xi$; (iii) $V^{(j)}_{\mathcal{N}_j} \neq 0$ and $\nabla_{\mathcal{N}_j}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda\tau_j V^{(j)}_{\mathcal{N}_j}/\|V^{(j)}_{\mathcal{N}_j}\|_2 + \lambda h_{\mathcal{N}_j} = 0$, where for $k \in \mathcal{N}_j$ either $V_k^{(j)} = 0$ and $|\nabla_k \mathcal{L}_n(\widehat{\boldsymbol{\beta}})| \leq \lambda\xi$, or $V_k^{(j)} \neq 0$ and $h_k = \xi\,\mathrm{sign}(V_k^{(j)})$. Throughout the proof, denote $\nabla_{\mathcal{N}_j}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) = X^{\mathrm{T}}_{\mathcal{N}_j}(b'(\mathbf{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{y})$.

Denote $\hat{H} = \{j : \|V^{(j)}_{\mathcal{N}_j}\|_2 \neq 0\}$. Then, we have $\nabla_{\mathcal{N}_j}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda(\tau_j V^{(j)}_{\mathcal{N}_j}/\|V^{(j)}_{\mathcal{N}_j}\|_2 + \xi W^{(j)}_{\mathcal{N}_j}) = 0$ for each $j \in \hat{H}$ and $\nabla_{\mathcal{N}_j}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda(\tau_j Z^{(j)}_{\mathcal{N}_j} + \xi W^{(j)}_{\mathcal{N}_j}) = 0$ for each $j \notin \hat{H}$, where $Z^{(j)}$ and $W^{(j)}$ are $p \times 1$ random vectors with $\|Z^{(j)}_{\mathcal{N}_j}\|_2 \leq 1$ and $|W^{(j)}_k| \leq 1$, $k \in \mathcal{N}_j$. As some predictors may reside in more than one neighborhoods, following [46], the estimate needs to satisfy the conditions:

(i) $\tau_{i_1} V^{(i_1)}_j/\|V^{(i_1)}_{\mathcal{N}_{i_1}}\|_2 + \xi W^{(i_1)}_j = \tau_{i_2} V^{(i_2)}_j/\|V^{(i_2)}_{\mathcal{N}_{i_2}}\|_2 + \xi W^{(i_2)}_j$ for each $i_1, i_2 \in \hat{H}$ and $j \in \mathcal{N}_{i_1} \cap \mathcal{N}_{i_2}$;

(ii) $\tau_{i_1} V^{(i_1)}_j/\|V^{(i_1)}_{\mathcal{N}_{i_1}}\|_2 + \xi W^{(i_1)}_j = \tau_{i_2} Z^{(i_2)}_j$ for each $i_1 \in \hat{H}$, $i_2 \notin \hat{H}$ and $j \in \mathcal{N}_{i_1} \cap \mathcal{N}_{i_2}$;

(iii) $\tau_{i_1} Z^{(i_1)}_j = \tau_{i_2} Z^{(i_2)}_j$ for each for each $i_1, i_2 \notin \hat{H}$ and $j \in \mathcal{N}_{i_1} \cap \mathcal{N}_{i_2}$.

Then, any solution $\widehat{\boldsymbol{\beta}}$ satisfies the following equation

$$\nabla\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda(\hat{f} + \hat{h}) = 0, \tag{A.7}$$

where for each $1 \leq j \leq p$, $\hat{f}_j = \tau_j V^{(j)}_j/\|V^{(j)}_{\mathcal{N}_j}\|_2$ if $j \in \hat{H}$ and $\hat{f}_j = \tau_j Z^{(j)}_j$ if $j \notin \hat{H}$, and $\hat{h}_j = \xi\,\mathrm{sign}(V^{(j)}_j)$, if $V^{(j)}_j \neq 0$ and $\hat{h}_j = \xi W^{(j)}_j$ with $|W^{(j)}_j| \leq 1$, if $V^{(j)}_j = 0$.

Define events

$$\Omega_1 = \{\|\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}\|_\infty < \beta^0_{\min}\},$$
$$\Omega_2 = \{\|\hat{f}_{\mathcal{N}_j}\|_2 < \tau_j, j \in \mathcal{J}_0^c\}.$$

When event Ω_1 occurs, we have $\mathrm{sign}(\hat{\beta}_j) = \mathrm{sign}(\beta^0_j)$ for each $j \in \mathcal{J}_0$. If event Ω_2 occurs, we obtain $V^{(j)}_{\mathcal{N}_j} = 0$ for each $j \in \mathcal{J}_0^c$. Furthermore, we know that $V^{(j)}_{\mathcal{N}_j^c} = 0$ for each j. Then, by Assumption 1, $\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0^c} = \sum_{j\in\mathcal{J}_0^c} V^{(j)}_{\mathcal{J}_0^c} = 0$. Thus, it suffices to show that $P(\Omega_1 \cap \Omega_2) \to 1$, which implies $P(\mathrm{sign}(\widehat{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta}^0)) \to 1$ as $n \to \infty$.

Note that if events Ω_1 and Ω_2 occur, from equation (A.7), we have

$$\nabla_{\mathcal{J}_0}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda(\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}) = 0,$$
$$\nabla_{\mathcal{N}_j}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda(\hat{f}_{\mathcal{N}_j} + \hat{h}_{\mathcal{N}_j}) = 0, \qquad j \in \mathcal{J}_0^c.$$

Thus, by a Taylor expansion and Assumption 1, we have

$$\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0) + \nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}_{\mathcal{J}_0}) + \lambda(\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}) = 0, \tag{A.8}$$
$$\nabla_{\mathcal{N}_j}\mathcal{L}_n(\boldsymbol{\beta}^0) + \nabla^2_{\mathcal{N}_j\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}_{\mathcal{J}_0}) + \lambda(\hat{f}_{\mathcal{N}_j} + \hat{h}_{\mathcal{N}_j}) = 0, \qquad j \in \mathcal{J}_0^c, \tag{A.9}$$

where $\bar{\boldsymbol{\beta}}_{\mathcal{J}_0}$ lies on the line segment joining $\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0}$ and $\boldsymbol{\beta}^0_{\mathcal{J}_0}$. Then, by Assumption 4

$$\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}_{\mathcal{J}_0} = -(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1}[\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0) + \lambda(\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0})], \tag{A.10}$$

$$\begin{aligned}\hat{f}_{\mathcal{N}_j} &= -\frac{1}{\lambda}\Big(\nabla_{\mathcal{N}_j}\mathcal{L}_n(\boldsymbol{\beta}^0) - \nabla^2_{\mathcal{N}_j\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}})(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1}\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\Big)\\ &\quad + \Big(\nabla^2_{\mathcal{N}_j\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}})(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1}\hat{h}_{\mathcal{J}_0} - \hat{h}_{\mathcal{N}_j}\Big)\\ &\quad + \nabla^2_{\mathcal{N}_j\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}})(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1}\hat{f}_{\mathcal{J}_0}. \qquad \text{(A.11)}\end{aligned}$$

By equation (A.10), we obtain

$$\begin{aligned}&\|\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}\|_\infty\\ &\quad\le \|((\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1} - (\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}}))^{-1})(\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0) + \lambda(\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}))\|_\infty\\ &\quad + \|(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}}))^{-1}(\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0) + \lambda(\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}))\|_\infty\\ &\quad\le \sqrt{s_0}\|(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1} - (\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}}))^{-1})\|_2(\|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty\\ &\quad + \lambda(\|\hat{f}_{\mathcal{J}_0}\|_\infty + \|\hat{h}_{\mathcal{J}_0}\|_\infty)) + \|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \lambda(\|\hat{f}_{\mathcal{J}_0}\|_\infty + \|\hat{h}_{\mathcal{J}_0}\|_\infty)\\ &\quad\precsim \sqrt{s_0}\|(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1} - (\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}}))^{-1})\|_2\\ &\quad\times \Big(\|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \lambda(\max_{j\in\mathcal{J}_0}\tau_j + \xi)\Big)\\ &\quad + \sqrt{s_0}\Big(\|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \lambda(\max_{j\in\mathcal{J}_0}\tau_j + \xi)\Big),\end{aligned}$$

where we use the fact that $|\hat{f}_j| \le \tau_j$ and $|\hat{h}_j| \le \xi$ for each j in the last inequality.

Because the X_i's are sub-Gaussian and $b''(\cdot)$ is uniformly bounded by assumption, then for any v, $w \in \mathbb{R}^p$, the expression

$$v^{\mathrm{T}}\nabla^2\mathcal{L}_n(\bar{\boldsymbol{\beta}})w = \frac{1}{n}\sum_{i=1}^n b''(X_i^{\mathrm{T}}\bar{\boldsymbol{\beta}})v^{\mathrm{T}}X_i \cdot w^{\mathrm{T}}X_i$$

is the i.i.d. average of the product of sub-Gaussian variables $b''(X_i^{\mathrm{T}}\bar{\boldsymbol{\beta}})v^{\mathrm{T}}X_i$ and $w^{\mathrm{T}}X_i$. By Lemma A.2, a standard discretization argument of the s_0-dimensional unit sphere yields

$$\|\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}) - \nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}})\|_2 \precsim \sqrt{\frac{s_0}{n}}$$

with probability at least $1 - c_1\exp(-c_2 s_0)$. Moreover, for any invertible matrix A and B, if $\|A^{-1}\|\|A - B\|_2 \le 1/2$, then $\|A^{-1} - B^{-1}\|_2 = O(\|A^{-1}\|\|A - B\|_2$. Therefore, we have

$$\|(\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}))^{-1} - (\nabla^2_{\mathcal{J}_0\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}}))^{-1}\|_2 \precsim \sqrt{\frac{s_0}{n}}$$

as well. A similar argument shows that

$$\max_{j\in\mathcal{J}_0^c}\|e_j^{\mathrm{T}}(\nabla^2_{\mathcal{J}_0^c\mathcal{J}_0}\mathcal{L}_n(\bar{\boldsymbol{\beta}}) - \nabla^2_{\mathcal{J}_0^c\mathcal{J}_0}\mathcal{L}(\bar{\boldsymbol{\beta}}))\|_2 \precsim \max\left\{\sqrt{\frac{s_0}{n}}, \sqrt{\frac{\log p}{n}}\right\}$$

with probability at least $1 - c_1 \exp(-c_2 \min\{s_0, \log p\})$.

Then based on Assumption 3 and 4, $\sqrt{s_0^2/n} \to 0$. According to the proof of Theorem 4.1, we have $\|\nabla_{\mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty \leq \sqrt{\frac{\log p}{n}}$ with probability at least $1 - c_1' \exp(-c_2' \log p)$. Thus,

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}\|_\infty \precsim \frac{\sqrt{s_0}}{C_{\min}} \left(\sqrt{\frac{\log p}{n}} + \lambda (\max_{j \in \mathcal{J}_0} \tau_j + \xi) \right)$$

with probability at least $1 - c_1' \exp(-c_2' \log p)$.

Hence, according to the assumption that $\beta^0_{\min} > \frac{\sqrt{s_0}}{C_{\min}} \left(\sqrt{\frac{\log p}{n}} + \lambda (\max_{j \in \mathcal{J}_0} \tau_j + \xi) \right)$, we have

$$P(\Omega_1) = P(\|\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}\|_\infty < \beta^0_{\min}) \geq 1 - c_1' \exp(-c_2' \log p) \to 1. \tag{A.12}$$

Next, by (A.8) and (A.9), we have

$$\begin{aligned}
\mathcal{Q}^n_{\mathcal{J}_0 \mathcal{J}_0} (\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}) &= -\nabla_{\mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0) - \lambda (\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}) - R^n_{\mathcal{J}_0} \\
\mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}) &= -\nabla_{\mathcal{N}_j} \mathcal{L}_n(\boldsymbol{\beta}^0) - \lambda (\hat{f}_{\mathcal{N}_j} + \hat{h}_{\mathcal{N}_j}) - R^n_{\mathcal{N}_j},
\end{aligned}$$

where $Q^n_{\mathcal{J}_0 \mathcal{J}_0} = \nabla^2_{\mathcal{J}_0 \mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0)$, $Q^n_{\mathcal{N}_j \mathcal{J}_0} = \nabla^2_{\mathcal{N}_j \mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0)$ and the remainder $R^n = (\nabla^2 \mathcal{L}_n(\bar{\boldsymbol{\beta}}) - \nabla^2 \mathcal{L}_n(\boldsymbol{\beta}^0))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. Hence,

$$\begin{aligned}
\hat{f}_{\mathcal{N}_j} = &\frac{1}{\lambda} \mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (Q^n_{\mathcal{J}_0 \mathcal{J}_0})^{-1} (\nabla_{\mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0) + R^n_{\mathcal{J}_0}) + \mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (Q^n_{\mathcal{J}_0 \mathcal{J}_0})^{-1} (\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}) \\
&- \frac{1}{\lambda} (\nabla_{\mathcal{N}_j} \mathcal{L}_n(\boldsymbol{\beta}^0) + R^n_{\mathcal{N}_j}) - \hat{h}_{\mathcal{N}_j}.
\end{aligned}$$

Then, for each $j \in \mathcal{J}_0^c$,

$$\begin{aligned}
\frac{\|\hat{f}_{\mathcal{N}_j}\|_2}{\tau_j} \leq &\frac{1}{\lambda \tau_j} \left\| \mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (Q^n_{\mathcal{J}_0 \mathcal{J}_0})^{-1} \left(\nabla_{\mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0) + R^n_{\mathcal{J}_0} \right) \right\|_2 \\
&+ \frac{1}{\tau_j} \left\| \mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (Q^n_{\mathcal{J}_0 \mathcal{J}_0})^{-1} (\hat{f}_{\mathcal{J}_0} + \hat{h}_{\mathcal{J}_0}) \right\|_2 \\
&+ \frac{1}{\lambda \tau_j} \left\| \nabla_{\mathcal{N}_j} \mathcal{L}_n(\boldsymbol{\beta}^0) + R^n_{\mathcal{N}_j} \right\|_2 + \frac{1}{\tau_j} \|\hat{h}_{\mathcal{N}_j}\|_2 \\
&\leq \frac{1}{\lambda \tau_j} \left\| \mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (Q^n_{\mathcal{J}_0 \mathcal{J}_0})^{-1} \right\|_2 (\|\nabla_{\mathcal{J}_0} \mathcal{L}_n(\boldsymbol{\beta}^0)\|_2 + \|R^n_{\mathcal{J}_0}\|_2) \\
&+ \frac{1}{\tau_j} \left\| \mathcal{Q}^n_{\mathcal{N}_j \mathcal{J}_0} (Q^n_{\mathcal{J}_0 \mathcal{J}_0})^{-1} \right\|_2 (\|\hat{f}_{\mathcal{J}_0}\|_2 + \|\hat{h}_{\mathcal{J}_0}\|_2) \\
&+ \frac{\sqrt{d_j}}{\lambda \tau_j} \left\| \nabla_{\mathcal{N}_j} \mathcal{L}_n(\boldsymbol{\beta}^0) + R^n_{\mathcal{N}_j} \right\|_\infty + \frac{\sqrt{d_j}}{\tau_j} \|\hat{h}_{\mathcal{N}_j}\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\sqrt{d_j}}{\lambda\tau_j}\left\|\mathcal{Q}^n_{\mathcal{N}_j\mathcal{J}_0}(Q^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\right\|_\infty \sqrt{s_0}(\|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \|R^n_{\mathcal{J}_0}\|_\infty)\\
&+ \frac{\sqrt{d_j}}{\lambda\tau_j}(\|\nabla_{\mathcal{N}_j}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \|R^n_{\mathcal{N}_j}\|_\infty)\\
&+ \frac{\sqrt{d_j}}{\tau_j}\left\|\mathcal{Q}^n_{\mathcal{N}_j\mathcal{J}_0}(Q^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\right\|_\infty \sqrt{s_0}(\max_{j\in\mathcal{J}_0}\tau_j + \xi) + \frac{\sqrt{d_j}}{\tau_j}\xi.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\max_{j\in\mathcal{J}_0^c}\frac{\|\hat{f}_{\mathcal{N}_j}\|_2}{\tau_j} &\leq \left\|\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0}(Q^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\right\|_\infty \sqrt{s_0}(\|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \|R^n_{\mathcal{J}_0}\|_\infty)\max_{j\in\mathcal{J}_0^c}\frac{\sqrt{d_j}}{\lambda\tau_j}\\
&+ (\|\nabla_{\mathcal{J}_0^c}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty + \|R^n_{\mathcal{J}_0^c}\|_\infty)\max_{j\in\mathcal{J}_0^c}\frac{\sqrt{d_j}}{\lambda\tau_j}\\
&+ \left\|\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0}(Q^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\right\|_\infty \max_{j\in\mathcal{J}_0^c}\frac{\sqrt{d_j}\sqrt{s_0}(\max_{j\in\mathcal{J}_0}\tau_j + \xi)}{\tau_j}\\
&+ \max_{j\in\mathcal{J}_0^c}\frac{\sqrt{d_j}\xi}{\tau_j}.
\end{aligned}
$$

By the proof of Theorem 4.1, we have $\|\nabla_{\mathcal{J}_0}\mathcal{L}_n(\boldsymbol{\beta}^0)\|_\infty \leq \sqrt{\frac{\log p}{n}}$ with probability at least $1 - c_1'\exp(-c_2'\log p)$.

Next, we consider the rate of $\|R^n\|_\infty$,

$$
\begin{aligned}
R^n &= (\nabla^2\mathcal{L}_n(\bar{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}_n(\boldsymbol{\beta}^0))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\\
&= \frac{1}{n}\sum_{i=1}^n [b''(X_i^{\mathrm{T}}\bar{\boldsymbol{\beta}}) - b''(X_i^{\mathrm{T}})]X_iX_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)
\end{aligned}
$$

for some point $\bar{\boldsymbol{\beta}} = t\widehat{\boldsymbol{\beta}} + (1-t)\boldsymbol{\beta}^0$. Using the mean value theorem and the bounded condition of $b'''(\cdot)$ gives

$$
R^n = \frac{1}{n}\sum_{i=1}^n b'''(X_i^{\mathrm{T}}\bar{\bar{\boldsymbol{\beta}}})X_i(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^{\mathrm{T}}X_iX_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0),
$$

where $\bar{\bar{\boldsymbol{\beta}}}$ lies on the line segment between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^0$. Then, for each j,

$$
R^n_j = \frac{1}{n}\sum_{i=1}^n b'''(X_i^{\mathrm{T}}\bar{\bar{\boldsymbol{\beta}}})X_{ij}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^{\mathrm{T}}X_iX_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0).
$$

Let $a_i = b'''(X_i^{\mathrm{T}}\bar{\bar{\boldsymbol{\beta}}})X_{ij}$ and $d_i = (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^{\mathrm{T}}X_iX_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$, we have

$$
|R^n_j| = \frac{1}{n}|\sum_{i=1}^n a_id_i| \leq \frac{1}{n}\|\mathbf{a}\|_\infty\|\mathbf{d}\|_1.
$$

By assumption $\|\mathbf{a}\|_\infty \le M$ for some constant $M > 0$, and note that $\mathrm{supp}(\widehat{\boldsymbol{\beta}}) \subseteq \mathcal{J}_0$, then

$$\begin{aligned}\frac{1}{n}\|\mathbf{d}\|_1 &\precsim t_j \left\| \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^{\mathrm{T}} \right)_{\mathcal{J}_0\mathcal{J}_0} \right\|_2 \|\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}\|_2^2 \\ &\precsim \left(\Lambda_{\max}(\Sigma) + \sqrt{\frac{s_0}{n}} \right) \|\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0} - \boldsymbol{\beta}^0_{\mathcal{J}_0}\|_2^2 \\ &\precsim \left(\Lambda_{\max}(\Sigma) + \sqrt{\frac{s_0}{n}} \right) \lambda^2 \left(\frac{\mathcal{K}}{\tau_{\min} \wedge \xi} \right)^2\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \min\{s_0, \log p\})$, where the second inequality holds by a standard spectral norm bound on the sample covariance matrix and the last inequality holds by Theorem 4.1. Therefore,

$$\frac{\|R^n\|_\infty}{\lambda} \precsim \lambda \left(\frac{\mathcal{K}}{\tau_{\min} \wedge \xi} \right)^2. \tag{A.13}$$

Finally, by Assumption 4 and 5, it remains to show that there are universal constants (c_1, c_2) such that

$$P(\|\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0}(Q^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_\infty \ge 1 - \zeta/2) \le c_1 \exp(-c_2 \min\{s_0, \log p\}). \tag{A.14}$$

We begin by decomposing the sample matrix as the sum $\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0}(Q^n_{\mathcal{J}_0\mathcal{J}_0})^{-1} = T_1 + T_2 + T_3 + T_4$ where we define

$$T_1 := \mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0}((\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0})^{-1} - (\mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})^{-1}), \tag{A.15}$$

$$T_2 := (\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0})(\mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})^{-1}, \tag{A.16}$$

$$T_3 := (\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0})((\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0})^{-1} - (\mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})^{-1}), \tag{A.17}$$

$$T_4 := \mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0}(\mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})^{-1}. \tag{A.18}$$

By the incoherence condition 5, we have

$$\|T_4\|_\infty \le 1 - \zeta.$$

For T_1,

$$\begin{aligned}\|T_1\|_\infty &= \|\mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0}(Q_{\mathcal{J}_0\mathcal{J}_0})^{-1}(\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})(\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_\infty \\ &\le \|\mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0}(Q_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_\infty \|\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0}\|_\infty \|(\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_\infty \\ &\le s_0(1-\zeta)\|\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0}\|_2 \|(\mathcal{Q}^n_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_2 \\ &\le (1-\zeta)\sqrt{\frac{s_0^3}{n}} \frac{1}{C_{\min} + \sqrt{\frac{s_0}{n}}} \le (1-\zeta)\sqrt{\frac{s_0^3}{n}} \frac{2}{C_{\min}}\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 s_0)$, where the third inequality holds by Lemma A.2. By Assumption 3, we have $\|T_1\|_\infty \le \zeta/6$ with probability at least $1 - c_1 \exp(-c_2 s_0)$. For T_2,

$$\|T_2\|_\infty \le \sqrt{s_0} \max_{j \in \mathcal{J}_0^c} \|e_j^{\mathrm{T}}(\mathcal{Q}^n_{\mathcal{J}_0^c\mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0^c\mathcal{J}_0})(\mathcal{Q}_{\mathcal{J}_0\mathcal{J}_0})^{-1}\|_2$$

$$\leq \sqrt{s_0} \max_{j \in \mathcal{J}_0^c} \|e_j^{\mathrm{T}} (\mathcal{Q}^n_{\mathcal{J}_0^c \mathcal{J}_0} - \mathcal{Q}_{\mathcal{J}_0^c \mathcal{J}_0}) \|_2 \|(\mathcal{Q}_{\mathcal{J}_0 \mathcal{J}_0})^{-1}\|_2$$
$$\leq \frac{2\sqrt{s_0}}{C_{\min}} \max\left\{ \sqrt{\frac{s_0}{n}}, \sqrt{\frac{\log p}{n}} \right\}$$

with probability at least $1 - c_1 \exp(-c_2 \min\{s_0, \log p\})$, where the third inequality holds by Lemma A.2. Based on Assumption 3, we have $\|T_2\|_\infty \leq \zeta/6$ with probability at least $1 - c_1 \exp(-c_2 \min\{s_0, \log p\})$. For T_3, a similar argument shows that,

$$\|T_3\|_\infty \leq \sqrt{s_0} \max\left\{ \sqrt{\frac{s_0}{n}}, \sqrt{\frac{\log p}{n}} \right\}^2.$$

According to Assumption 3, we have $\|T_3\|_\infty \leq \zeta/6$ with probability greater than or equal to $1 - c_1 \exp(-c_2 \min\{s_0, \log p\})$. Combining all the results above, we conclude that (A.14) holds.

Therefore,

$$\begin{aligned}
\max_{j \in \mathcal{J}_0^c} \frac{\|\hat{f}_{\mathcal{N}_j}\|_2}{\tau_j} \leq\ & (1 - \zeta/2) \frac{1}{\lambda} \sqrt{\frac{s_0 \log p}{n}} \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j}}{\tau_j} + \frac{1}{\lambda} \sqrt{\frac{\log(p - s_0)}{n}} \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j}}{\tau_j} \\
& + (1 - \zeta/2) \sqrt{s_0} \left(\frac{\mathcal{K}}{\tau_{\min} \wedge \xi} \right)^2 \lambda \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j}}{\tau_j} \\
& + \left(\frac{\mathcal{K}}{\tau_{\min} \wedge \xi} \right)^2 \lambda \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j}}{\tau_j} \\
& + \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j} \xi}{\tau_j} + (1 - \zeta/2) \frac{\sqrt{s_0} (\max_{j \in \mathcal{J}_0} \tau_j + \xi)}{\min_{j \in \mathcal{J}_0^c} m_j}
\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \min\{s_0, \log p\})$. By the conditions $\lambda(\tau_{\min} + \xi) \geq c\sqrt{d_{\max} \log p / n}$, $\sqrt{s_0}(\sqrt{s_0} \max_{j \in \mathcal{J}_0} m_j \vee \xi) = o(\min_{\mathcal{J}_0^c} m_j)$ and

$$\frac{s_0^{1/2} \lambda \left(\mathcal{K} / (\tau_{\min} \wedge \xi) \right)^2 + \xi}{\min_{j \in \mathcal{J}_0^c} m_j} \to 0,$$

we obtain

$$\frac{1}{\lambda} \sqrt{\frac{s_0 \log p}{n}} \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j}}{\tau_j} \leq c \sqrt{\frac{s_0}{d_{\max}}} \frac{\tau_{\min} + \xi}{\min_{j \in \mathcal{J}_0^c} m_j} \leq \frac{\sqrt{s_0} (\min_{1 \leq j \leq p} m_j \vee \xi)}{\min_{j \in \mathcal{J}_0^c} m_j} \to 0,$$

$$\frac{\sqrt{s_0} (\max_{j \in \mathcal{J}_0} \tau_j + \xi)}{\min_{j \in \mathcal{J}_0^c} m_j} \leq \frac{\sqrt{s_0} (\sqrt{s_0} \max_{j \in \mathcal{J}_0} m_j + \xi)}{\min_{j \in \mathcal{J}_0^c} m_j} \to 0,$$

and

$$\sqrt{s_0} \left(\frac{\mathcal{K}}{\tau_{\min} \wedge \xi} \right)^2 \lambda \max_{j \in \mathcal{J}_0^c} \frac{\sqrt{d_j}}{\tau_j} \leq \frac{s_0^{1/2} \lambda}{\min_{j \in \mathcal{J}_0^c} m_j} \left(\frac{\mathcal{K}}{\tau_{\min} \wedge \xi} \right)^2 \to 0.$$

Therefore,

$$P(\Omega_2) = P\left(\max_{j\in\mathcal{J}_0^c} \frac{\|\hat{f}_{\mathcal{N}_j}\|_2}{\tau_j} < 1\right) \geq 1 - c_1 \exp(-c_2 \min\{s_0, \log p\}) \to 1. \quad \text{(A.19)}$$

By (A.12) and (A.19), we conclude that $P(\Omega_1 \cap \Omega_2) \to 1$ and $P(\text{sign}(\widehat{\boldsymbol{\beta}}) = \text{sign}(\boldsymbol{\beta}^0)) \to 1$ as $n \to \infty$. □

Appendix B: Parallel Dykstra-like proximal algorithm

In this section, we describe the Parallel Dykstra-like proximal algorithm. Following [46], in order to find the proximity operator in step 4 of Algorithm 1, we combined an alternative method via proximal splitting methods based on Parallel Dykstra-like proximal algorithm [5]. Following the proofs of Theorem 1 in [47], the $\text{prox}_{\lambda(\|\boldsymbol{\beta}\|_{G,\tau}+\|\boldsymbol{\beta}\|_{S,\xi})/L}(\cdot)$ can be directly derived from $\text{prox}_{\lambda\|\boldsymbol{\beta}\|_{G,\tau}/L}(\cdot)$ by soft thresholding. By the Lemma 1 and Lemma 2 in [42], the proximity operator amounts to a projection operator onto the intersection of active groups. Thus we can use the Parallel Dykstra-like proximal algorithm to find the projection. From the step 5 in Algorithm 2, the projection first enforces sparsity within a neighbourhood by performing the element-wise soft thresholding and then imposes sparsity among the neighbourhoods by performing the group soft thresholding. Denote $\mathcal{T}^{(t)} = \{j : \|\text{ST}(h_{\mathcal{N}_j}^{(t)}, \lambda\xi/L)\|_2 > \lambda\tau_j/L\}$ as an active set, where $\text{ST}(t,\lambda) = \text{sign}(t)(|t| - \lambda)_+$ is a soft thresholding and $(x)_+ := \max\{0, x\}$ is the positive part function. $h_{\mathcal{N}_j}^{(t)}$ is defined in Algorithm 1. The closed form solution is given by

$$\begin{aligned} p_{\mathcal{N}_j}^{j,n} &= \text{Proj}_{\frac{\lambda}{L}(\xi\|\cdot\|_1 + \tau_j\|\cdot\|_2)}(z_{\mathcal{N}_j}^{j,n}) \\ &= \text{Proj}_{\frac{\lambda\tau_j}{L}\|\cdot\|_2}\left(\text{ST}(z_{\mathcal{N}_j}^{j,n}, \lambda\xi/L)\right) \\ &= w_{\mathcal{N}_j}^{j,n}\mathbf{1}(\|w_{\mathcal{N}_j}^{j,n}\|_2 \leq \lambda\tau_j/L) + \frac{\lambda\tau_j w_{\mathcal{N}_j}^{j,n}}{L\|w_{\mathcal{N}_j}^{j,n}\|_2}\mathbf{1}(\|w_{\mathcal{N}_j}^{j,n}\|_2 > \lambda\tau_j/L), \end{aligned}$$

where $w_{\mathcal{N}_j}^{j,n} = \text{ST}(z_{\mathcal{N}_j}^{j,n}, \lambda\xi/L)$. Specifically, we note that the proximate operator can be solved efficiently when the predictor graph G comprises disconnected components.

Acknowledgments

The authors are grateful to the referees, the associate editor and the editor for their insightful comments and suggestions.

Algorithm 2 Parallel Dykstra-like proximal algorithm

Set $x_0 = h^{(t)}$, $z^{j,0} = x_0$ for $j = 1, \ldots, |\mathcal{T}^{(t)}|$, $n = 0$.
repeat
  for $j = 1, \ldots, |\mathcal{T}^{(t)}|$ **do**
    $p^{j,n}_{\mathcal{N}_j^c} = z^{j,n}_{\mathcal{N}_j^c}$;
    $p^{j,n}_{\mathcal{N}_j} = \text{Proj}_{\frac{\lambda}{L}(\xi\|\cdot\|_1 + \tau_j\|\cdot\|_2)}(z^{j,n}_{\mathcal{N}_j})$;
    $x_{n+1} = \sum_{i=1}^{|\mathcal{T}^{(t)}|} \frac{p^{j,n}}{|\mathcal{T}^{(t)}|}$;
  for $i = 1, \ldots, |\mathcal{T}^{(t)}|$ **do**
    $z^{j,n} = x_{n+1} + z^{(j,n)} - p^{(j,n)}$;
  $n = n + 1$;
until convergence;
return x.

References

[1] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2** 183–202. MR2486527

[2] Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123. MR2422825

[3] Botticelli, A., Putignani, L., Zizzari, I., Del Chierico, F., Reddel, S., Di Pietro, F., Quagliarello, A., Onesti, C. E., Raffaele, G., Mazzuca, F. et al. (2018). Changes of microbiome profile during nivolumab treatment in NSCLC patients.

[4] Bunea, F. et al. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics* **2** 1153–1194. MR2461898

[5] Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering* 185–212. Springer. MR2858838

[6] Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics* 505–530. MR1105834

[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360. MR1946581

[8] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. MR2065194

[9] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[10] Gao, X. and Carroll, R. J. (2017). Data integration with high dimensionality. *Biometrika* **104** 251–272. MR3698252

[11] Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. *The*

Annals of Statistics **34** 2367–2386. MR2291503
[12] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.
[13] Horie, M., Miura, T., Hirakata, S., Hosoyama, A., Sugino, S., Umeno, A., Murotomi, K., Yoshida, Y. and Koike, T. (2017). Comparative analysis of the intestinal flora in type 2 diabetes and nondiabetic mice. *Experimental animals* 17–0021.
[14] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics* **27**. MR3025130
[15] Laria, J. C., Carmen Aguilera-Morillo, M. and Lillo, R. E. (2019). An iterative sparse-group lasso. *Journal of Computational and Graphical Statistics* 1–10. MR4007753
[16] Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.
[17] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327. MR1137117
[18] Liu, J., Yu, G. and Liu, Y. (2019). Graph-based sparse linear discriminant analysis for high-dimensional classification. *Journal of Multivariate Analysis* **171** 250–269. MR3898276
[19] Loh, P.-L. et al. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics* **45** 866–896. MR3650403
[20] Loh, P.-L. and Wainwright, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics* **41** 3022–3049. MR3161456
[21] Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *Journal of Machine Learning Research* **16** 559–616. MR3335800
[22] Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J. J. and Gajewski, T. F. (2018). The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science* **359** 104–108.
[23] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* **37**. CRC Press. MR3223057
[24] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of statistics* **34** 1436–1462. MR2278363
[25] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of statistics* **37** 246–270. MR2488351
[26] Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B. et al. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557. MR3025133
[27] Obozinski, G., Jacob, L. and Vert, J.-P. (2011). Group lasso with

overlaps: the latent group lasso approach. *arXiv preprint 1110.0413*. MR3211304

[28] Rao, N., Nowak, R., Cox, C. and Rogers, T. (2015). Classification with the sparse group lasso. *IEEE Transactions on Signal Processing* **64** 448–463. MR3446222

[29] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics* **38** 1287–1319. MR2662343

[30] Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics* **40** 639–665. MR2933661

[31] Schirmer, M., Smeekens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., ter Horst, R., Jansen, T., Jacobs, L., Bonder, M. J. et al. (2016). Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167** 1125–1136.

[32] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22** 231–245. MR3173712

[33] Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T. H. et al. (2017). Influence of diet on the gut microbiome and implications for human health. *Journal of translational medicine* **15** 73.

[34] Spantini, A., Bigoni, D. and Marzouk, Y. (2018). Inference via low-dimensional couplings. *The Journal of Machine Learning Research* **19** 2639–2709. MR3899768

[35] Stephenson, M. (2018). Doubly Sparse Regularized Regression Incorporating Graphical Structure Among Predictors, PhD thesis, University of Guelph.

[36] Stephenson, M., Ali, R. A., Darlington, G. A. and Initiative, A. D. N. (2019). Doubly sparse regression incorporating graphical structure among predictors. *Canadian Journal of Statistics*. MR4035798

[37] Stephenson, M., Ali, R. A., Darlington, G. A., Schenkel, F. S. and Squires, E. J. (2019). DSLRIG: Leveraging predictor structure in logistic regression. *Communications in Statistics-Simulation and Computation* 1–13.

[38] Temraz, S., Nassar, F., Nasr, R., Charafeddine, M., Mukherji, D. and Shamseddine, A. (2019). Gut Microbiome: A Promising Biomarker for Immunotherapy in Colorectal Cancer. *International journal of molecular sciences* **20** 4155.

[39] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

[40] Tibshirani, R. and Friedman, J. (2019). A pliable lasso. *Journal of Computational and Graphical Statistics* **just-accepted** 1–18. MR4085876

[41] Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36** 614–645. MR2396809

[42] Villa, S., Rosasco, L., Mosci, S. and Verri, A. (2014). Proximal

methods for the latent group lasso penalty. *Computational Optimization and Applications* **58** 381–407. MR3201966

[43] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press. MR3967104

[44] Yan, X. and Bien, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science* **32** 531–560. MR3730521

[45] Yang, E., Ravikumar, P., Allen, G. I. and Liu, Z. (2013). On Poisson graphical models. In *NIPS* 1718–1726.

[46] Yu, G. and Liu, Y. (2016). Sparse regression incorporating graphical structure among predictors. *Journal of the American Statistical Association* **111** 707–720. MR3538699

[47] Yuan, L., Liu, J. and Ye, J. (2013). Efficient methods for overlapping group lasso. *IEEE transactions on pattern analysis and machine intelligence* **35** 2104–2116.

[48] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574

[49] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* **38** 894–942. MR2604701

[50] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research* **7** 2541–2563. MR2274449

[51] Zhou, S., Zhou, J., Zhang, B. et al. (2019). High-dimensional generalized linear models incorporating graphical structure among predictors. *Electronic Journal of Statistics* **13** 3161–3194. MR4010596

[52] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101** 1418–1429. MR2279469

[53] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67** 301–320. MR2137327