Methodological Advances in Structured Statistical Learning

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Hongru Zhao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Adam J. Rothman, Adviser

June 2025

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Adam J. Rothman, for his profound knowledge, remarkable patience, and invaluable guidance throughout my Ph.D. journey. His mentorship has been essential to my academic and professional growth.

I am deeply thankful to my research collaborators from the University of Minnesota: Professors Sara Algeri, Tiefeng Jiang, Galin Jones, Xiaoou Li, Vuk Mandic, Aaron J. Molstad, and Xiaotong Shen, as well as Professor Jinchao Xu from King Abdullah University of Science and Technology, for their insightful discussions and collaborative support. The research presented in Chapter 2 was conducted in collaboration with Professors Adam J. Rothman and Aaron J. Molstad, while the work in Chapter 3 was carried out in collaboration with Professor Xiaotong Shen. My sincere appreciation also extends to my committee members, Professors Aaron J. Molstad, William Leeb, and Xiaotong Shen, whose valuable feedback and suggestions significantly enriched my dissertation.

I am grateful for the support and camaraderie of my peers in the School of Statistics: Yuanhao Cai, Li Chen, Dongyu Cui, Erik Floden, Jinwen Fu, Dae Young Ham, Ruoyu He, Arnold Johnsen, Ben Lewis, Chunlin Li, Yifei Liu, Tuan Pham, Fangqiao Tian, R. Oliver Vandenberg, Kun Wang, Yinzhao Wang, Yuyuan Wang, Xiangyu Zhang, and Christina Zhao, who have made my graduate experience both memorable and enjoyable.

Finally, I owe special thanks to my family, especially my father, Haisheng Zhao, and my mother, Yi Li, whose unwavering support, encouragement, and understanding have sustained me throughout my life.

Abstract

This dissertation explores methodological advances in statistical learning, specifically addressing two fundamental challenges: the interpretability of complex association structures in multivariate categorical data analysis, and privacy-preserving distributed inference across heterogeneous datasets.

First, we propose a penalized likelihood framework tailored for multivariate categorical response regression, encompassing classical discrete graphical models such as the Ising model, Potts model, and hypergraph models. Utilizing a distinctive subspace decomposition, the method explicitly captures mutual, joint, and conditionally independent associations between categorical variables, facilitating interpretable representations of association structures. We derive theoretical guarantees, establishing error bounds that hold particularly in high-dimensional contexts. Comprehensive simulation studies demonstrate that our approach achieves greater interpretability and improved predictive accuracy compared to existing methods.

Second, we tackle data-sharing challenges under stringent privacy constraints and site heterogeneity, common in multi-site clinical trials. We introduce a robust distributed algorithm for high-dimensional inference and structure learning. Our heterogeneous model integrates global and site-specific effects, employing nonconvex regularization via difference of convex programming under an ℓ_0 constraint, ensuring selection consistency and computational feasibility. Despite the underlying optimization being NP-hard, our method converges globally in polynomial time under realistic conditions. By exclusively penalizing nuisance parameters, our approach maintains valid statistical inference, directly addressing practical data-sharing constraints.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Multivariate Categorical Response Regression

We consider a multivariate response regression where each of the response variables is categorical. Specifically, let $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ be the predictor vector and let $\boldsymbol{Z} = (Z_1, \cdots, Z_q)^\top$ be the multivariate categorical response. The kth component of the response, Z_k, has J_k numerically coded outcome categories with $J_k \geq 2$ for $k \in [q]$, where $[m]$ is defined as $\{1, \ldots, m\}$ for positive integer m. The essential problem is to model the conditional distribution $\boldsymbol{Z}|\boldsymbol{X} = \boldsymbol{x}$ whose joint probability mass function is given by

$$\pi_{\boldsymbol{j}}(\boldsymbol{x}) := \mathbb{P}(Z_1 = j_1, \cdots, Z_q = j_q \big| \boldsymbol{X} = \boldsymbol{x}) \geq 0, \tag{1.1}$$

for any $\boldsymbol{j} = (j_1, \cdots, j_q) \in \boldsymbol{J} := [J_1] \times \cdots \times [J_q]$, where $j_l \in [J_l]$ for all $l \in [q]$. For a given $\boldsymbol{x}$, $\boldsymbol{Z}$ has a multivariate version of the single-trial multinomial distribution. If, for a given $\boldsymbol{x}$, one were to observe $v \geq 1$ independent realizations of $\boldsymbol{Z}$, say $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_v$, then the probability mass function corresponding to (1.1) would be given by $\frac{v!}{\prod_{\boldsymbol{j} \in \boldsymbol{J}} y_{\boldsymbol{j}}!} \prod_{\boldsymbol{j} \in \boldsymbol{J}} \{\pi_{\boldsymbol{j}}(\boldsymbol{x})\}^{y_{\boldsymbol{j}}}$, where $y_{\boldsymbol{j}} := \sum_{i=1}^{v} \mathbf{1}(\boldsymbol{z}_i = \boldsymbol{j})$ for each $\boldsymbol{j} \in \boldsymbol{J}$.

For a given $\boldsymbol{x}$, if v is sufficiently large, one could model (1.1) using standard methods for the analysis of q-way contingency tables, a classical problem in categorical data analysis (McCullagh and Nelder, 1989; Christensen, 1997; Agresti, 2002). However, when one needs to model (1.1) for all $\boldsymbol{x} \in \mathcal{X}$, methods for contingency tables cannot be applied. For example, in many applications, for every subject in the study we observe (or impose) a distinct $\boldsymbol{x}$, and observe the outcome of only a single trial, $v = 1$. Instead, one could model (1.1) using existing methods for multinomial regression (or in statistical learning terminology, multiclass classification). Notice that (1.1) could be equivalently defined in terms of a "univariate" categorical response variable, $\boldsymbol{Z}^{\star}$, with the cardinality of $\boldsymbol{J}$, $|\boldsymbol{J}|$, many outcome categories: one corresponding to each distinct element of $\boldsymbol{J}$. Letting $f : \boldsymbol{J} \to [|\boldsymbol{J}|]$ be any bijective function, it would thus be natural to model $\mathbb{P}(\boldsymbol{Z}^{\star} = f(\boldsymbol{j}) \mid \boldsymbol{X} = \boldsymbol{x}) = \pi_{\boldsymbol{j}}(\boldsymbol{x}) = \mathbb{P}(Z_1 = j_1, \cdots, Z_q = j_q \big| \boldsymbol{X} = \boldsymbol{x})$ using multinomial logistic regression (Agresti, 2002; Vincent and Hansen, 2014); linear or quadratic discriminant analysis (Hastie et al., 2009; Mai et al., 2019); or nonparametric methods. Modeling the conditional distribution $\boldsymbol{Z}^{\star} \mid \boldsymbol{X}$ using one of these methods is appealing because they allow for arbitrary dependence among the q categorical response variables.

However, off-the-shelf application of methods designed for a univariate categorical response may be problematic. In particular, these methods would fail to exploit that $\boldsymbol{Z}^{\star}$ is constructed from q distinct response variables. This negatively affects both estimation efficiency and interpretability of the fitted model. Moreover, for even moderate q, $|\boldsymbol{J}|$ will be large. As a consequence, with small sample sizes, many outcome category combinations $\boldsymbol{j}$ will not be observed in the training data. If, for example, one used a multinomial logistic regression in this situation, the maximum likelihood estimator would not exist. In this work, we propose a new method for fitting (1.1) that allows practitioners to discover parsimonious and interpretable dependence structures among responses, and exploits the multivariate nature of the response.

To motivate our approach, consider a multinomial logistic regression model for

(1.1) with $\boldsymbol{x} \in \mathbb{R}$ (i.e., $p = 1$),

$$\pi_{\boldsymbol{j}}(\boldsymbol{x}) = \mathbb{P}\left(Z_1 = j_1, \ldots, Z_q = j_q | \boldsymbol{X} = \boldsymbol{x}\right) = \frac{\exp(\boldsymbol{x} \cdot \boldsymbol{\zeta}_{\boldsymbol{j}})}{\sum_{\boldsymbol{j} \in \boldsymbol{J}} \exp(\boldsymbol{x} \cdot \boldsymbol{\zeta}_{\boldsymbol{j}})}, \quad \boldsymbol{j} \in \boldsymbol{J}, \quad \sum_{\boldsymbol{j} \in \boldsymbol{J}} \boldsymbol{\zeta}_{\boldsymbol{j}} = 0, \tag{1.2}$$

where $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_{\boldsymbol{j}}\}_{\boldsymbol{j} \in \boldsymbol{J}}$ is an unknown tensor. In full generality, $\boldsymbol{\zeta} \in \{\boldsymbol{v} \in \mathbb{R}^{[J_1] \times \cdots \times [J_q]} : \sum_{\boldsymbol{j} \in \boldsymbol{J}} \boldsymbol{v}_{\boldsymbol{j}} = 0\}$, which implies no restrictions on the dependence amongst responses: their dependence can be arbitrarily complex. Restrictions on the dependence between responses under (1.1) can often be represented as constraints on the space of the coefficients $\boldsymbol{\zeta}$. For example, in the case that $q = 2$, $J_1 = J_2 = 2$, if $\boldsymbol{\zeta} \in \mathfrak{C}^0$, where $\mathfrak{C}^0 = \{\boldsymbol{\zeta} \in \mathbb{R}^{[2] \times [2]} : \boldsymbol{\zeta}_{(1,1)} + \boldsymbol{\zeta}_{(2,2)} - \boldsymbol{\zeta}_{(1,2)} - \boldsymbol{\zeta}_{(2,1)} = 0\}$, then $Z_1 \perp\!\!\!\perp Z_2 \mid \boldsymbol{X}$. Intuitively, $\mathfrak{C}^0$ is the set of coefficients for which the log odds ratio between the two responses is zero for all $\boldsymbol{x}$. This observation motivated Molstad and Rothman (2023) to propose a regularized maximum likelihood estimator of $\boldsymbol{\zeta}$ that shrinks coefficients towards the set $\mathfrak{C}^0$. For applications with $J_l \geq 2$ and $q \leq 3$, Molstad and Rothman (2023) generalized the set $\mathfrak{C}^0$ to correspond to coefficients with all local log odds ratios equal to zero. Their approach thus allowed practitioners to discover only whether responses are mutually independent ($\boldsymbol{\zeta} \in \mathfrak{C}^0$) or are arbitrarily dependent ($\boldsymbol{\zeta} \notin \mathfrak{C}^0$). When $q \geq 3$, however, there are many other parsimonious dependence structures that are "intermediate" to mutual independence and arbitrary dependence. In this work, we generalize the approach of Molstad and Rothman (2023), allowing practitioners to discover much more complex, interpretable dependence structures.

As we just described, to learn the association structure for (1.1), it is crucial to identify whether the regression coefficients reside within a specific subspace. Representing the linear subspace $\mathfrak{L}$ of $\mathbb{R}^k$ can be approached in two ways: externally and internally. For the external representation, consider $\mathfrak{L} = \ker(\boldsymbol{A}) = \{\boldsymbol{v} \in \mathbb{R}^k; \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}\}$ for some matrix $\boldsymbol{A}$. Then regularization of $\boldsymbol{v}$ towards the subspace

$\ker(\boldsymbol{A})$ can be achieved by penalizing the term $\|\boldsymbol{A}\boldsymbol{v}\|_2$. In this sense, Molstad and Rothman (2023) achieves structure learning via an external subspace representation. In contrast, for the internal representation, we can set $\mathfrak{L} = \text{span}\left(\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_s\}\right)$ and $\boldsymbol{v} = v_1\boldsymbol{e}_1 + v_2\boldsymbol{e}_2 + \cdots + v_k\boldsymbol{e}_k$, where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_k$ form an orthonormal basis for $\mathbb{R}^k$. Then regularization of $\boldsymbol{v}$ towards the subspace span $\left(\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_s\}\right)$ can be achieved by penalizing the terms $v_{s+1}, \cdots, v_k$. This is the approach we take in this paper by selecting an orthonormal basis and penalizing the coordinates to achieve association structure learning.

For example, if $J_1 = J_2 = 2$, we can define $\boldsymbol{\zeta}^* = v_{11}g_1g_1^\top + v_{12}g_1g_2^\top + v_{21}g_2g_1^\top + v_{22}g_2g_2^\top$, where $g_i = \frac{1}{\sqrt{2}}\left(1, (-1)^{i+1}\right)^\top, i \in \{1, 2\}$. Here, $g_1g_1^\top$ represents the *overall association*, $g_1g_2^\top$ denotes the *main association* of the first response, $g_2g_1^\top$ is for the main association for the second response, and $g_2g_2^\top$ captures the *interaction association* between the two responses.

Because $\mathfrak{C}^0 = \text{span}\{g_1g_1^\top, g_1g_2^\top, g_2g_1^\top\}$, if $v_{22} = 0$, then $\boldsymbol{\zeta}^* \in \mathfrak{C}^0$. That is, by carefully constructing the internal subspace representation, sparsity in the corresponding coefficients can imply parsimonious association structures among responses. This observation is central to our methodological developments, and one of our main contributions is the explicit construction of a flexible, interpretable internal subspace representation.

Multivariate categorical response regression without predictors serves as an extension of contingency table analysis, allowing for a more comprehensive examination of categorical variable interrelations. The Poisson log-linear model is used for association structure modeling of multiple categorical responses without predictors, with the connection between log-linear models for frequencies and multinomial response models for proportions being extensively studied (McCullagh and Nelder, 1989; Christensen, 1997).

In this paper, we will study structure learning via an internal subspace representation. We present a reparameterization via subspace decomposition and obtain a unifying framework for both multinomial and Poisson categorical response regression models in high dimensions. Complex dependencies between response variables can be systematically modeled, encompassing all possible association structures, including mutual independence, joint independence, and conditional independence among response variables. To achieve structure learning, we utilize variations of the group lasso (Yuan and Lin, 2006) and overlapping group lasso (Zhao et al., 2009; Jenatton et al., 2011) penalties. We apply accelerated proximal gradient descents algorithm to solve the corresponding convex optimization problems. We prove an error bound that illustrates our estimator's performance in high-dimensional settings. A key theoretical development is our derivation of restricted strong convexity conditions specific to multivariate categorical response regression, which notably incorporates the Rademacher complexity associated with general norm penalties. Finally, simulation studies validate our method's effectiveness in terms of interpretability and prediction accuracy.

1.2 Distributed Algorithm for Heterogeneous Data

Multicenter research, especially with clinical data, provides advantages over single-center studies, including larger sample sizes for enhanced generalizability and collaborative resource-sharing (Sidransky et al., 2009; Cheng et al., 2017). However, privacy regulations often restrict access to individual-level data, complicating efforts to pool data across centers (Barrows Jr and Clayton, 1996). Consequently, there is a pressing need for efficient statistical tools that synthesize evidence while maintaining privacy.

In parallel, federated learning (Konecnỳ et al., 2016; McMahan et al., 2017) aims to train machine learning models on decentralized data without explicitly sharing them. Many recent studies extend federated algorithms to handle heterogeneous data and improve stability (Yu et al., 2024; Khaled et al., 2020; Wang et al., 2019; Han et al., 2025; Guo et al., 2025; Yu et al., 2024).

A key challenge in distributed computation is integrating statistical inference to manage uncertainty with heterogeneous data across different sites. (Duan et al., 2022) introduces a distributed algorithm that considers heterogeneous distributions by including site-specific nuisance parameters essential for reflecting site-specific variations. However, this approach relies on the efficient score function to mitigate the impact of inaccurate estimations of these parameters, which may falter when the number of nuisance parameters exceeds the sample size. Due to the complexities of multiple sites and limited sample sizes at each site, previous research often utilizes regularization to prevent overfitting (Wang et al., 2017; Battey et al., 2018; Jordan et al., 2019). These studies propose communication-efficient distributed algorithms for optimization and regression, underlining the statistical inference complexities in decentralized settings. Yet, they do not account for site-specific nuisance parameters crucial for depicting heterogeneity across sites. Our paper addresses this gap by integrating site-specific nuisance parameters and regularization in a high-dimensional context, facilitating the management of overparametrized settings where the number of parameters substantially exceeds the sample size.

This paper will focus on statistical inference for distributed algorithms in linear models to assimilate heterogeneous data involving regularization. This exploration addresses the crucial requirement for integrating inference with distributed computation, enhancing the precision and reliability of statistical methods within distributed environments. Our approach distinguishes itself from existing methods by employing a likelihood approach for higher efficacy rather than relying on surrogate methods. Specifically, we introduce a linear regression framework designed to estimate the global effect across heterogeneous data sets by integrating data from multiple sites while managing site-specific effects individually. This integration is achieved through the application of regularization techniques. By pooling information from multiple sites to estimate a global effect, the overall sample size increases, leading to more efficient estimation and improved inference quality compared to using data from individual sites alone. Furthermore, we develop algorithms to execute this process utilizing nonlinear regularization via an ℓ_0-constraint. As showed in Theorem 3.1, our constrained Difference of Convex (DC) algorithm with the ℓ_0 projection attains a global minimizer in polynomial time, with probability tending to one under the data generation distribution. This result is in contrast to a negative result that in the worst case scenario there does not exist an algorithm that can resolve this nonconvex minimization in polynomial time (Chen et al., 2017, 2019).

In the context of composite hypotheses, we present a hypothesis test that preserves the parameters of interest without regularization, while applying an ℓ_0-constraint on nuisance parameters, such as numerous site-specific parameters, to enhance the power of the test. We derive the asymptotic distribution of the global effect for inference. Additionally, we establish a theoretical guarantee of the validity of the proposed algorithms. Our key result demonstrates that the algorithm achieves selection consistency, ensuring that the supports of the oracle estimators are subsets of the estimated supports with high probability. Moreover, when the sparsity tuning parameter precisely aligns with the true sparsity level, our estimator achieves support recovery, guaranteeing accurate identification of the true model structure. These theoretical findings underscore the effectiveness of our methodology in high-dimensional settings.

Chapter 2

Subspace Decompositions for Association Structure Learning in Multivariate Categorical Response Regression

2.1 Association structure learning via subspace decomposition

2.1.1 Overview

We will introduce the notation that will be used throughout the rest of the article. Let $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote the maximum and minimum eigenvalues of the real symmetric matrix $\boldsymbol{A}$. For any vector (resp. matrix) $\boldsymbol{X}$, define the Euclidean (resp. Frobenius) norm $\|\boldsymbol{X}\| = \sqrt{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})}$. Let $\mathbf{1}_m$ denote a vector of ones of length m. For matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the same size, define the Frobenius inner product $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{Y})$ and define the operator norm $\|\boldsymbol{X}\|_{\mathrm{op}} = \sqrt{\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X})}$. Define the maximum norm $\|\boldsymbol{X}\|_\infty = \max_{i,j} |x_{ij}|, \boldsymbol{X} = \{x_{i,j}\}_{1 \le i \le n, 1 \le j \le m}$. Let $\|\boldsymbol{\beta}\|_0 = \sum_{j,k} \mathbf{1}(\boldsymbol{\beta}_{j,k} \neq 0)$ for matrix $\boldsymbol{\beta}$. Let I be the identity matrix. When $\boldsymbol{X}$ and $\boldsymbol{Y}$ are matrices, let $\boldsymbol{X} \otimes \boldsymbol{Y}$ denote the Kronecker product between $\boldsymbol{X}$ and $\boldsymbol{Y}$. When $\boldsymbol{U}$ and $\boldsymbol{V}$ are vector spaces, let $\boldsymbol{U} \otimes \boldsymbol{V}$ be the tensor product of $\boldsymbol{U}$ and $\boldsymbol{V}$. Finally, let $\otimes_t(\boldsymbol{u}, \boldsymbol{v})$ denote the tensor product between vector $\boldsymbol{u}$ and $\boldsymbol{v}$.

Assume the response has $q \geq 2$ categorical components with $J_1, \cdots, J_q$ categories, respectively. Define the Cartesian product of an indexed family of sets $\boldsymbol{J} = [J_1] \times [J_2] \times \cdots \times [J_q]$. The cardinality of set $\boldsymbol{J}$ is $|\boldsymbol{J}| = \prod_{i=1}^{q} J_i$. Let $\mathbb{R}^{\boldsymbol{J}}$ and $\mathbb{N}^{\boldsymbol{J}}$ denote the spaces of $\boldsymbol{J}$ arrays with entries that are real numbers and whole numbers, respectively. That is, $\boldsymbol{y} = \{y_{\boldsymbol{j}}\}_{\boldsymbol{j} \in \boldsymbol{J}} \in \mathbb{F}^{\boldsymbol{J}}$ if and only if $y_{\boldsymbol{j}} \in \mathbb{F}$ for any $\boldsymbol{j} \in \boldsymbol{J}$, where field $\mathbb{F}$ can take $\mathbb{R}$ and $\mathbb{N}$. For a q-way array of shape $\boldsymbol{J}$, let $\boldsymbol{y}^{\boldsymbol{J}} = \{y_{\boldsymbol{j}}\}_{\boldsymbol{j} \in \boldsymbol{J}} \in \mathbb{R}^{\boldsymbol{J}}$, where $\boldsymbol{j} = (j_1, \cdots, j_q), j_1 \in [J_1], \cdots, j_q \in [J_q]$. Define the $\boldsymbol{J}$-vectorization of $\boldsymbol{y}^{\boldsymbol{J}}$ as

$$\mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}^{\boldsymbol{J}}) := (y_{1,1,\cdots,1}, y_{2,1,\cdots,1}, \cdots, y_{J_1,1,\cdots,1}, y_{1,2,\cdots,1}, \cdots, y_{J_1,J_2,\cdots,J_q})^\top \in \mathbb{R}^{|\boldsymbol{J}|}. \quad (2.1)$$

Define the inverse $\boldsymbol{J}$-vectorization for vector $\mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}^{\boldsymbol{J}})$ as $\mathrm{vec}_{\boldsymbol{J}}^{-1}(\mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}^{\boldsymbol{J}})) = \boldsymbol{y}^{\boldsymbol{J}}$.

Let the sample data from the ith observational unit be denoted $(\boldsymbol{x}_i, \boldsymbol{y}_i^{\boldsymbol{J}})$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{y}_i^{\boldsymbol{J}} \in \mathbb{N}^{\boldsymbol{J}}$, and let $\boldsymbol{y}_i = \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}}) \in \mathbb{N}^{|\boldsymbol{J}|}$. We assume that $\boldsymbol{y}_1^{\boldsymbol{J}}, \cdots, \boldsymbol{y}_n^{\boldsymbol{J}}$ are independent. Similar to Molstad and Rothman (2023), we use a multinomial logistic regression model for the q response variables. Specifically, we assume that $\boldsymbol{y}_i$ is a multinomial random vector with index $n_i \geq 1$ and category probabilities

$$\mathrm{vec}_{\boldsymbol{J}}\{\boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x}_i)\} = \frac{\exp(\boldsymbol{\theta}\boldsymbol{x}_i)}{\langle \mathbf{1}_{|\boldsymbol{J}|}, \exp(\boldsymbol{\theta}\boldsymbol{x}_i)\rangle} \in \mathbb{R}^{|\boldsymbol{J}|}, \quad (2.2)$$

where $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}| \times p}$. In certain settings, we may treat the elements of $\boldsymbol{y}_i$ as independent Poisson random variables with mean

$$\mathrm{vec}_{\boldsymbol{J}}\{\boldsymbol{\mu}^{\boldsymbol{J}}(\boldsymbol{x}_i)\} = \exp(\boldsymbol{\theta}\boldsymbol{x}_i) \in \mathbb{R}^{|\boldsymbol{J}|}, \quad (2.3)$$

and category probabilities $\mathrm{vec}_{\boldsymbol{J}}\{\boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x}_i)\} = \exp(\boldsymbol{\theta}\boldsymbol{x}_i)\{\langle \mathbf{1}_{|\boldsymbol{J}|}, \exp(\boldsymbol{\theta}\boldsymbol{x}_i)\rangle\}^{-1}$, where

$$\boldsymbol{\mu}^{\boldsymbol{J}}(\boldsymbol{x}) = \{\mu_{\boldsymbol{j}}(\boldsymbol{x})\}_{\boldsymbol{j}\in\boldsymbol{J}} \text{ and } \boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x}) = \{\pi_{\boldsymbol{j}}(\boldsymbol{x})\}_{\boldsymbol{j}\in\boldsymbol{J}}$$

are q-dimensional arrays.

The ith observational unit's contribution to the negative log-likelihood of multinomial and Poisson categorical response models are

$$\ell_{\mathrm{Mult}}(\boldsymbol{\theta}\boldsymbol{x}_i, \boldsymbol{y}_i) = -\langle \boldsymbol{y}_i, \boldsymbol{\theta}\boldsymbol{x}_i\rangle + n_i \cdot \log\left(\langle \mathbf{1}_{|\boldsymbol{J}|}, \exp(\boldsymbol{\theta}\boldsymbol{x}_i)\rangle\right), \qquad n_i = \langle \mathbf{1}_{|\boldsymbol{J}|}, \boldsymbol{y}_i\rangle, \tag{2.4}$$

and

$$\ell_{\mathrm{Pois}}(\boldsymbol{\theta}\boldsymbol{x}_i, \boldsymbol{y}_i) = -\langle \boldsymbol{y}_i, \boldsymbol{\theta}\boldsymbol{x}_i\rangle + \langle \mathbf{1}_{|\boldsymbol{J}|}, \exp(\boldsymbol{\theta}\boldsymbol{x}_i)\rangle. \tag{2.5}$$

The function in (2.4) is sometimes referred to as cross-entropy loss.

When considering the multinomial categorical response model, we impose the constraint $\mathbf{1}_{|\boldsymbol{J}|}^\top\boldsymbol{\theta} = \mathbf{0}$ for identifability. Define the linear subspace $\mathcal{V} = \{\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{J}|} : \mathbf{1}_{|\boldsymbol{J}|}^\top\boldsymbol{\alpha} = \mathbf{0}\}$, and define the orthogonal projection matrix $\boldsymbol{P}_{\mathcal{V}} = (I - |\boldsymbol{J}|^{-1}\mathbf{1}_{|\boldsymbol{J}|}\mathbf{1}_{|\boldsymbol{J}|}^\top)$. Notice that $\ell_{\mathrm{Mult}}(\boldsymbol{\theta}\boldsymbol{x}, \boldsymbol{y}) = \ell_{\mathrm{Mult}}(\boldsymbol{\theta}'\boldsymbol{x}, \boldsymbol{y})$ for any $(\boldsymbol{x}, \boldsymbol{y})$ if and only if $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta} = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}'$ (see Lemma 4.1).

2.1.2 Subspace decomposition

We now introduce the subspace decomposition that allows us to parsimoniously model the mass function of interest. Naturally, the dependence between response variables is arbitrarily complex when $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|\times p}$ without additional constraints. To discover parsimonious association structures, we decompose $\boldsymbol{\theta}$ into a sum of components, each of which spans a particular subspace. Returning to an example from the introduction, when $q = 2$, we can decompose $\boldsymbol{\theta} = \boldsymbol{H}_{\{0\}}\boldsymbol{\beta}_{\{0\}} + \boldsymbol{H}_{\{1\}}\boldsymbol{\beta}_{\{1\}} + \boldsymbol{H}_{\{2\}}\boldsymbol{\beta}_{\{2\}} + \boldsymbol{H}_{\{1,2\}}\boldsymbol{\beta}_{\{1,2\}}$, where each $\boldsymbol{H}$ is a basis matrix and $\boldsymbol{\beta}$ are the corresponding coefficients. With appropriately constructed $\boldsymbol{H}$, if $\boldsymbol{\beta}_{\{1,2\}} = 0$, then the two response variables are independent. We demonstrate how to construct such $\boldsymbol{H}$ in the following.

Example 2.1 (Subspace decomposition of a $J_1 \times J_2$ contingency table)**.** Consider the intercept only model (i.e., $p = 1$ with $\boldsymbol{x}_i = 1$) with $q = 2$, and the categorical responses having $J_1 = 2$ categories for the first response and $J_2 = 3$ categories for the second. We can write $\boldsymbol{\theta} = (a_{1,1}, \ldots, a_{J_1,J_2})^\top \in \mathbb{R}^{J_1 J_2}$ as $\left(\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})\right)_{j_1,j_2} = a_{j_1,j_2}$, where $a_{j_1,j_2} \in \mathbb{R}$ for any $(j_1, j_2) \in \boldsymbol{J}$. Accordingly, we can write

$$\boldsymbol{\theta} = \sum_{j_2=1}^{J_2}\sum_{j_1=1}^{J_1} a_{j_1,j_2}\boldsymbol{e}_{j_2}^{J_2} \otimes \boldsymbol{e}_{j_1}^{J_1} \text{ and } \mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta}) = \sum_{j_2=1}^{J_2}\sum_{j_1=1}^{J_1} a_{j_1,j_2}\boldsymbol{E}_{j_1j_2}^{\boldsymbol{J}},$$

where $\boldsymbol{e}_{j_i}^{J_i}$ is the j_ith standard basis vector for $\mathbb{R}^{J_i}$ (i.e., the j_ith column of I_{J_i}) for $i \in [2]$, and $\boldsymbol{E}_{j_1j_2}^{\boldsymbol{J}}$ denotes the standard basis 2-way array for $\mathbb{R}^{J_1 \times J_2}$, which is defined as the array whose (j_1, j_2)-th entry is 1 and all other entries are 0.

Define $\boldsymbol{U}_m$ as a matrix such that $[\frac{1}{\sqrt{m}}\mathbf{1}_m, \boldsymbol{U}_m]$ is an orthogonal matrix of order m. Let $\mathcal{R}(\boldsymbol{U})$ denote the column space of the matrix $\boldsymbol{U}$. Then, we can rewrite $\mathbb{R}^{J_i}$ as the internal direct sum between $\mathcal{R}(\boldsymbol{U}_{J_i})$ and $\mathcal{R}(\mathbf{1}_{J_i})$, denoted as $\mathbb{R}^{J_i} = \mathcal{R}(\boldsymbol{U}_{J_i}) \oplus \mathcal{R}(\mathbf{1}_{J_i})$. The definitions of the internal direct sum and tensor product can be found in Section 4.1.

Due to the bilinearity of the tensor product, $\mathbb{R}^{J_2} \otimes \mathbb{R}^{J_1}$ can be decomposed into an internal direct sum

$$\begin{aligned}\mathbb{R}^{J_2} \otimes \mathbb{R}^{J_1} =& \{\mathcal{R}(\boldsymbol{U}_{J_2}) \oplus \mathcal{R}(\mathbf{1}_{J_2})\} \otimes \{\mathcal{R}(\boldsymbol{U}_{J_1}) \oplus \mathcal{R}(\mathbf{1}_{J_1})\} \\ =& \{\mathcal{R}(\mathbf{1}_{J_2}) \otimes \mathcal{R}(\mathbf{1}_{J_1})\} \oplus \{\mathcal{R}(\mathbf{1}_{J_2}) \otimes \mathcal{R}(\boldsymbol{U}_{J_1})\} \\ & \oplus \{\mathcal{R}(\boldsymbol{U}_{J_2}) \otimes \mathcal{R}(\mathbf{1}_{J_1})\} \oplus \{\mathcal{R}(\boldsymbol{U}_{J_2}) \otimes \mathcal{R}(\boldsymbol{U}_{J_1})\}.\end{aligned}$$

Consider an isomorphism T from the tensor product $\mathbb{R}^{J_2} \otimes \mathbb{R}^{J_1}$ to $\mathbb{R}^{|\boldsymbol{J}|}$. The isomorphism T is uniquely determined by the change of basis $T(\otimes_t(\boldsymbol{e}_{j_2}^{J_2}, \boldsymbol{e}_{j_1}^{J_1})) = \boldsymbol{e}_{j_2}^{J_2} \otimes \boldsymbol{e}_{j_1}^{J_1}$, where the tensor product $\otimes_t(\boldsymbol{u}, \boldsymbol{v})$ denotes the bilinear map of $(\boldsymbol{u}, \boldsymbol{v})$ from the Cartesian product $\mathbb{R}^{J_2} \times \mathbb{R}^{J_1}$, whose basis can be chosen as $\{\otimes_t(\boldsymbol{e}_{j_2}^{J_2}, \boldsymbol{e}_{j_1}^{J_1}); 1 \leq j_1 \leq J_1, 1 \leq j_2 \leq J_2\}$.

Applying the isomorphism T onto each subspace, we obtain that

$$T(\mathcal{R}(\boldsymbol{V}_2) \otimes \mathcal{R}(\boldsymbol{V}_1)) = \mathcal{R}(\boldsymbol{V}_2 \otimes \boldsymbol{V}_1), \text{ where } \boldsymbol{V}_i \in \left\{\frac{1}{\sqrt{J_i}}\boldsymbol{1}_{J_i}, \boldsymbol{U}_{J_i}\right\} \text{ for } i \in [2].$$

Hence, for $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|}$, we can write

$$\boldsymbol{\theta} = \sum_{\boldsymbol{V}_1 \in \left\{\frac{1}{\sqrt{J_1}}\boldsymbol{1}_{J_1}, \boldsymbol{U}_{J_1}\right\}} \sum_{\boldsymbol{V}_2 \in \left\{\frac{1}{\sqrt{J_2}}\boldsymbol{1}_{J_2}, \boldsymbol{U}_{J_2}\right\}} \Big(\boldsymbol{V}_2 \otimes \boldsymbol{V}_1\Big)\boldsymbol{\alpha}_{\boldsymbol{V}_2, \boldsymbol{V}_1}$$

for vectors $\boldsymbol{\alpha}_{\boldsymbol{V}_2, \boldsymbol{V}_1}$ of appropriate size. More simply, we may write $\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}$, with

$$\begin{aligned} &\boldsymbol{H}_{\{0\}} = \frac{1}{\sqrt{|\boldsymbol{J}|}}\boldsymbol{1}_{|\boldsymbol{J}|}, \quad \boldsymbol{H}_{\{1\}} = \frac{1}{\sqrt{J_2}}\boldsymbol{1}_{J_2} \otimes \boldsymbol{U}_{J_1}, \quad \boldsymbol{H}_{\{2\}} = \boldsymbol{U}_{J_2} \otimes \frac{1}{\sqrt{J_1}}\boldsymbol{1}_{J_1} \\ &\boldsymbol{H}_{\{1,2\}} = \boldsymbol{U}_{J_2} \otimes \boldsymbol{U}_{J_1}, \quad \mathcal{K} = \{\{0\}, \{1\}, \{2\}, \{1, 2\}\}, \end{aligned}$$

where each $\boldsymbol{\beta}_{\boldsymbol{k}}$ is simply the corresponding $\boldsymbol{\alpha}_{\boldsymbol{V}_2, \boldsymbol{V}_1}$. As we will formalize in Lemma 2.1, Lemma 2.2, and Theorem 2.1, $\operatorname{span}(\boldsymbol{H}_{\{0\}})$ is the subspace for overall association, $\operatorname{span}(\boldsymbol{H}_{\{1\}}), \operatorname{span}(\boldsymbol{H}_{\{2\}})$ are the subspaces for marginal association on Z_1 and Z_2, respectively, and $\operatorname{span}(\boldsymbol{H}_{\{1,2\}})$ is the subspace for joint association on (Z_1, Z_2). Because of this, sparsity in the coefficients corresponding to each subspace can imply an interpretable restriction on $\boldsymbol{\theta}$.

The discussion outlined above can be generalized to any $\boldsymbol{J}$, with the corresponding isomorphism, denoted as $T_{\boldsymbol{J}}$, being applicable to each subspace.

Lemma 2.1 (Isomorphism)**.** *Define an isomorphism $T_{\boldsymbol{J}}$ from the tensor product space $\mathbb{R}^{J_q} \otimes \cdots \otimes \mathbb{R}^{J_1}$ to $\mathbb{R}^{|\boldsymbol{J}|}$, which is uniquely determined by the change of basis $T_{\boldsymbol{J}}(\otimes_t(\boldsymbol{e}_{j_q}^{J_q}, \cdots, \boldsymbol{e}_{j_1}^{J_1})) = \boldsymbol{e}_{j_q}^{J_q} \otimes \cdots \otimes \boldsymbol{e}_{j_1}^{J_1}$. Here, $\{\otimes_t(\boldsymbol{e}_{j_q}^{J_q}, \cdots, \boldsymbol{e}_{j_1}^{J_1})\}_{(j_1,\cdots,j_q)\in\boldsymbol{J}}$ and $\{\boldsymbol{e}_{j_q}^{J_q} \otimes \cdots \otimes \boldsymbol{e}_{j_1}^{J_1}\}_{(j_1,\cdots,j_q)\in\boldsymbol{J}}$ denote the basis of $\mathbb{R}^{J_q} \otimes \cdots \otimes \mathbb{R}^{J_1}$ and $\mathbb{R}^{|\boldsymbol{J}|}$, respectively. Then,*

(i) for any vector $\boldsymbol{v}_i \in \mathbb{R}^{J_i}, i \in [q]$, we have

$$T_{\boldsymbol{J}}\Big(\otimes_t (\boldsymbol{v}_q, \cdots, \boldsymbol{v}_1)\Big) = \boldsymbol{v}_q \otimes \cdots \otimes \boldsymbol{v}_1, \tag{2.6}$$

(ii) and for any $\boldsymbol{V}_i \in \{\frac{1}{\sqrt{J_i}}\boldsymbol{1}_{J_i}, \boldsymbol{U}_{J_i}\}$, $i \in [q]$, we have

$$T_{\boldsymbol{J}}\Big(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)\Big) = \mathcal{R}\Big(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1\Big). \tag{2.7}$$

In Lemma 2.1, the isomorphism $T_{\boldsymbol{J}}$ is not only a vector space isomorphism between $\mathbb{R}^{J_q} \otimes \cdots \otimes \mathbb{R}^{J_1}$ and $\mathbb{R}^{|\boldsymbol{J}|}$, i.e., $T_{\boldsymbol{J}}(\boldsymbol{v}+\boldsymbol{w}) = T_{\boldsymbol{J}}(\boldsymbol{v})+T_{\boldsymbol{J}}(\boldsymbol{w})$, $T_{\boldsymbol{J}}(a\boldsymbol{v}) = a\cdot T_{\boldsymbol{J}}(\boldsymbol{v})$, $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{J_q} \otimes \cdots \otimes \mathbb{R}^{J_1}$, $a \in \mathbb{R}$, but also preserves the bilinear structure, i.e., equation (2.6) holds. Here, the Kronecker product serves as the bilinear function.

To summarize Example 2.1 and Lemma 2.1, we define $\boldsymbol{U}_m$ as any matrix such that $[\frac{1}{\sqrt{m}}\boldsymbol{1}_m, \boldsymbol{U}_m]$ is an orthogonal matrix of order m. Without loss of generality, we take

$$\boldsymbol{U}_m = \left[\frac{(1,-1,0,\cdots,0)^\top}{\sqrt{2}}, \frac{(1,1,-2,0,\cdots,0)^\top}{\sqrt{6}}, \cdots, \frac{(1,\cdots,1,-(m-1))^\top}{\sqrt{(m-1)m}}\right].$$

Define the index space $\mathcal{K}_s = \{\boldsymbol{k} = \{k_1, \cdots, k_s\} \subset [q] : 1 \leq k_1 < k_2 < \cdots < k_s \leq q\}$ for $s \in [q]$. Define the order of the $\boldsymbol{k}$-interaction as $\|\boldsymbol{k}\|_0 = s$ and its number of parameters as $|\boldsymbol{k}|_J = (J_{k_1} - 1)\cdots(J_{k_s} - 1)$, if $\boldsymbol{k} \in \mathcal{K}_s$. Thus, the space $\mathcal{K}_s$ defines the set of all possible joint associations of order s. Let $\mathcal{K}_0 = \{\{0\}\}$, and define

$\|\boldsymbol{k}\|_0 = 0$ and $|\boldsymbol{k}|_{\boldsymbol{J}} = 1$ when $\boldsymbol{k} = \{0\}$. In this context, $\boldsymbol{k} = \{0\}$ is treated as an empty set. The number of parameters (per predictor) for all sth order associations is given by $L_s = \sum_{1\le i_1<\cdots<i_s\le q}(J_{i_1}-1)\cdots(J_{i_s}-1), s \in [q]$. Let $L_0 = 1$. Define $\boldsymbol{H}_0 = \frac{\boldsymbol{1}_{|\boldsymbol{J}|}}{\sqrt{|\boldsymbol{J}|}} = \frac{\boldsymbol{1}_{J_q}}{\sqrt{J_q}} \otimes \frac{\boldsymbol{1}_{J_{q-1}}}{\sqrt{J_{q-1}}} \otimes \cdots \otimes \frac{\boldsymbol{1}_{J_2}}{\sqrt{J_2}} \otimes \frac{\boldsymbol{1}_{J_1}}{\sqrt{J_1}}$.

For any $\boldsymbol{k} = \{k_1, \cdots, k_s\} \in \mathcal{K}_s, s \ge 1$, define

$$\boldsymbol{H}_{\boldsymbol{k}} = \boldsymbol{V}_q \otimes \boldsymbol{V}_{q-1} \otimes \cdots \otimes \boldsymbol{V}_2 \otimes \boldsymbol{V}_1, \qquad \boldsymbol{V}_i = \begin{cases} \boldsymbol{U}_{J_i} & i \in \boldsymbol{k} \\ \frac{\boldsymbol{1}_{J_i}}{\sqrt{J_i}} & i \in [q]\backslash\boldsymbol{k} \end{cases}. \tag{2.8}$$

Following from Example 2.1, we see that span$(\boldsymbol{H}_{\boldsymbol{k}})$ is the subspace corresponding to the $\boldsymbol{k} = \{k_1, k_2, \cdots, k_s\}$-joint associations. Importantly, it can be verified that the columns of $\boldsymbol{H}_{\boldsymbol{k}}$ are orthonormal.

Lemma 2.2 (Subspace decomposition)**.** *We can express $\mathbb{R}^{|\boldsymbol{J}|}$ as the orthogonal direct sum of the family $\{\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})\}_{\boldsymbol{k}\in\cup_{s=0}^{q}\mathcal{K}_s}$ of subspaces of $\mathbb{R}^{|\boldsymbol{J}|}$, where $\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})$ denotes the column space of $\boldsymbol{H}_{\boldsymbol{k}}$ and the orthogonal direct sum is defined in Section 4.1. Furthermore, for any $\boldsymbol{k} \in \cup_{s=0}^{q}\mathcal{K}_s$, the orthogonal projection matrix onto $\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})$ is given by $\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top$.*

Lemma 2.2 has two key implications. First, it ensures invariance of our proposed estimator (see Section 2.3.2), and second, it allows us to identify the association encoded by the span of each $\boldsymbol{H}_{\boldsymbol{k}}$, as we discuss in the following subsection.

2.1.3 Reparameterization via subspace decomposition

When fitting (2.2), it is common to restrict the hypothesis space of models to include joint associations of at most order d for some $1 \le d \le q$. For example, if $q = 3$, we only want to consider models with all possible main associations and two-way interaction associations. In this case, we would take $d = 2$. As such, define the space of possible associations as $\mathcal{K} = \cup_{s=0}^{d}\mathcal{K}_s$. We call $\mathcal{K}$ the association index space.

Though $\mathcal{K}$ is a function of d, the maximal order of association considered, we omit notation indicating this dependence for notational simplicity.

The following theorem elucidates how the reparameterization of $\boldsymbol{\theta}$ through our subspace decomposition neatly characterizes relationships of mutual, joint, and conditional independence among categorical responses.

Theorem 2.1 (Sparsity and interpretable models)**.** *Let $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ be a partition of $[q]$. For any $\boldsymbol{k} \in \mathcal{K}_s$, let $\boldsymbol{\beta}_{\boldsymbol{k}}$ be a matrix in $\mathbb{R}^{|\boldsymbol{k}|_J \times p}$. For any $\mathcal{I} = \{i_1, \cdots, i_s\} \subset [q]$ such that $1 \leq i_1 < \cdots < i_s \leq q$, define $\boldsymbol{J}^{\mathcal{I}} = [J_{i_1}] \times \cdots \times [J_{i_s}]$.*

1. *(Mutual and joint independence) Let $\mathcal{S}_{\text{joint}} = \{\boldsymbol{k} \in \mathcal{K} : \exists\ i \in [m]$ such that $\boldsymbol{k} \subset \mathcal{I}_i\}$. If $\boldsymbol{\beta}_{\boldsymbol{k}} = \mathbf{0}$ for all $\boldsymbol{k} \notin \mathcal{S}_{\text{joint}}$, i.e., the parameter $\boldsymbol{\theta}$ under either Poisson (2.3) categorical response model or multinomial (2.2) categorical response model is given by $\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}} = \sum_{\boldsymbol{k} \in \mathcal{S}_{\text{joint}}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}$, then*

$$\pi_{\boldsymbol{j}}(\boldsymbol{x}) = \prod_{l=1}^{m} \pi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}), \tag{2.9}$$

   where $\boldsymbol{j}_{\mathcal{I}_l} = \{j_{i_1^l}, \cdots, j_{i_s^l}\} \in \boldsymbol{J}^{\mathcal{I}^l}$, $\mathcal{I}^l = \{i_1^l, \cdots, i_s^l\}$ for some s (which may depend on l), and $\pi_{\boldsymbol{j}_{\mathcal{I}_l},+}$ is the marginal probability mass function of the responses corresponding to $\mathcal{I}_l$.

2. *(Conditional independence) Let $\mathcal{S}_{\text{joint}|\mathcal{I}_m} = \{\boldsymbol{k} \in \mathcal{K} : \exists\ i \in [m-1]$ such that $\boldsymbol{k} \subset \mathcal{I}_i \cup \mathcal{I}_m\}$. If $\boldsymbol{\beta}_{\boldsymbol{k}} = \mathbf{0}$ for all $\boldsymbol{k} \notin \mathcal{S}_{\text{joint}|\mathcal{I}_m}$, i.e., the parameter $\boldsymbol{\theta}$ under either Poisson (2.3) categorical response model or multinomial (2.2) categorical response model is given by $\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}} = \sum_{\boldsymbol{k} \in \mathcal{S}_{\text{joint}|\mathcal{I}_m}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}$, then*

$$\pi_{\boldsymbol{j}_{\mathcal{I}_1}, \boldsymbol{j}_{\mathcal{I}_2}, \cdots, \boldsymbol{j}_{\mathcal{I}_{m-1}} | \boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \prod_{l=1}^{m-1} \pi_{\boldsymbol{j}_{\mathcal{I}_l},+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}), \tag{2.10}$$

   where $\pi_{\boldsymbol{j}_{\mathcal{I}_1}, \cdots, \boldsymbol{j}_{\mathcal{I}_{m-1}} | \boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \frac{\pi_{\boldsymbol{j}}(\boldsymbol{x})}{\pi_{\boldsymbol{j}_{\mathcal{I}_m},+}(\boldsymbol{x})}$, and $\pi_{\boldsymbol{j}_{\mathcal{I}_l},+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \frac{\pi_{\boldsymbol{j}_{\mathcal{I}_l \cup \mathcal{I}_m},+}(\boldsymbol{x})}{\pi_{\boldsymbol{j}_{\mathcal{I}_m},+}(\boldsymbol{x})}$.

To illustrate the practical implications and applications of Theorem 2.1, we present the following example.

Example 2.2. Suppose $q = 4$. The following types of dependence structures—akin to those in Chapter 6 of McCullagh and Nelder (1989)—are encoded in the sparsity of the $\boldsymbol{\beta}_{\boldsymbol{k}}$. Recall that the random multivariate categorical response is $(Z_1, \cdots, Z_q) \in \boldsymbol{J}$.

1. **Mutual independence.** If $\boldsymbol{\theta} = \boldsymbol{H}_{\{0\}}\boldsymbol{\beta}_{\{0\}} + \boldsymbol{H}_{\{1\}}\boldsymbol{\beta}_{\{1\}} + \boldsymbol{H}_{\{2\}}\boldsymbol{\beta}_{\{2\}} + \boldsymbol{H}_{\{3\}}\boldsymbol{\beta}_{\{3\}} + \boldsymbol{H}_{\{4\}}\boldsymbol{\beta}_{\{4\}}$, then Z_1, Z_2, Z_3 and Z_4 are mutually independent for any given $\boldsymbol{x}$, i.e., for all $\boldsymbol{x} \in \mathcal{X}$, $\pi_{j_1,j_2,j_3,j_4}(\boldsymbol{x}) = \pi_{j_1,+,+,+}(\boldsymbol{x}) \cdot \pi_{+,j_2,+,+}(\boldsymbol{x}) \cdot \pi_{+,+,j_3+,}(\boldsymbol{x}) \cdot \pi_{+,+,+,j_4}(\boldsymbol{x})$, for all $(j_1, j_2, j_3, j_4) \in \boldsymbol{J}$.

2. **Joint independence.** If $\boldsymbol{\theta} = \sum_{i=0}^{4} \boldsymbol{H}_{\{i\}}\boldsymbol{\beta}_{\{i\}} + \Big(\boldsymbol{H}_{\{2,3\}}\boldsymbol{\beta}_{\{2,3\}} + \boldsymbol{H}_{\{2,4\}}\boldsymbol{\beta}_{\{2,4\}} + \boldsymbol{H}_{\{3,4\}}\boldsymbol{\beta}_{\{3,4\}} + \boldsymbol{H}_{\{2,3,4\}}\boldsymbol{\beta}_{\{2,3,4\}}\Big)$, then the variable Z_1 is jointly independent of $\{Z_2, Z_3, Z_4\}$ for any given $\boldsymbol{x}$, i.e., for all $\boldsymbol{x} \in \mathcal{X}$, $\pi_{j_1,j_2,j_3,j_4}(\boldsymbol{x}) = \pi_{j_1,+,+,+}(\boldsymbol{x}) \cdot \pi_{+,j_2,j_3,j_4}(\boldsymbol{x}),$ for all $(j_1, j_2, j_3, j_4) \in \boldsymbol{J}$.

3. **Conditional independence.** If $\boldsymbol{\theta} = \sum_{i=0}^{4} \boldsymbol{H}_{\{i\}}\boldsymbol{\beta}_{\{i\}} + \Big(\sum_{2 \le i < j \le 4} \boldsymbol{H}_{\{i,j\}}\boldsymbol{\beta}_{\{i,j\}} + \boldsymbol{H}_{\{2,3,4\}}\boldsymbol{\beta}_{\{2,3,4\}}\Big) + \boldsymbol{H}_{\{1,4\}}\boldsymbol{\beta}_{\{1,4\}}$, then the variable Z_1 and $\{Z_2, Z_3\}$ are conditionally independent for any given $\boldsymbol{x}$ and Z_4, i.e., for all $\boldsymbol{x} \in \mathcal{X}$, $\pi_{j_1,j_2,j_3|j_4}(\boldsymbol{x}) = \pi_{j_1,+,+|j_4}(\boldsymbol{x}) \cdot \pi_{+,j_2,j_3|j_4}(\boldsymbol{x}),$ for all $(j_1, j_2, j_3, j_4) \in \boldsymbol{J}$.

The neat interpretations in Example 2 rely partly on a hierarchical structure of the associations. That is, high-order associations are included only if all the corresponding low-order associations are included. Formally, if association $\boldsymbol{k}$ is included in the model, then all $\boldsymbol{k}' \in \mathcal{K}$ such that $\boldsymbol{k}' \subset \boldsymbol{k}$ must also be included in the model. For example, with $q = 3$, if the joint association $\{1, 2, 3\}$ is included in the model, then for the hierarchy to be enforced, the associations $\{0\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}$, and $\{2, 3\}$ must all be included in the model.

Formally, given an association index space $\mathcal{K}$, the corresponding class of *hierarchical association index space* is the collection of all sets $\mathcal{N} \subset \mathcal{K}$ such that if $\boldsymbol{k} \in \mathcal{N}$, then $\mathcal{P}(\boldsymbol{k}) \subset \mathcal{N}$, where $\mathcal{P}(\boldsymbol{k})$ denotes the powerset of $\boldsymbol{k}$ (with the null set replaced with $\{0\}$).

To restrict attention only to models that respect such a hierarchy, it is natural to consider a class of hierarchical hypotheses spaces

$$\left\{\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|\times p} : \mathcal{R}(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}\in\mathcal{N}} \mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}}) \subset \mathbb{R}^{|\boldsymbol{J}|},\ \ \mathcal{N} \subset \mathcal{K} \ \text{ s.t. } \ \boldsymbol{k} \in \mathcal{N} \implies \mathcal{P}(\boldsymbol{k}) \subset \mathcal{N}\right\}. \tag{2.11}$$

In the next subsection, we will propose a penalized maximum likelihood estimator that allows to explore models in $\mathcal{K}$ or its corresponding hierarchical association index space.

2.2 Penalized likelihood-based association learning

2.2.1 Penalized maximum likelihood estimation

Define the negative log-likelihoods as L_n^{Mult}, and its reparametarized versions as $\mathcal{L}_n^{\text{Mult}}$ where $L_n^{\text{Mult}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell_{\text{Mult}}(\boldsymbol{\theta}\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $\mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}) = L_n^{\text{Mult}}(\boldsymbol{H}\boldsymbol{\beta})$. Similarly define $\mathcal{L}_n^{\text{Pois}}(\boldsymbol{\beta}) = L_n^{\text{Pois}}(\boldsymbol{H}\boldsymbol{\beta})$. To simplify the notation and unify the statements and analysis, set $\mathcal{L}_n(\boldsymbol{\beta}) = \mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta})$ when $\ell = \ell_{\text{Mult}}$, and set $\mathcal{L}_n(\boldsymbol{\beta}) = \mathcal{L}_n^{\text{Pois}}(\boldsymbol{\beta})$ when $\ell = \ell_{\text{Pois}}$ As described in the previous section, due to our subspace decomposition, association structure learning is achieved by learning the sparsity pattern of $\boldsymbol{\beta} \in \mathbb{R}^{\sum_{s=0}^d L_s \times p}$. For this, we will use penalized maximum likelihood estimators of the form

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \left\{\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega(\boldsymbol{\beta})\right\}, \tag{2.12}$$

for convex penalties $\Omega : \mathbb{R}^{\sum_{s=0}^{d} L_s \times p} \to [0, \infty)$ to be discussed in the next subsection.

2.2.2 Global versus local association learning

Given $d \in [q]$ and association index space $\mathcal{K}$ (determined by d), to take the advantages of the subspace decomposition in Section 2.1, we parameterize $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}} =: \boldsymbol{H}\boldsymbol{\beta}, \tag{2.13}$$

with $\boldsymbol{H} = \{\boldsymbol{H}_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathcal{K}} \in \mathbb{R}^{|\boldsymbol{J}| \times \sum_{s=0}^{d} L_s}$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathcal{K}} \in \mathbb{R}^{\sum_{s=0}^{d} L_s \times p}$. Let $\boldsymbol{x}_{i(j)} \in \mathbb{R}^{p_j}$ be the jth subvector of $\boldsymbol{x}_i$, $j \in [t]$, where $\sum_{j=1}^{t} p_j = p$. Without loss of generality, we partition the matrix $\boldsymbol{\beta}_{\boldsymbol{k}} = [\boldsymbol{\beta}_{\boldsymbol{k},1}, \boldsymbol{\beta}_{\boldsymbol{k},2}, \ldots, \boldsymbol{\beta}_{\boldsymbol{k},t}]$ and vector $\boldsymbol{x}_i = (\boldsymbol{x}_{i(1)}^\top, \ldots, \boldsymbol{x}_{i(t)}^\top)^\top \in \mathbb{R}^p$ so that

$$\boldsymbol{\theta}\boldsymbol{x}_i = \sum_{\boldsymbol{k} \in \mathcal{K}} \sum_{j=1}^{t} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k},j} \boldsymbol{x}_{i(j)}, \qquad \boldsymbol{\beta}_{\boldsymbol{k},j} \in \mathbb{R}^{|\boldsymbol{k}|_{\boldsymbol{J}} \times p_j}.$$

As discussed in the previous section, if $\boldsymbol{\beta}_{\boldsymbol{k}} = 0$, then the corresponding association defined by $\boldsymbol{k}$ is not included in our model. Our predictor grouping structure allows us to perform association learning at distinct resolutions: *global association learning* or *local association learning* (i.e., predictor-wise association learning).

The goal of global association learning is to discover associations such that all predictors contribute to the association, or none contribute to the association. For global association learning, we take $t = 1$. To encourage sparsity in our fitted model so as to discover a small number of global associations, we use a group lasso-type penalty (Yuan and Lin, 2006) with a positive set of (user-specified) weights $\{w_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathcal{K}}$

for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively, as

$$\Omega_{\text{global}}(\boldsymbol{\beta}) = \sum_{\boldsymbol{k}\in\mathcal{K}} w_{\boldsymbol{k}} \left\|\boldsymbol{\beta}_{\boldsymbol{k}}\right\|, \tag{2.14}$$

$$\Phi_{\text{global}}(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}=\boldsymbol{H}\boldsymbol{\beta}} \Omega_{\text{global}}(\boldsymbol{\beta}) = \Omega_{\text{global}}(\boldsymbol{H}^\top\boldsymbol{\theta}). \tag{2.15}$$

Given that $\boldsymbol{\theta} = \boldsymbol{H}\boldsymbol{\beta}$ uniquely determines a $\boldsymbol{\beta} \in \mathbb{R}^{\sum_{s=0}^{d} L_s \times p}$, the infimum in (2.15) can be omitted. Because the Frobenius norm is nondifferentiable at the matrix of zeros, using Ω_{global} as a penalty can encourage estimates of the $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$ such that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}} = 0$ for many $\boldsymbol{k} \in \mathcal{K}$.

In local association learning, we relax the assumption that all predictors either contribute to an association, or no predictors contribute to an association. For example, when $q = 2$, it is possible that for the majority of predictors (but not all), a change in the predictor's value does not lead to a change in any of the local odd-ratios between response variables (i.e., these predictors only affect the marginal distributions of the response). This was exactly the type of association learning performed by Molstad and Rothman (2023). Our local association learning is much more general: we can discover which predictors modify certain high-order associations, and which predictors (or groups of predictors) only affect lower-order associations.

To achieve this type of learning, define the set $\mathcal{G}_{\text{local}} = \{(\boldsymbol{k}, j) : \boldsymbol{k} \in \mathcal{K}, j \in [p]\}$, let $\{w_{\boldsymbol{k},j}\}_{(\boldsymbol{k},j)\in\mathcal{G}_{\text{local}}}$ be a positive sequence, and define the penalty function

$$\Omega_{\text{local}}(\boldsymbol{\beta}) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}_{\text{local}}} w_{\boldsymbol{k},j} \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\|, \tag{2.16}$$

and similarly for Φ_{local}. In contrast to Ω_{global}, Ω_{local} has nondifferentiabilities when $\boldsymbol{\beta}_{\boldsymbol{k},j} = 0$ for any $(\boldsymbol{k}, j) \in \mathcal{G}_{\text{local}}$. As such, this penalty can encourage estimates such that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k},j} = 0$ for many $j \in [p]$, but if $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k},j'} \neq 0$ for any j', then the $\boldsymbol{k}$-joint association is included in the model.

Defining the set $\mathcal{G} = \{(\boldsymbol{k}, j) : \boldsymbol{k} \in \mathcal{K}, j \in [t]\}$, and defining

$$\Omega_{\mathcal{G}}(\boldsymbol{\beta}) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j} \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\|,$$

we generalize both global association learning ($t = 1$) and predictor-wide local association learning ($t = p$). More generally, we can perform a version of local association learning with predictors partitioned into t sets. This may be useful, for example, if predictors are categorical and encoded via multiple dummy variables.

2.2.3 Association learning with hierarchical constraints

As mentioned in Section 2.1.3, it is often desirable to enforce a hierarchical structure for the associations. To this end, we can modify both our global and local association structure learning penalties to enforce the hierarchy. Recall that for the hierarchy to be enforced, we must have that for every association $\boldsymbol{k}$ included in the model, all elements of $\mathcal{P}(\boldsymbol{k})$ must also be included in the model.

To achieve model fits of this type, we utilize the overlapping group lasso penalty. This penalty is defined by

$$\Omega^H_{\mathcal{G}}(\boldsymbol{\beta}) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j} \sqrt{\sum_{\boldsymbol{k}':\boldsymbol{k}\subset\boldsymbol{k}'} \left\|\boldsymbol{\beta}_{\boldsymbol{k}',j}\right\|^2} \tag{2.17}$$

and

$$\Phi^H_{\mathcal{G}}(\boldsymbol{\theta}) = \Omega^H_{\mathcal{G}}(\boldsymbol{H}^\top\boldsymbol{\theta}). \tag{2.18}$$

The term $\sqrt{\sum_{\boldsymbol{k}':\boldsymbol{k}\subset\boldsymbol{k}'} \left\|\boldsymbol{\beta}_{\boldsymbol{k}',j}\right\|^2}$ is a group lasso penalty on the entire set of coefficients corresponding to associations that include $\boldsymbol{k}$ in their powerset. For example, if $q = 3$ and $\boldsymbol{k} = \{1\}$, then

$$\sqrt{\sum_{\boldsymbol{k}':\boldsymbol{k}\subset\boldsymbol{k}'} \|\boldsymbol{\beta}_{\boldsymbol{k}',j}\|^2} = \sqrt{\|\boldsymbol{\beta}_{\{1\},j}\|^2 + \|\boldsymbol{\beta}_{\{1,2\},j}\|^2 + \|\boldsymbol{\beta}_{\{1,3\},j}\|^2 + \|\boldsymbol{\beta}_{\{1,2,3\},j}\|^2}.$$

Consequently, this penalty essentially precludes the possibility that $\widehat{\boldsymbol{\beta}}_{\{1,2\},j} \neq 0$ but $\widehat{\boldsymbol{\beta}}_{\{1\},j} = 0$, for example, because the penalty enforces $\widehat{\boldsymbol{\beta}}_{\{1\},j} = 0$ (via nondifferentiability at the origin) only when all higher order associations $\widehat{\boldsymbol{\beta}}_{\{1,2\},j} = \widehat{\boldsymbol{\beta}}_{\{1,3\},j} = \widehat{\boldsymbol{\beta}}_{\{1,2,3\},j} = 0$ as well. See Yan and Bien (2017) for a comprehensive review of how hierarchical structures can be enforced with the overlapping group lasso and related penalties.

2.3 Relation to existing work

2.3.1 Alternative parametric links

Multivariate categorical response regression is a classical problem in categorical data analysis (e.g., see Chapter 6 of McCullagh and Nelder, 1989). The majority of existing methods designed specifically for this task utilize parametric links between predictors and responses that can yield interpretable fitted models. To best describe these methods, we will first consider the case that $p = 1$ and $\boldsymbol{x}_i = 1$ for all $i \in [n]$ (i.e., the analysis of a q-way contingency table).

One particularly neat parametric link is the multivariate logistic transform. This transform maps probabilities $\boldsymbol{\pi} \in \mathbb{R}^{|\boldsymbol{J}|}$ to a set of parameters $\boldsymbol{\eta}$. These parameters represent the logarithms of the marginal odds, pairwise odds ratios, and higher-order odds ratios, which are derived from all possible joint marginals of subsets $Z_1, \ldots, Z_q$ (McCullagh and Nelder, 1989; Glonek, 1996; Molenberghs and Lesaffre, 1999). For a given $\boldsymbol{\pi}$, the transformation $\boldsymbol{\pi} \to \boldsymbol{\eta}$ can be expressed as the equation $\boldsymbol{\eta} = C\log(M\boldsymbol{\pi})$, where C is a contrast matrix, and M is a marginalizing matrix that computes the joint marginals from the cell probabilities. A more general class of log-linear models (where C and M are more general, and $\boldsymbol{\eta} = Z\boldsymbol{\eta}_z$ for design matrix Z), was proposed by Lang (1996). According to the definition of Bergsma and Rudas (2002), a numerical value assigned to $\boldsymbol{\eta}$ is considered *strongly compatible* if there exists a valid probability distribution $\boldsymbol{\pi}$ that corresponds to it. Palmgren (1989) showed that excluding the cases when $q = 2$ with $J_1 = J_2$, no explicit solution is available. Glonek and McCullagh (1995) pointed out the difficulty in solving $\boldsymbol{\eta} = C\log(M\boldsymbol{\pi})$ for the analysis of contingency tables, stating that "no readily computable criterion, for determining whether a particular $\boldsymbol{\eta}$ is valid, is available". If there are more than two categorical variables, it can happen that no solution exists because of incompatibility of the lower dimensional marginals. Evidently, it remains unclear how to determine whether a specific $\boldsymbol{\eta}$ is strongly compatible. For Bernoulli response $Z_1, \ldots, Z_q$, Qaqish and Ivanova (2006) can determine the strong compatibility of $\boldsymbol{\eta}$, and compute $\boldsymbol{\pi}$ from a strongly compatible $\boldsymbol{\eta}$ using a noniterative algorithm. When any $J_l \geq 3$, however, their results cannot be applied.

Matters become even more challenging when we consider the more general log-linear regression model $f(\boldsymbol{x}_i) = C\log\{M\boldsymbol{\pi}(\boldsymbol{x}_i)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ for linear function f. The goal of our work is to provide an alternative to log-linear models that (i) has parameters that can be interpreted in the same way as log-linear models and (ii) can be easily computed. Desideratum (i) is addressed by Theorem 2.1, and as we will show in a later section, because our estimator is the solution to a convex optimization problem, we can readily employ modern first order methods for (ii).

2.3.2 Generalizing log-linear models for contingency tables

In this section, we will explain how our method generalizes log-linear models used for the analysis of contingency tables. The key is that our method has the interpretability of "standard" log-linear models, but our specific subspace decomposition leads to an invariance property that is essential for penalized maximum likelihood-based association learning.

Log-linear models are a class of statistical models used to describe the relationship between categorical variables by modeling the expected cell counts in a contingency table. These models express the logarithm of expected frequencies as a linear combination of parameters corresponding to main associations and interactions of the variables. Specifically, for a contingency table (i.e., the intercept only model with $p = 1$) with variables Z_1 and Z_2, the model can be written as

$$\log(\mu_{j_1 j_2}) = \mathbf{\Lambda}_{j_1,j_2} := \mu + \mu_{j_1}^{Z_1} + \mu_{j_2}^{Z_2} + \mu_{j_1 j_2}^{Z_1 Z_2} \tag{2.19}$$

where $\mu_{j_1 j_2}$ denotes the expected count in cell (j_1, j_2), μ is the overall mean, $\mu_{j_1}^{Z_1}$ and $\mu_{j_2}^{Z_2}$ represent the main associations of variables Z_1 and Z_2, respectively, and $\mu_{j_1 j_2}^{Z_1 Z_2}$ denotes the interaction association between Z_1 and Z_2. Under a multinomial sampling scheme, the model can be written as

$$\pi_{j_1 j_2} = \frac{\exp(\mathbf{\Lambda}_{j_1,j_2})}{\sum_{j_1,j_2} \exp(\mathbf{\Lambda}_{j_1,j_2})}. \tag{2.20}$$

The log-linear model and the multinomial model share the same linear structure of $\mathbf{\Lambda}_{j_1,j_2}$.

To ensure the parameters in a log-linear model are uniquely estimable, certain constraints must be imposed. Commonly, sum-to-zero constraints are used, where the sum of the main associations and interaction associations for each variable is set to zero. For example, for the main associations, the constraints are: $\sum_{j_1=1}^{J_1} \mu_{j_1}^{Z_1} = 0$

and $\sum_{j_2=1}^{J_2} \mu_{j_2}^{Z_2} = 0$. Similarly, for the interaction associations:

$$\sum_{j_1} \mu_{j_1 j_2}^{Z_1 Z_2} = 0 \quad \text{for each } j_2 \quad \text{and} \quad \sum_{j_2} \mu_{j_1 j_2}^{Z_1 Z_2} = 0 \quad \text{for each } j_1.$$

Alternatively, one could define $\mu_1^{Z_1} = 0, \mu_1^{Z_2} = 0$ and $\mu_{j_1 j_2}^{Z_1 Z_2} = 0$ if $j_1 = 1$ or $j_2 = 1$. For maximum likelihood estimation (without penalization), the choice of constraint does not matter due to the invariance property of the maximum likelihood estimator. If, on the other hand, one wanted to impose sparsity inducing penalties on the μ, the choice of constraint may affect the solution.

To see this, recall that $\boldsymbol{\mu}^{\boldsymbol{J}} = \{\mu_{j_1,j_2}\}_{(j_1,j_2)\in \boldsymbol{J}}$ for log-linear model. Let $\boldsymbol{U}'_m = [\boldsymbol{e}_2^m, \ldots, \boldsymbol{e}_m^m]$. Similar to $\boldsymbol{H}_{\boldsymbol{k}}$ defined in (2.8), for any $\boldsymbol{k} = \{k_1, \ldots, k_s\} \in \mathcal{K}_s, s \geq 1$, define

$$\boldsymbol{H}'_{\boldsymbol{k}} = \boldsymbol{V}_q \otimes \boldsymbol{V}_{q-1} \otimes \cdots \otimes \boldsymbol{V}_2 \otimes \boldsymbol{V}_1, \qquad \boldsymbol{V}_i = \begin{cases} \boldsymbol{U}'_{J_i} & i \in \boldsymbol{k} \\ \boldsymbol{1}_{J_i} & i \in [q]\backslash \boldsymbol{k} \end{cases}. \tag{2.21}$$

We can thus rewrite (2.19) in matrix form as

$$\mathrm{vec}_{\boldsymbol{J}}\{\log(\boldsymbol{\mu}^{\boldsymbol{J}})\} = \boldsymbol{H}'_{\{0\}}\boldsymbol{\beta}'_{\{0\}} + \boldsymbol{H}'_{\{1\}}\boldsymbol{\beta}'_{\{1\}} + \boldsymbol{H}'_{\{2\}}\boldsymbol{\beta}'_{\{2\}} + \boldsymbol{H}'_{\{1,2\}}\boldsymbol{\beta}'_{\{1,2\}},$$

where $\{\boldsymbol{H}'_{\boldsymbol{k}}\}_{\boldsymbol{k}}$ are defined in (2.21) with $q = 2$. Similarly, recall that $\boldsymbol{\pi}^{\boldsymbol{J}} = \{\pi_{j_1,j_2}\}_{(j_1,j_2)}$ for the multinomial log-linear model so that

$$\mathrm{vec}_{\boldsymbol{J}}\{\boldsymbol{\pi}^{\boldsymbol{J}}\} = \frac{\exp(\boldsymbol{\theta})}{\langle \exp(\boldsymbol{\theta}), \boldsymbol{1}_{J_1 J_2}\rangle}, \quad \boldsymbol{\theta} = \boldsymbol{H}'_{\{0\}}\boldsymbol{\beta}'_{\{0\}} + \boldsymbol{H}'_{\{1\}}\boldsymbol{\beta}'_{\{1\}} + \boldsymbol{H}'_{\{2\}}\boldsymbol{\beta}'_{\{2\}} + \boldsymbol{H}'_{\{1,2\}}\boldsymbol{\beta}'_{\{1,2\}}.$$

Here, $\boldsymbol{H}'_{\{0\}}\boldsymbol{\beta}'_{\{0\}}, \boldsymbol{H}'_{\{1\}}\boldsymbol{\beta}'_{\{1\}}, \boldsymbol{H}'_{\{2\}}\boldsymbol{\beta}'_{\{2\}}, \boldsymbol{H}'_{\{1,2\}}\boldsymbol{\beta}'_{\{1,2\}}$ are the matrix forms of μ, $\mu_{j_1}^{Z_1}$, $\mu_{j_2}^{Z_2}$, and $\mu_{j_1 j_2}^{Z_1 Z_2}$, respectively for both log-linear model and multinomial model. Evidently the log-linear model can be parameterized as $\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}'_{\boldsymbol{k}}\boldsymbol{\beta}'_{\boldsymbol{k}}$. If we wanted to impose sparsity on the $\boldsymbol{\beta}'$, it would be tempting to use the same group lasso penalty as defined before,

However, when considering $\Phi'(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}'_{\boldsymbol{k}}\|$, we see that $\Phi'(\cdot)$ is not invariant to the choice of identifability constraints. To be more specific, if $\max_{i\in[q]} J_i > 2$, $\boldsymbol{U}''_m = [\boldsymbol{e}^m_1, \ldots, \boldsymbol{e}^m_{m-1}]$, and we define $\boldsymbol{H}''$ accordingly, then $\boldsymbol{H}'_{\boldsymbol{k}}\boldsymbol{\beta}'_{\boldsymbol{k}} \not\equiv \boldsymbol{H}''_{\boldsymbol{k}}\boldsymbol{\beta}''_{\boldsymbol{k}}, \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}'_{\boldsymbol{k}}\| \not\equiv \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}''_{\boldsymbol{k}}\|$, where $\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}'_{\boldsymbol{k}}\boldsymbol{\beta}'_{\boldsymbol{k}} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}''_{\boldsymbol{k}}\boldsymbol{\beta}''_{\boldsymbol{k}}$. Choosing $\boldsymbol{H}''$ instead of $\boldsymbol{H}'$ changes how the $\boldsymbol{k}$-joint association influences the categorical response, leading to results that may depend on this arbitrary selection rather than reflecting an inherent property.

To address the invariance issue, one might consider using an overparameterized version of the log-linear model with penalization of the parameters. However, this leads to an explosion in the number of parameters, and the parameters are more difficult to interpret. Moreover, statistical analysis of such an estimator is fundamentally more difficult than the analysis of our estimator.

In our reparameterization $\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}$, the corresponding group lasso penalty $\Phi(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}_{\boldsymbol{k}}\|$ is invariant under different choices of $\boldsymbol{U}_m$ such that $[\frac{1}{\sqrt{m}}\mathbf{1}_m, \boldsymbol{U}_m]$ is a real orthogonal matrix. To be more specific, if we let $\mathcal{U}_m$ be another real matrix such that $[\frac{1}{\sqrt{m}}\mathbf{1}_m, \mathcal{U}_m]$ is a real orthogonal matrix, and define $\boldsymbol{H}^{\mathcal{U}}_{\boldsymbol{k}}$ by replacing $\boldsymbol{U}_{J_i}$ with $\mathcal{U}_{J_i}$ in (2.8), then

$$\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} \equiv \boldsymbol{H}^{\mathcal{U}}_{\boldsymbol{k}}\boldsymbol{\beta}^{\mathcal{U}}_{\boldsymbol{k}}, \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}_{\boldsymbol{k}}\| \equiv \sum_{\boldsymbol{k}} \left\|\boldsymbol{\beta}^{\mathcal{U}}_{\boldsymbol{k}}\right\|, \quad \text{where } \boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}^{\mathcal{U}}_{\boldsymbol{k}}\boldsymbol{\beta}^{\mathcal{U}}_{\boldsymbol{k}}.$$

2.3.3 Modern approaches to multivariate categorical response regression in high dimensions

Existing methods for multivariate categorical response regression with a large number of predictors, responses, and/or a large number categories per response typically rely on latent variable models (e.g., the regularized latent class model of Molstad and Zhang, 2022), or classifier chains (Read et al., 2021).

The latent class model is able to capture complex relationships between responses by assuming that given a latent variable W, Z_m and $Z_{m'}$ are independent given $\boldsymbol{X}$, i.e., $Z_m \perp\!\!\!\perp Z_{m'} \mid \boldsymbol{X}, \boldsymbol{W}$. Thus, fitted model coefficients cannot be straightforwardly interpreted in terms of the distribution of interest $Z_1, \dots, Z_q \mid \boldsymbol{X}$, as can the coefficients from our fitted model. Moreover, the order of associations in the latent class method cannot, generally speaking, be easily identified unless the association is null.

Along similar lines, it is common to decompose the joint mass function of interest into simpler, estimable parts. Methods utilizing to this approach include those most popular in the machine learning literature on "multilabel classification" (Herrera et al., 2016), namely, classifier chains (Read et al., 2021). A classifier chain estimates $Z_1, \dots, Z_q \mid \boldsymbol{X}$ by fitting a model for $Z_1 \mid \boldsymbol{X}$, then $Z_2 \mid \boldsymbol{X}, Z_1$, then $Z_3 \mid \boldsymbol{X}, Z_1, Z_2$, and so on, and using their product as an estimate of the mass function of interest. This approach requires many ad-hoc decisions that can have a significant impact on how the model performs (e.g., in what order to fit the chain and how to model each specific conditional distribution). Like the latent class model approach, classifier chains cannot be used to identify the order of associations in a straightforward way, which is the primary motivation for our work.

Recently, to handle large p, Deng et al. (2024) proposed to model multivariate categorical responses under a multivariate generalization of the normal linear discriminant analysis model. Though their method can handle large p, it is specifically designed for a bivariate response, and does not provide the interpretability of our method.

2.3.4 Connection with Ising, Potts, and Hypergraph Models

In the classical Ising model, each vertex of a graph is assigned a binary state $j_i \in \{0, 1\}$, with interactions limited to pairs of vertices connected by an edge. Cheng et al. (2014) extend the Ising model allow interactions to depend on covariates. They propose a sparse pseudo-likelihood estimation framework tailored to the binary Ising model, which does not extend to multi-level categorical variables. In contrast, in the Potts model, each vertex is assigned a variable $j_i \in \{1, 2, \dots, J_i\}$, with pairwise interactions specified by edges (Shimagaki and Weigt, 2019; Wipf et al., 2007; Tanaka et al., 2011; Hirakida et al., 2016). Such models capture the essential physics of systems whose energy depends exclusively on pairwise interactions. Broadly speaking, maximum pseudo-likelihood estimation is widely used under both Ising and generalized Potts models without predictors (Razaee and Amini, 2020; Ekeberg et al., 2013; Okabayashi et al., 2011; Cheng et al., 2014; Keetelaar et al., 2024; Levada et al., 2009). However, as illustrated in Keetelaar et al. (2024), there is a key trade-off to be considered: while pseudo-likelihood is computationally tractable and approximates MLE well with sufficiently large sample sizes, it becomes less reliable under with small or moderate n. In contrast, our approach relies on the complete likelihood.

To the best of our knowledge, no existing study has introduced covariates into the Potts model. As we will show below, the generalized Potts model—already a special case of a hypergraph model—also emerges from our approach when the maximum interaction order is set to 2. Our method thus advances Potts modeling by incorporating covariates, extending them beyond the purely pairwise framework.

Many real-world systems exhibit interactions among more than two components simultaneously. For example, in certain magnetic materials or biological networks, the collective behavior depends on multi-body interactions. To model such scenarios, it is natural to generalize the notion of a graph to that of a (undirected) hypergraph (Bretto, 2013; Klamt et al., 2009; Yu et al., 2024; Robiglio et al., 2024; Mukherjee et al., 2022; Bhowal and Mukherjee, 2024). In Section 4.3, we demonstrate that both the generalized Potts model and the hypergraph model coincide with our proposed multivariate categorical response regression model—under the reparameterization $\boldsymbol{\theta} = \sum_{\boldsymbol{k}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}$—when there are no predictors.

2.4 Computation

In this section, we outline a proximal gradient descent algorithm—described in Chapter 4 of Parikh et al. (2014)—for computing the group lasso and the overlapping group lasso-penalized estimators.

The proximal gradient descent algorithm can be understood from the perspective of the majorize-minimize principle (Lange, 2016). If there exists some $L > 0$ such that for arbitrary kth iterate $\boldsymbol{\beta}^k$,

$$\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega(\boldsymbol{\beta}) \leq \mathcal{L}_n(\boldsymbol{\beta}^k) + \left\langle \nabla\mathcal{L}_n(\boldsymbol{\beta}^k), \boldsymbol{\beta} - \boldsymbol{\beta}^k \right\rangle + \frac{L}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^k\right\|^2 + \lambda\Omega(\boldsymbol{\beta}) \quad (2.22)$$

for all $\boldsymbol{\beta}$, then, if we define the $(k+1)$th iterate as

$$\boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2}\left\|\boldsymbol{\beta} - \left\{\boldsymbol{\beta}^k - \frac{1}{L}\nabla\mathcal{L}_n(\boldsymbol{\beta}^k)\right\}\right\|^2 + \frac{\lambda}{L}\Omega(\boldsymbol{\beta})\right], \quad (2.23)$$

we are ensured that the objective function at $\boldsymbol{\beta}^{k+1}$ is no greater than the objective function at $\boldsymbol{\beta}^k$ (i.e., the sequence of iterates $\{\boldsymbol{\beta}^k\}_{k=1}^{\infty}$ have the descent property). When Ω is the group lasso penalty (i.e., $\Omega_{\mathcal{G}}$), then the proximal operator (2.23) has closed form solution

$$\boldsymbol{\beta}_g^{k+1} = \max\left(1 - \frac{\lambda w_g}{\left\|L \cdot \boldsymbol{\beta}_g^k - \frac{\partial}{\partial \boldsymbol{\beta}_g}\mathcal{L}_n(\boldsymbol{\beta}^k)\right\|}, 0\right)\left(\boldsymbol{\beta}_g^k - \frac{1}{L}\frac{\partial}{\partial \boldsymbol{\beta}_g}\mathcal{L}_n(\boldsymbol{\beta}^k)\right), \quad g \in \mathcal{G}.$$

When Ω is the overlapping group lasso penalty (i.e., $\Omega_{\mathcal{G}}^{H}$), we solve the proximal operator (2.23) using a blockwise coordinate algorithm (Jenatton et al., 2011; Yan and Bien, 2017).

In Lemma 4.7, we show that for all $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$, $\left\|\nabla\mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}) - \nabla\mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}')\right\| \leq (2n)^{-1}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X})\,\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|$ with $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$, which implies that with $L \geq (2n)^{-1}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X})$, (2.22) will hold. However, in the case of a Poisson categorical response model, the inequality (2.22) cannot hold globally for any L. Therefore, we use a proximal gradient descent algorithm with the step size determined adaptively at each iteration by a backtracking line search.

For the sake of space, more details about tuning parameter selection, as well as the formulation of an accelerated variation of the proximal gradient descent algorithm, can be found in Section 4.4. More details about the accelerated proximal gradient descent algorithm can be found in Section 4.3 of Parikh et al. (2014) and Algorithm 2 of Tseng (2008). We present our algorithm and all needed sub-algorithms in Section 4.4.

2.5 Statistical properties

In this section, we examine the statistical properties of the group lasso estimator, as defined in (2.12), considering variations in n, p, and $\boldsymbol{J}$. Let $\boldsymbol{\theta}^* = \boldsymbol{H}_{\text{full}}\boldsymbol{\beta}^*$ represent the data generation parameter, where $\boldsymbol{H}_{\text{full}} = \{\boldsymbol{H}_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\cup_{s=0}^{q}\mathcal{K}_s} \in \mathbb{R}^{|\boldsymbol{J}|\times|\boldsymbol{J}|}$. To establish an error bound, it is necessary to define an identifiable estimand: the parameter $\boldsymbol{\beta}^\dagger$. Let the set $\mathcal{F}_{\boldsymbol{\theta}^*}$ denote the set of all $\boldsymbol{\beta}$, which leads to the same probability distribution, that is for multinomial and Poisson categorical response models,

$$\begin{aligned}\mathcal{F}_{\boldsymbol{\theta}^*} =& \left\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{|\boldsymbol{J}|\times p}; \ell(\boldsymbol{\theta}^*\boldsymbol{x}, \boldsymbol{y}) = \ell(\boldsymbol{H}_{\text{full}}\widetilde{\boldsymbol{\beta}}\boldsymbol{x}, \boldsymbol{y}), \forall(\boldsymbol{x}, \boldsymbol{y})\right\} \\ =& \left\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{|\boldsymbol{J}|\times p}; \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}^* = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\text{full}}\widetilde{\boldsymbol{\beta}}\right\},\end{aligned} \tag{2.24}$$

where $\boldsymbol{P}_{\mathcal{V}} = I - |\boldsymbol{J}|^{-1}\mathbf{1}_{|\boldsymbol{J}|}\mathbf{1}_{|\boldsymbol{J}|}^\top$ for multinomial categorical response model, and $\boldsymbol{P}_{\mathcal{V}} = I$ for Poisson categorical response model.

Define $\boldsymbol{\theta}^\dagger = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^\dagger = \boldsymbol{H}_{\text{full}}^\top\boldsymbol{\theta}^\dagger$. By Lemma 4.9, we know that

$$\boldsymbol{\beta}^\dagger \in \underset{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}}{\arg\min}\{\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta})\} = \underset{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}}{\arg\min}\,\Omega_{\mathcal{G}}(\boldsymbol{\beta}). \tag{2.25}$$

Now, we introduce our assumptions. The first is a standard scaling assumption on the predictors.

Assumption 2.1 (Predictor scaling)**.** The predictors are scaled so that for any $i \in [n], j \in [t]$, and $\|\boldsymbol{k}\|_0 \le d$, $\left\|\boldsymbol{x}_{i(j)}\right\| \le w_{\boldsymbol{k},j}C$ for finite constant C.

The following assumption regards the data generating process.

Assumption 2.2. The responses $\boldsymbol{y}_i^{\boldsymbol{J}} = \{y_{\boldsymbol{j}}^i\}_{\boldsymbol{j}\in\boldsymbol{J}}, 1 \le i \le n$ are independent given $\{\boldsymbol{x}_i\}_{i=1}^n$ and generated under (i) the Poisson categorical response model or (ii) the multinomial categorical response model with ($n_i = 1$ for $i \in [n]$, without loss of generality).

Assumption 2.3 (Poisson categorical response model)**.** Under (i), the Poisson categorical response model with $\mathcal{V} = \mathbb{R}^{|\boldsymbol{J}|}$, there exists a finite constant C_1 such that $\Lambda := \max_{i\in[n]} \|\exp(\boldsymbol{\theta}^\dagger\boldsymbol{x}_i)\|_\infty \le C_1$.

Note that under (ii), the multinomial categorical response model, $\mathcal{V} = \{\boldsymbol{\theta}' \in \mathbb{R}^{|\boldsymbol{J}|} : \mathbf{1}_{|\boldsymbol{J}|}^\top\boldsymbol{\theta}' = \mathbf{0}\}$. This is not an assumption, but rather a definition.

Next, we make an assumption on the curvature of the negative log-likelihood in certain directions: this is commonly known as restricted strong convexity (Wainwright, 2019, Definition 9.15 and Theorem 9.36). Let $\mathcal{E}_n(\Delta\boldsymbol{\theta}) := L_n(\boldsymbol{\theta}^\dagger + \Delta\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}^\dagger) - \left\langle\nabla L_n(\boldsymbol{\theta}^\dagger), \Delta\boldsymbol{\theta}\right\rangle$.

Assumption 2.4 (Restricted strong convexity)**.** Let $\Phi_{\mathcal{G}}(\boldsymbol{\theta}) = \Omega_{\mathcal{G}}(\boldsymbol{H}^\top \boldsymbol{\theta})$ be the reparameterized group lasso penalty for the association learning. The quantity $\mathcal{E}_n(\Delta\boldsymbol{\theta})$ satisfies restricted strong convexity (RSC) condition with radius $R > 0$, constants A and C, and curvature $\kappa > 0$, i.e., $\Delta\boldsymbol{\theta} \in \mathcal{M}_H := \{\boldsymbol{\theta} : \boldsymbol{\theta} = \sum_{g \in \mathcal{G}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_g\}$,

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) \geq \frac{\kappa}{2} \|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\|^2 - A \cdot C^2 \Big(\frac{\log|\mathcal{G}|}{n} + \frac{m}{n}\Big) \cdot \inf_{\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta} = \boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}'} \Phi^2_{\mathcal{G}}(\Delta\boldsymbol{\theta}'), \quad \|\Delta\boldsymbol{\theta}\| \leq R, \tag{2.26}$$

where $|\mathcal{G}|$ is the cardinality of $\mathcal{G}$, and $m = \max_{(\boldsymbol{k},j)\in\mathcal{G}} |\boldsymbol{k}|_{\boldsymbol{J}} \cdot p_j$. Under (i), the Poisson categorical response model, $\boldsymbol{P}_{\mathcal{V}} = I$, and denote $\kappa = \kappa^{Pois}_{\boldsymbol{J}}$, whereas under (ii), the multinomial categorical response model, $\boldsymbol{P}_{\mathcal{V}} = I - |\boldsymbol{J}|^{-1}\mathbf{1}_{|\boldsymbol{J}|}\mathbf{1}^\top_{|\boldsymbol{J}|}$, and denote $\kappa = \kappa^{Mult}_{\boldsymbol{J}}$.

The restricted strong convexity condition is a well-understood condition in penalized regression (Negahban et al., 2012, Section 2.4). Effectively, this condition requires that in a neighborhood of the true parameter, the negative log-likelihood has sufficient curvature.

Remark 2.1. In Lemma 4.11, we verify that under mild assumptions on the distribution of predictors, restricted strong convexity holds with high probability for (i) Poisson and (ii) multinomial categorical response models.

Define the support of $\boldsymbol{\beta}^\dagger$ as $\mathcal{S} = \{g \in \mathcal{G}; \boldsymbol{\beta}^\dagger_g \neq 0\}$ and define $\Psi(\mathcal{S})^2 = \sum_{(\boldsymbol{k},j)\in\mathcal{S}} w^2_{\boldsymbol{k},j}$. Clearly, if $w_g = 1$ for all $g \in \mathcal{G}$, then $\Psi(\mathcal{S})^2 = |\mathcal{S}|$. Note that $\Psi(\mathcal{S})$ is essentially the subspace compatibility constant (Wainwright, 2019, Definition 9.18): a quantity that often appears in error bounds for regularized M-estimators.

Note that the dimensionality of $\widehat{\boldsymbol{\beta}}$ depends on the user-specified d, whereas $\boldsymbol{\beta}^\dagger \in \mathbb{R}^{|\boldsymbol{J}|\times p}$. Thus, to simplify notation, let $\widehat{\boldsymbol{\beta}}_0$ denote the version of $\widehat{\boldsymbol{\beta}}$ where all associations of order higher than d have been set to zero (i.e., $\widehat{\boldsymbol{\beta}}_0^\top = [\widehat{\boldsymbol{\beta}}^\top, \mathbf{0}] \in \mathbb{R}^{p\times|\boldsymbol{J}|}$). We are now prepared to present our error bound for $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\dagger\| = \|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^\dagger\|$. Recall that

$\mathcal{K} = \cup_{s=0}^{d} \mathcal{K}_s$, where d denotes the maximal number of association between response variables. Define the true maximal number of association as $d^* = \{\|\boldsymbol{k}\|_0 \,; (\boldsymbol{k}, j) \in \mathcal{S}, \boldsymbol{\beta}_{\boldsymbol{k}}^{\dagger} \neq \mathbf{0}\}$.

Theorem 2.2. *Let B, B_1, B_2 and B' be positive absolute constants, and let $\xi \geq 1$ be fixed. Suppose that d is chosen so that $d^* \leq d$ and that Assumptions 2.1-2.4 hold.*

(i) Under the Poisson categorical response model, if

$$\lambda = \xi BC(\sqrt{\Lambda m/n} + \sqrt{\Lambda \log |\mathcal{G}|/n})$$

with $0 \leq (\xi - 1)(\sqrt{m/n} + \sqrt{\log |\mathcal{G}|/n}) \leq B_2$, $\lambda \leq R\kappa_{\boldsymbol{J}}^{Pois}\{6\sqrt{|\mathcal{S}|}\}^{-1}$, *and* $(m/n + \log |\mathcal{G}|/n) \leq B_1 \cdot \min\{1, \kappa_{\boldsymbol{J}}^{Pois}(AC^2|\mathcal{S}|)^{-1}\}$, *then*

$$\|\boldsymbol{\theta}^{\dagger} - \widehat{\boldsymbol{\theta}}\| = \|\boldsymbol{\beta}^{\dagger} - \widehat{\boldsymbol{\beta}}_0\| \leq \frac{6\xi BC\sqrt{|\mathcal{S}|}}{\kappa_{\boldsymbol{J}}^{Pois}} \left(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log |\mathcal{G}|}{n}}\right),$$

with probability at least $1 - \exp(-B'(\xi - 1)^2(m + \log |\mathcal{G}|))$.

(ii) Under the multinomial categorical response model, if

$$\lambda = \xi BC(\sqrt{m/n} + \sqrt{\log |\mathcal{G}|/n}),$$

$\lambda \leq R\kappa_{\boldsymbol{J}}^{Mult}\{6\sqrt{|\mathcal{S}|}\}^{-1}$, *and* $(m/n + \log |\mathcal{G}|/n) \leq B_1 \cdot \min\{1, \kappa_{\boldsymbol{J}}^{Mult}(AC^2|\mathcal{S}|)^{-1}\}$, *then*

$$\|\boldsymbol{\theta}^{\dagger} - \widehat{\boldsymbol{\theta}}\| = \|\boldsymbol{\beta}^{\dagger} - \widehat{\boldsymbol{\beta}}_0\| \leq \frac{6\xi BC\sqrt{|\mathcal{S}|}}{\kappa_{\boldsymbol{J}}^{Mult}} \left(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log |\mathcal{G}|}{n}}\right),$$

with probability at least $1 - \exp(-B'(\xi - 1)^2(m + \log |\mathcal{G}|))$.

In Lemma 2.3, we show that, under certain regularity assumptions, $\kappa_{\boldsymbol{J}}^{Mult} \asymp |\boldsymbol{J}|^{-1}\kappa_{\boldsymbol{J}}^{Pois}$ and $\kappa_{\boldsymbol{J}}^{Pois} = O(1)$. However, we cannot conclude that the Poisson-likelihood based estimator is better its multinomial counterpart because the data generating models being assumed are fundamentally different.

The result of Theorem 2.2 indicates that under the Poisson or multinomial sampling scheme, assuming $\kappa_{\boldsymbol{J}}^{Pois} = O(1)$ or $\kappa_{\boldsymbol{J}}^{Mult} = (|\boldsymbol{J}|^{-1})$, we can achieve a Frobenius norm error rate of $O(\sqrt{m/n} + \sqrt{\log |\mathcal{G}|/n})$. Call that $|\mathcal{G}|$ is the number of groups of parameters being penalized in (2.12) under general local association learning. This would seem to suggest that having fewer groups is beneficial, but this term is counterbalanced with m, which is the largest number of parameters per group. Hence, since a small number of groups would require a larger number of parameters per group, there is a clear tradeoff between the two. Importantly, both terms are multiplied by $|\mathcal{S}|$, so ideally, we will select a number of groups that leads to small $\mathcal{S}$ without inflating $|\mathcal{G}|$ or m.

Though not made explicit in our bounds, the association of a well-specified d is apparent in our error bounds. If $d = q \gg d^*$, then both m and $|\mathcal{G}|$ will be larger than if d were specified closer to d^*. Of course, if $d < d^*$, we could not expect consistent estimation since this will force estimates of truly nonzero associations to be zero.

The following corollary is a special case of Theorem 2.2 for multinomial categorical response model, letting $\mathcal{G} = \mathcal{G}_{\text{global}}$ or $\mathcal{G}_{\text{local}}$. Here, we replace the quantities from Theorem 2.2 with more explicit versions.

Corollary 2.1. *Under the conditions of Theorem 2.2 and assuming the multinomial categorical response model, if tuning parameters are chosen according to part (ii) of Theorem 2.2 and $w_g = 1$ for all $g \in \mathcal{G}$, then the following statements holds.*

1. *For global association learning, with probability as specified in Theorem 2.2,*

$$\|\boldsymbol{\theta}^\dagger - \widehat{\boldsymbol{\theta}}\| \leq \frac{6\xi BC}{\kappa_{\boldsymbol{J}}^{Mult}} \sqrt{\sum_{\boldsymbol{k}\in\mathcal{K}} \mathbf{1}(\boldsymbol{\beta}_{\boldsymbol{k}}^\dagger \neq 0)} \left\{ \sqrt{\frac{p\prod_{\ell=1}^d (J_{(\ell)} - 1)}{n}} + \sqrt{\frac{\log \sum_{l=0}^d \binom{q}{l}}{n}} \right\},$$

where $J_{(1)}, \cdots, J_{(q)}$ *is a permutation of* $J_1, \cdots, J_q$ *such that* $J_{(1)} \geq J_{(2)} \geq \cdots \geq J_{(q)}$.

2. *For local association learning* ($t \geq 2$)*, with probability as specified in Theorem 2.2,*

$$\|\boldsymbol{\theta}^{\dagger} - \widehat{\boldsymbol{\theta}}\| \leq \frac{6\xi BC}{\kappa_{\boldsymbol{J}}^{Mult}} \left\|\boldsymbol{\beta}^{\dagger}\right\|_{0,\mathcal{G}} \left\{ \sqrt{\frac{\max_{(\boldsymbol{k},j)\in\mathcal{G}} |\boldsymbol{k}|_{\boldsymbol{J}} \cdot p_j}{n}} + \sqrt{\frac{\log t + \log \sum_{l=0}^{d} \binom{q}{l}}{n}} \right\},$$

where $\left\|\boldsymbol{\beta}^{\dagger}\right\|_{0,\mathcal{G}} = \sqrt{\sum_{(\boldsymbol{k},j)\in\mathcal{G}} \mathbf{1}(\boldsymbol{\beta}^{\dagger}_{\boldsymbol{k},j} \neq 0)}$.

For the multinomial sampling scheme, as $|\boldsymbol{J}|$ increases, the upper bound of the estimation error worsens. This suggests that increasing the dimension of the response will lead to poorer estimation.

We continue by demonstrating the reasonableness of Assumption 2.4, particularly regarding its validity under the assumption of random predictors. In section 9 of Wainwright (2019), the restricted strong convexity condition has been derived under a GLM setting (See Theorem 9.36 in Wainwright (2019)). Here, we generalized their results to a multivariate GLM setting, and calibrate the Rademacher complexity term of the group lasso penalty according to multivariate GLM setting. We summarize the results in Section 4.9 and incorporate both the multinomial and Poisson categorical response settings into the following lemma.

Lemma 2.3. *Under Assumptions 2.1–2.3, assume that* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are independent and identically distributed with zero mean. Additionally, assume that for some positive constants* (α, β)*, we have* $\mathbb{E}\left\|\Delta\boldsymbol{\theta}\boldsymbol{x}_i\right\|^2 \geq \alpha$ *and* $\mathbb{E}\left\|\Delta\boldsymbol{\theta}\boldsymbol{x}_i\right\|^4 \leq \beta$*, for all vectors* $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$ *such that* $\|\Delta\boldsymbol{\theta}\| = 1$*. Then, the following results hold. For both multinomial and Poisson categorical response models with the reparameterized group lasso penalty* $\Phi_{\mathcal{G}}$*, the restricted strong convexity condition (2.26) in Assumption 2.4* *holds with probability at least* $1 - c_1 \exp(-c_2 n)$. *Furthermore,* $\kappa_{\boldsymbol{J}}^{Mult} \asymp |\boldsymbol{J}|^{-1} \kappa_{\boldsymbol{J}}^{Pois}$ *and* $\kappa_{\boldsymbol{J}}^{Pois} = O(1)$.

The above lemma justifies condition (2.26) and the typical behavior of the curvature κ in Assumption 2.4, showing that both will hold with high probability under mild assumptions.

2.6 Numerical studies

2.6.1 Data generating models and competitors

We present a series of simulations designed to evaluate the performance of the proposed methods and applicable variations of existing methods under various scenarios. We consider a range of parameters, including different sample sizes, dimensions, and three different model generation schemes. A detailed description of this study is in Section 4.5.

For $N_{\text{rep}} = 100$ independent replications, we simulate data from the multivariate multinomial response regression model, with $d = 4$, $q = 4$, and $(J_1, J_2, J_3, J_4) = (2, 2, 2, 3)$. With $n \in \{100, 300, 500, 1000, 2000\}$ training samples, each observation $\boldsymbol{x}_i$ (excluding its first element, because the first row of $\boldsymbol{x}_i$ is always 1) is drawn from a multivariate normal distribution $\mathrm{N}_p(\mathbf{0}, \Sigma)$, where the covariance entries $\Sigma_{jk} = 0.5^{|j-k|}$ are defined for all pairs $(j,k) \in [p-1] \times [p-1]$. Given a coefficient matrix $\boldsymbol{\beta}^* \in \mathbb{R}^{|\boldsymbol{J}| \times p}$, the probability vector is given by $\mathrm{vec}_{\boldsymbol{J}}\left(\boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x})\right) = \dfrac{\exp(\boldsymbol{H}\boldsymbol{\beta}^*\boldsymbol{x})}{\langle \mathbf{1}_{|\boldsymbol{J}|}, \exp(\boldsymbol{H}\boldsymbol{\beta}^*\boldsymbol{x})\rangle}$, from which we generate the response $\boldsymbol{y}_i \in \mathbb{R}^p$ as a realization of

$$\text{Multinomial}\Big(n_i, \mathrm{vec}_{\boldsymbol{J}}\left(\boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x}_i)\right)\Big), \quad n_i = 1. \tag{2.27}$$

This process is also extended to generate 1000 validation samples for model tuning and $N_{\text{test}} = 10000$ test samples to evaluate model performance. We conduct our simulations over a range of dimensions $p \in \{10, 50\}$ to assess scalability and robustness. Let $\mathcal{G}_{\text{global}} = \{(\boldsymbol{k}, j); \boldsymbol{k} \in \mathcal{K}, j \in \{1, 2\}\}$ with $p_1 = 1, p_2 = p - 1$ and $\mathcal{G}_{\text{local}} = \{(\boldsymbol{k}, j); \boldsymbol{k} \in \mathcal{K}, j \in \{1, \cdots, p\}\}$ with $p_1 = p_2 = \cdots = p_p = 1$.

We consider three distinct structures for $\boldsymbol{\beta}^*$. The parameter generation methods for $\boldsymbol{\beta}^*$ are designated as Scheme 1, Scheme 2, and Scheme 3, and are discussed in detail in Section 4.5. These correspond to the three interpretable models—mutual independence, joint independence, and conditional independence, respectively—as presented in Example 2.2. Moreover, in each scheme, the associations are local—only the intercept and two randomly selected predictors have nonzero associations. Hence, these schemes are most well suited for the versions of our methods using Ω^H_{local}.

In our simulation studies, we will examine the following eight estimators. The first six estimators—`O-Mult`, `O-Pois`, `L-Mult`, `L-Pois`, `G-Mult` and `G-Pois`—are derived using the reparameterization technique. Here `O`, `L`, and `G` denote estimators using overlapping group lasso with hierarchical structure built on local group $\mathcal{G}_{\text{local}}$, group lasso with local group $\mathcal{G}_{\text{local}}$, and group lasso with global group $\mathcal{G}_{\text{global}}$ penalties, respectively. Additionally, `Mult` and `Pois` refer to multinomial and Poisson multivariate categorical response models, respectively. Recall that the data-generating model is based on a multinomial model. Thus, the `O-Mult`, `L-Mult`, and `G-Mult` are penalized maximum likelihood estimators for a correctly specified model. In contrast, `O-Pois`, `L-Pois`, and `G-Pois` can be thought of as M-estimators. The seventh estimator, `G-Mult-`$\boldsymbol{\theta}$, employs the classical parameterization approach in $\boldsymbol{\theta}$. The eighth estimator, `Sep-Mult`, is designed to individually address each category in the multinomial vector, providing estimates of each response's probability mass function separately. The method denoted `Oracle` represents the true parameter, and is included to serve as a baseline.

We employ a train-validation split to select the tuning parameter λ in our simulation study, with the weight parameters $w_{\boldsymbol{k},j}$ set to 1. Specifically, we choose the candidate tuning parameter that minimizes the cross-entropy loss on the validation set.

2.6.2 Results

The estimators' performances, evaluated based on Hellinger distance and (joint) misclassification rate on a test set, is displayed in Figure 2.1 and Figure 2.2. The `Sep-Mult` estimator is correctly specified under Scheme 1, where the responses are mutually independent. Unsurprisingly, `Sep-Mult` outperforms all other estimators under this scheme. Under Schemes 2 and 3 where responses are dependent, we see `Sep-Mult` perform very poorly relative to the other methods.

The estimators `O-Mult`, `O-Pois`, `L-Mult`, `L-Pois`, `G-Mult`, and `G-Pois` are all based on our parameterization. Considering the overall performance based on the Hellinger distance and the misclassification rate, the `O-Mult` estimator is generally the most favorable. This is expected as this method is based on a correct specification of the model and can exploit the hierarchical structure of the local associations. The estimator `L-Mult` tends to perform second best when sample sizes are large. Notably, the estimator `O-Pois` performs reasonably well when $n = 100$: only `O-Mult` is evidently better. As n increases, however, `O-Pois` tends to be outperformed by the methods assuming a multinomial data generating model.

We caution that these results do not imply that estimators based on the multinomial negative log-likelihood are uniformly preferable to those utilizing the Poisson negative log-likelihood. In this case, the multinomial estimators assume a correctly specified model, and thus, as the sample size increases, tend to outperform their Poisson counterparts.

2.6.3 Poisson data generating model

Simulation study results under the Poisson data generating model are more difficult to interpret than those based the multinomial data generating model. This is partly because when fixing n, the effective sample size for the multinomial estimators is a random variable. Specifically, for each of the n samples, we draw a (possibly large) number of Poisson counts from the conditional distribution in (2.3). The multinomial estimators treat each count as an independent realization from a single-trial multinomial. Thus, the number of "samples" input into the multinomial estimators can be extremely large and vary greatly from simulation replicate to simulation replicate. For this reason, we exclude results under the Poisson from this manuscript. Nonetheless, to briefly summarize the results we observed in the simulation scenarios we considered (specifically Scheme 2 and 3), we found that under the Poisson data generating models, both `L-Pois` and `G-Pois` significantly outperformed `L-Mult` and `G-Mult`.

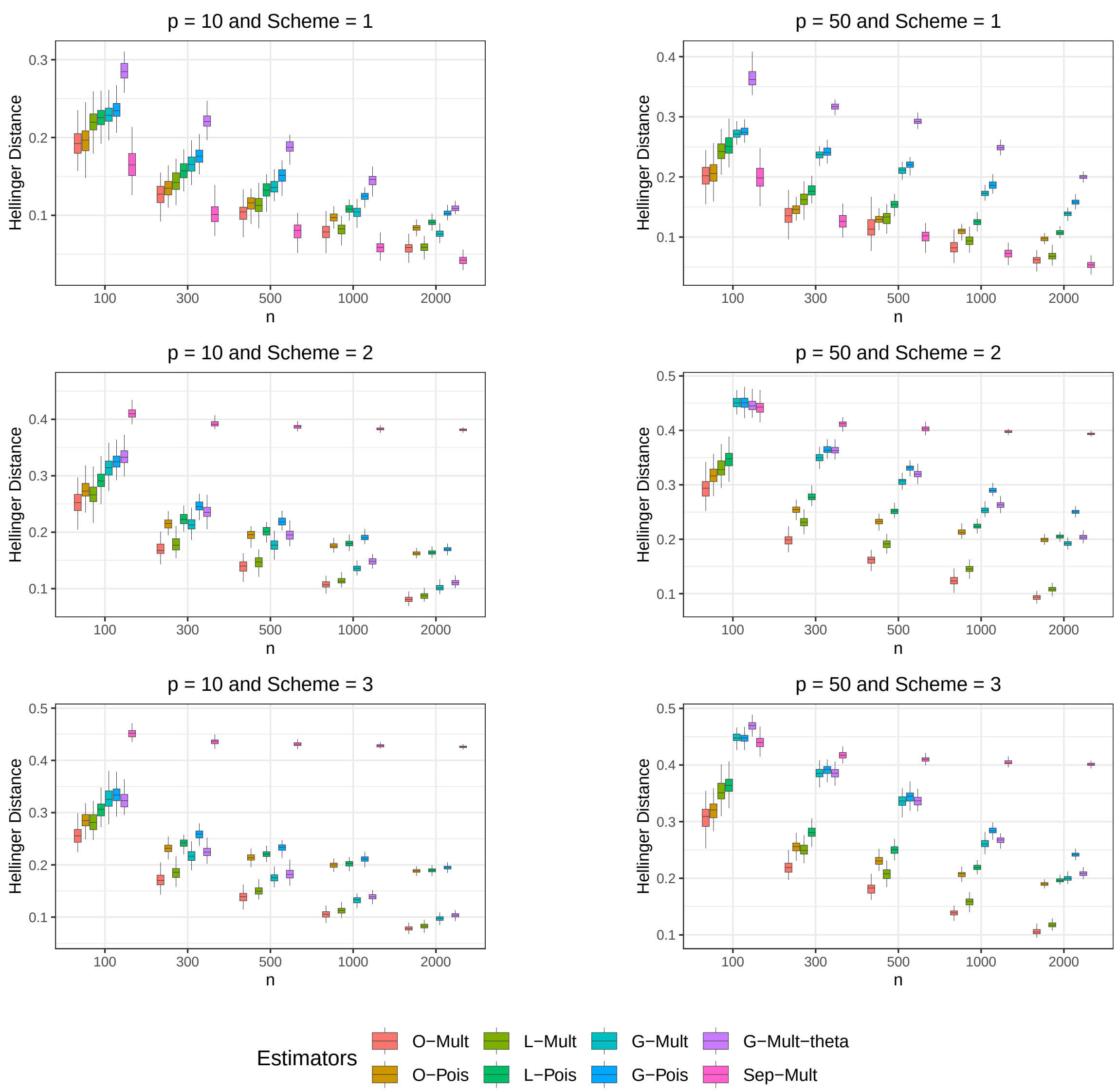

Figure 2.1: Hellinger distances for the competing estimators with $p \in \{10, 50\}$ and Scheme $\in \{1, 2, 3\}$ as n varies.

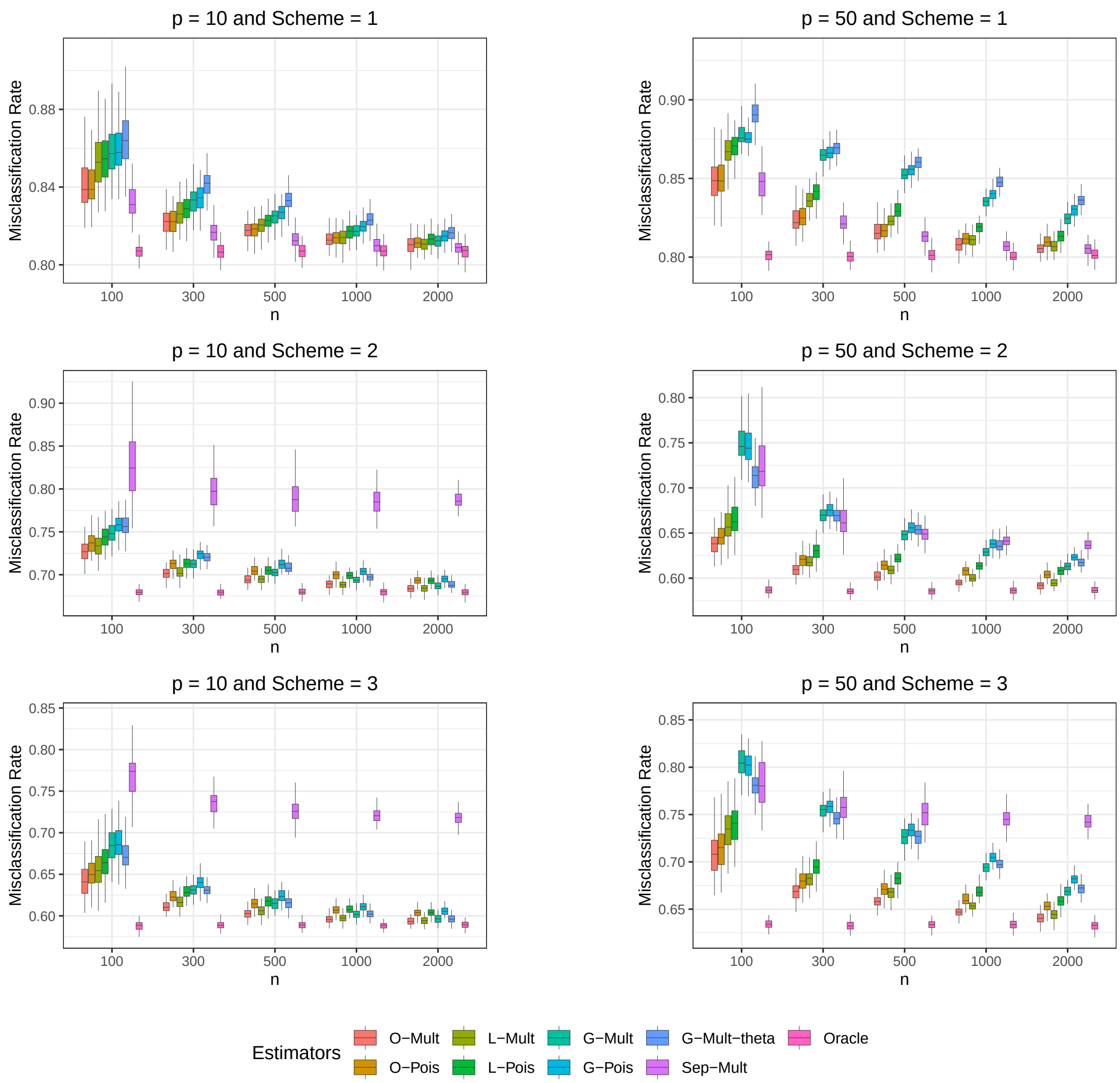

Figure 2.2: Misclassification rates for the competing estimators with $p \in \{10, 50\}$ and Scheme $\in \{1, 2, 3\}$ as n varies.

2.7 Colon tissue data analysis

We apply our proposed method to a dataset obtained from the Gene Expression Omnibus (accession number GDS3268). This dataset consists of 202 colon tissue samples, with each sample containing expression levels for tens of thousands of genes. Additionally, each sample includes three categorical response labels: (i) patient state (normal or ulcerative colitis), (ii) tissue inflammation status (inflamed or uninflamed), and (iii) anatomical location (ascending colon, descending colon, sigmoid colon, or terminal ileum). Thus, we have multiple response variables to jointly model.

First, the gene expression measurements are log-transformed and normalized. We then select a subset of the most variable genes, ranking them according to their median absolute deviation and retaining the top p genes. Highly correlated genes are pruned in order to avoid retaining genes with overlapping information. We consider several values for p; $100, 200, 300, 400$, and 500.

We randomly partition the 202 samples into three subsets: 100 for training, 50 for validation (used for tuning parameter selection), and 52 for testing. On the training set, we fit the candidate estimators introduced in Section 2.6, modeling joint associations of at most order $d = 1, 2, 3$, and compare these to a separate logistic regression baseline. The tuning parameters are selected by minimizing cross-entropy loss on the validation set. Specifically, we tune (λ, r) using the weight $w_{\boldsymbol{k},j} = r^{\max\{\|\boldsymbol{k}\|_0 - 1, 0\}}$, which penalizes higher-order associations more heavily. A standard fine grid is used for λ, whereas for r we choose from the smaller set $r \in \{1, 2, 4, 6, 10\}$ to improve computational efficiency.

The predictive performance of each fitted model is evaluated on the held-out test set using the misclassification rate, i.e., the proportion of samples in the test set whose predicted response combination does not match the observed response combination. We also evaluated predictive performance using empirical cross-entropy, estimating

$-\mathbb{E}_{X,Z} \log\{\widehat{P}(Z \mid X)\}$.

In Table 2.1, we report the misclassification rates and empirical cross-entropy averaged across 1000 random training–validation–testing splits. Overall, the estimators `G-Mult` with $d = 2$ and $d = 3$ consistently achieves the best performance on the misclassification rates, with most proposed methods outperforming the separate logistic regression baseline (`Sep-Mult`). Moreover, as p increases, our proposed methods continue to improve, the `Sep-Mult` method shows minimal change. The nearly identical results for $d = 2$ and $d = 3$ suggest that, although our model was designed to capture up to third-order associations, the higher-order interactions permitted by the model were effectively negligible in this dataset.

From the empirical cross-entropy perspective, the estimators `G-Mult` with $d = 2$ and $d = 3$, and `O-Mult`, perform equally well and outperform the others.

Table 2.1: Average misclassification rates and cross-entropy over 1000 independent training/validation/testing splits of the colon tissue dataset.

		Misclassification rate					Cross entropy				
		$p = 100$	$p = 200$	$p = 300$	$p = 400$	$p = 500$	$p = 100$	$p = 200$	$p = 300$	$p = 400$	$p = 500$
d=3	L-Mult	0.5033	0.4754	0.4712	0.4709	0.4662	1.4142	1.2983	1.2954	1.2942	1.2820
d=3	L-Pois	0.5129	0.4785	0.4732	0.4737	0.4687	2.6878	2.6415	2.6407	2.6411	2.6298
d=3	G-Mult	0.5017	0.4647	0.4520	0.4403	0.4332	1.3956	1.2917	**1.2606**	**1.2384**	**1.2082**
d=3	G-Pois	0.5072	0.4701	0.4637	0.4458	0.4342	2.7227	2.7053	2.7711	2.6384	2.6246
d=3	O-Mult	0.5070	0.4702	0.4586	0.4474	0.4416	1.4006	1.2961	1.2687	1.2481	1.2240
d=3	O-Pois	0.5111	0.4739	0.4682	0.4606	0.4533	2.7197	2.6270	2.5969	2.5932	2.5794
d=2	L-Mult	0.5022	0.4743	0.4700	0.4698	0.4652	1.4140	1.2981	1.2951	1.2937	1.2819
d=2	L-Pois	0.5117	0.4771	0.4721	0.4724	0.4676	2.6839	2.6388	2.6385	2.6396	2.6298
d=2	G-Mult	**0.5015**	**0.4643**	**0.4515**	**0.4400**	**0.4327**	1.3956	1.2918	**1.2606**	**1.2385**	**1.2082**
d=2	G-Pois	0.5067	0.4698	0.4634	0.4455	0.4341	2.7233	2.7100	2.7742	2.6387	2.6251
d=2	O-Mult	0.5034	0.4667	0.4542	0.4426	0.4358	**1.3924**	**1.2908**	1.2620	1.2403	1.2112
d=2	O-Pois	0.5070	0.4680	0.4581	0.4472	0.4388	2.7214	2.6324	2.6023	2.5974	2.5756
d=1	L-Mult	0.5227	0.4946	0.4924	0.4927	0.4873	1.4357	1.3238	1.3218	1.3200	1.3116
d=1	L-Pois	0.5346	0.4969	0.4903	0.4914	0.4843	3.0891	3.5829	3.3387	3.2668	3.5938
d=1	G-Mult	0.5184	0.4804	0.4677	0.4552	0.4469	1.4113	1.3087	1.2782	1.2530	1.2228
d=1	G-Pois	0.5276	0.4876	0.4818	0.4622	0.4499	3.3714	3.9934	3.2810	3.0970	3.0893
d=1	O-Mult	0.5171	0.4793	0.4663	0.4545	0.4469	1.4048	1.3045	1.2759	1.2517	1.2210
d=1	O-Pois	0.5244	0.4813	0.4707	0.4570	0.4467	3.3548	3.6178	3.0861	3.0439	3.0182
	G-Mult-θ	0.6023	0.5769	0.5741	0.5680	0.5604	1.7416	1.6496	1.6420	1.6358	1.6162
	Sep-Mult	0.5117	0.5044	0.5102	0.5174	0.5152	1.4367	1.3849	1.4108	1.4290	1.4175
	Null model	0.9375	0.9375	0.9375	0.9375	0.9375					

Data availability statement

Open-source R code implementing the proposed methods described in this paper is available via our GitHub repository. This repository provides clearly documented R scripts to reproduce all numerical results and figures presented. Additionally, it includes instructions for obtaining and preprocessing the publicly available gene expression dataset from the Gene Expression Omnibus (GEO) database (accession number: GDS3268).

Data availability statement

Open-source R code implementing the proposed methods described in this paper is available via our GitHub repository. This repository provides clearly documented R scripts to reproduce all numerical results and figures presented. Additionally, it includes instructions for obtaining and preprocessing the publicly available gene expression dataset from the Gene Expression Omnibus (GEO) database (accession number: GDS3268).

Chapter 3

Distributed Algorithms for High-Dimensional Statistical Inference and Structure Learning with Heterogeneous Data

3.1 Introduction

The paper is organized as follows. Section 3.2 introduces the heterogeneous linear model and establishes the necessary notation. Section 3.3 presents the constrained optimization approach using the ℓ_0-constraint and provides the general computational algorithm and the distributed version of the algorithm. In Section 3.4, we demonstrate the convergence and consistency of our proposed algorithm in general linear model setting. Section 3.5 establishes the theoretical properties of our estimator and the constrained likelihood ratio test, including the generalized Wilks' phenomenon. Finally, Section 3.6 summarizes our findings and discusses the implications of our work.

3.1.1 Our Contribution

Our main contributions are four-folded.

1. We introduce a new statistical framework specifically designed for the distributed processing of heterogeneous data, enabling comprehensive global analysis through nonconvex regularization techniques. Our research is dedicated to developing linear regression methods that effectively handle heterogeneous data, facilitating structure learning, and distinguishing between global and site-specific effects. By aggregating information from multiple sites to ascertain a global effect, we increase the overall sample size. This leads to more efficient estimation and superior inference quality compared to analyzing data from individual sites alone.

2. We develop efficient algorithms to execute the proposed methodology, utilizing nonlinear regularization with an ℓ_0-constraint. Although finding an approximately optimal solution for our optimization problem has been shown to be NP-hard in the worst-case scenario, we demonstrate that our constrained minimization approach using DC programming and the ℓ_0 projection algorithm can obtain the global minimizer with probability tending to one under the data generation distribution.

3. We present a hypothesis testing strategy for composite hypotheses that preserves the parameters of interest without regularization, while applying an ℓ_0-constraint on other parameters, such as numerous site-specific parameters, to ensure adequate control of their sparsity. We establish the asymptotic properties of the constrained likelihood ratio test, including the generalized Wilks' phenomenon, facilitating accurate inference in high-dimensional settings.

4. We demonstrate the convergence and consistency of our proposed algorithm in a general linear model setting. Our key result shows that the algorithm achieves selection consistency, ensuring that the supports of the oracle estimators are subsets of the estimated supports with high probability. Moreover, when the

sparsity tuning parameter aligns precisely with the true sparsity level, our estimator attains support recovery, guaranteeing the accurate identification of the true model structure. These theoretical findings highlight the effectiveness of our methodology in high-dimensional settings.

3.2 Heterogeneous Linear Model

In this section, we formally introduce our heterogeneous linear model and the notation used throughout. We aim to develop linear regression methods that account for heterogeneity across K sites, facilitating structure learning and distinguishing global vs. site-specific effects. At each site j, we consider loss function $L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j)$, where $\boldsymbol{\beta}_0$ denotes the global effect parameter vector and $\boldsymbol{\beta}_j$ denotes the site-specific effect nuisance parameter vector.

If we pool all patient-level data together, the combined loss function is given by

$$L(\boldsymbol{\beta}) = L_{pooled}(\boldsymbol{\beta}) := \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \boldsymbol{\beta}^T = [\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T], \tag{3.1}$$

where unknown central server parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ and site-specific nuisance parameters $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}, j = 1, \cdots, K$.

Let $\mathcal{S} = \{(k, j) : 1 \le k \le p_j, 0 \le j \le K\}$ denote the index set of parameter vector $\boldsymbol{\beta}$. Define the true parameters as $\boldsymbol{\beta}_j^0 = (\beta_{1j}^0, \beta_{2j}^0, \ldots, \beta_{p_j j}^0)^T$ for $j = 0, 1, 2, \ldots, K$. Let $A^0 = \{(k, j) \in \mathcal{S} : \beta_{kj}^0 \neq 0\}$ represent the support of the true parameter vector $\boldsymbol{\beta}^0$.

3.3 Constrained Optimization Approach

To address the challenge of heterogeneous data in high-dimensional settings, we propose a constrained optimization approach using the ℓ_0 penalty. We aim to reconstruct the oracle estimator–the least squares estimator $\widehat{\boldsymbol{\beta}}^{ol} = \left(\widehat{\boldsymbol{\beta}}^{ol}_{A^0}, \mathbf{0}\right)^T$ supported on A^0. The following optimization problems have been described in Shen et al. (2013).

Constrained ℓ_0-method

Consider the constrained least squares regression

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \\ \text{subj to:} \quad & \sum_{(k,j)\in\mathcal{S}} I\left(\beta_{kj} \neq 0\right) \leq \kappa, \end{aligned} \tag{3.2}$$

where $\kappa > 0$ is an integer-valued tuning parameter. Denote the global minimizer of (3.2) as $\widehat{\boldsymbol{\beta}}^{\ell_0} = \left(\widehat{\boldsymbol{\beta}}^{\ell_0}_{\widehat{A}^{\ell_0}}, \mathbf{0}\right)^T$. Theorem 2 in Shen et al. (2013) demonstrates that the global minimizer consistently reconstructs the oracle estimator at a degree of separation level slightly higher than the minimum required.

Inspired by the works of Shen et al. (2013), Shi et al. (2019), and Zhu et al. (2020), we employ a constrained minimization algorithm via DC programming and ℓ_0 projection to address the ℓ_0 optimization problem as formulated in (3.2).

3.3.1 Algorithm

Set the tuning parameters $(\lambda, \tau, \kappa) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{N} \cup \{0\}$. At $(t+1)$-th iteration, we solve a weighted Lasso problem,

$$\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\boldsymbol{\Gamma}}^{[t]}), \tag{3.3}$$

where

$$S(\boldsymbol{\beta}; \boldsymbol{\beta}^{[t]}) = \frac{1}{n} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) + \lambda\tau \sum_{(k,j)\in\mathcal{S}} I\left(\left|\beta^{[t]}_{kj}\right| \leq \tau\right) |\beta_{kj}|,$$

$\lambda > 0$ is a tuning parameter and $\widetilde{\mathbf{\Gamma}}^{[t]}$ is the solution of (3.3) at the t-th iteration. The DC algorithm terminates at $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{\Gamma}}^{[t]}$ if $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) \leq S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]}) +$ Machine tolerance, or if t reaches a large pre-specified maximum number of iterations. Then, we obtain the solution $\widehat{\mathbf{\Gamma}}$ of (3.2) by projection $\widetilde{\mathbf{\Gamma}}$ onto the ℓ_0-constrained set $\{\|\mathbf{\Gamma}\|_0 \leq \kappa\}$, where $\|\mathbf{\Gamma}\|_0 = \sum_{(k,j)\in\mathcal{S}} I(\Gamma_{kj} \neq 0)$. We summarize the general constrained minimization via DC programming and ℓ_0 projection algorithm in Algorithm 1.

Algorithm 1 Constrained minimization via DC programming & ℓ_0 projection

---

1: **Initialization:** Specify $\lambda > 0$, $\tau > 0$, and $\kappa \geq 1$. Set $t = 0$. Initialize $\widetilde{\mathbf{\Gamma}}^{[0]} = \left\{\widetilde{\Gamma}^{[0]}_{kj}\right\}_{(k,j)\in\mathcal{S}}$.

2: **Weighted Lasso Update:** Use a weighted Lasso solver to solve (3.3).

3: **Check Convergence:** If $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) - S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]})$ has not converged, set $t \leftarrow t+1$ and return to line 2.

4: **Identify the Top-κ Indices:** Let

$$C = \left\{(k', j') \in \mathcal{S} : \sum_{(k,j)\in\mathcal{S}} I\left(|\widetilde{\Gamma}^{[t]}_{kj}| \geq |\widetilde{\Gamma}^{[t]}_{k'j'}|\right) \leq \kappa\right\}.$$

Without loss of generality (WLOG), assume $|C| = \kappa$. Otherwise, if $|C| < \kappa$, then select $\kappa - |C|$ more elements from $\arg\max_{(k,j)\in\mathcal{S}\backslash C} \left|\widetilde{\Gamma}^{[t]}_{kj}\right|$.

5: **ℓ_0-Projected Estimator:** Compute the ℓ_0 projection estimator $\widehat{\mathbf{\Gamma}}$:

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \text{ s.t. } \beta_{kj} = 0 \text{ for } (k,j) \in \mathcal{S}\backslash C. \tag{3.4}$$

6: **Output:** The ℓ_0-projected estimator $\widehat{\mathbf{\Gamma}}$.

For the weighted Lasso problem (3.3) in step 2 of Algorithm 1, we can consider a first-order iterative algorithm, such as ISTA Daubechies et al. (2004) and FISTA Beck and Teboulle (2009). Denote the first order iterative solver with weights $\boldsymbol{w} =$

$\{w_{k,j}; (k,j) \in \mathcal{S}\}$,

$$\widehat{\boldsymbol{\beta}}^{(l+1)} = \text{solver}\left(\widehat{\boldsymbol{\beta}}^{(l)}, \frac{\partial S(\widehat{\boldsymbol{\beta}}^{(l)})}{\partial \boldsymbol{\beta}}; \boldsymbol{w}\right).$$

In multicenter research, individual-level data are often protected and cannot be shared across sites. Therefore, it is essential that our weighted Lasso solver is designed to operate under these constraints. Specifically, the central server parameter $\boldsymbol{\beta}_0$ from the previous iteration and its partial derivative $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0}$ should be communicated to the central server. Meanwhile, the site-specific nuisance parameters at the jth site, $\boldsymbol{\beta}_j$, from the previous iteration and their partial derivatives $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}$ should remain local to the jth site.

Define the central server weight and the site weights $\boldsymbol{w}^j = \{w_{k',j'}; (k',j') \in \mathcal{S}, j' = j\}, j = 0, 1, \cdots, K$. The central server solver and the site solvers are given by

$$\text{central server update} : \widehat{\boldsymbol{\beta}}_0^{(l+1)} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_0^{(l)}, \sum_{j=1}^{K} \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_0}; \boldsymbol{w}^0\right), \text{ and} \quad (3.5)$$

$$\text{site server update} : \widehat{\boldsymbol{\beta}}_j^{(l+1)} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_j^{(l)}, \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_j}; \boldsymbol{w}^j\right), 1 \le j \le K, \quad (3.6)$$

where for any $j = 1, \cdots, K$, $S_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j)$.

At the fixed t-th iteration in Algorithm 1, the weight for the l-th iteration of the weighted Lasso solver (step 2) is given by

$$\boldsymbol{w} = \left\{\lambda\tau \cdot I\left(\left|\widetilde{\Gamma}_{kj}^{[t]}\right| \le \tau\right); (k,j) \in \mathcal{S}\right\},$$

where $\widetilde{\boldsymbol{\Gamma}}^{[t]}$ is the solution at the t-th iteration.

For the central server and each site $j \in \{0, 1, \cdots, K\}$, the corresponding weights are

$$\boldsymbol{w}^j = \left\{ \lambda\tau \cdot I\left(\left| \widetilde{\Gamma}^{[t]}_{k'j'} \right| \leq \tau \right); (k', j') \in \mathcal{S}, j' = j \right\}.$$

Identifying the top-κ indices in Step 4 of Algorithm 1 might appear to require transmitting all site-specific nuisance parameters to the central server for ranking. However, a threshold-based selection algorithm (see Algorithm 4 in Section 5.5) enables this step to be performed in a fully distributed manner: each site sends only a summarized count to the central server, thereby avoiding the need to share individual parameter estimates.

We summarize the constrained minimization algorithm, which employs DC programming and ℓ_0 projection, in the distributed algorithm setting, as presented in Algorithm 2.

Remark 3.1. The initial $\widetilde{\boldsymbol{\Gamma}}^{[0]}$ needs to be sparse, such as $\mathbf{0}$ or a sparse estimator obtained through penalized methods.

Algorithm 2 Constrained Minimization in the Distributed Algorithm Setting

1: **Initialization:** Specify $\lambda > 0$, $\tau > 0$, $\kappa \geq 1$, and $t = 0$. Initialize $\widetilde{\mathbf{\Gamma}}^{[0]} = \{\widetilde{\Gamma}^{[0]}_{k,j}\}_{(k,j)\in\mathcal{S}}$.

2: For each $j = 0, \ldots, K$, define $\boldsymbol{w}^j = \left\{\lambda\tau \cdot I\left(\left|\widetilde{\Gamma}^{[t]}_{k'j'}\right| \leq \tau\right); (k', j') \in \mathcal{S}, j' = j\right\}$.

Set the inner iteration counter $l = 0$, and initialize $\widehat{\boldsymbol{\beta}}_j^{(l)}$ for all $0 \leq j \leq K$.

3: **Site-by-Site Parameter Updates:**

4: For each $j \in \{1, 2, \ldots, K\}$:

- Update the site-specific parameter $\widehat{\boldsymbol{\beta}}_j^{(l)}$ using the weighted Lasso solver in (3.6).
- Pass $\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_0}$ to the central server.
- The central server updates $\widehat{\boldsymbol{\beta}}_0^{(l)}$ according to (3.5).

5: **Check Convergence of Inner Iterations:** If

$$\max_{1\leq j\leq K}\left\|\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_j}\right\|_2 \quad \text{and} \quad \left\|\sum_{j=1}^{K}\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_0}\right\|_2$$

are below prespecified tolerances, proceed; otherwise set $l \leftarrow l + 1$ and return to Step 3.

6: **Update Overall Parameter Estimates:** Set $\widetilde{\mathbf{\Gamma}}^{[t+1]} \leftarrow \widehat{\boldsymbol{\beta}}_l$. If

$$S\big(\widetilde{\mathbf{\Gamma}}^{[t]}, \widetilde{\mathbf{\Gamma}}^{[t]}\big) \; - \; S\big(\widetilde{\mathbf{\Gamma}}^{[t+1]}, \widetilde{\mathbf{\Gamma}}^{[t]}\big)$$

has not converged, set $t \leftarrow t + 1$ and return to Step 2.

7: **Identify Top-κ Indices (Threshold-Based Selection):** Apply the threshold-based selection algorithm (see Algorithm 4 in Section 5.5) to $\widetilde{\mathbf{\Gamma}}^{[t]}$, obtaining a set $C \subset \mathcal{S}$ such that $|C| = \kappa$.

8: **ℓ_0-Projected Estimator:**

$$\widehat{\mathbf{\Gamma}} \; = \; \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \quad \text{s.t.} \quad \beta_{k,j} = 0 \;\text{ for } (k, j) \in \mathcal{S} \setminus C.$$

9: **Output:** The ℓ_0-projected estimator $\widehat{\mathbf{\Gamma}}$.

3.4 Convergence and Consistency Results

3.4.1 Problem Setup and Notations

Before presenting our main theoretical results, we first introduce the linear model setup and necessary notation. Assume the training data are given by $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \boldsymbol{\beta} \in \mathbb{R}^p$ and $y_1, \cdots, y_n \in \mathbb{R}$. Furthermore, assume that y_i, given $\boldsymbol{x}_i$, has density $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$. For $B \subset [p]$, consider hypothesis testing

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0. \tag{3.7}$$

Here, $\boldsymbol{\beta}_B = 0$ if and only if $\beta_i = 0$ for any $i \in B$.

	Index Set	**Parameter Dimension**	**DC Algorithm**
Non-distributed	$i \in [p]$	$\boldsymbol{\beta} \in \mathbb{R}^p \quad (p = \sum_{j=0}^K p_j)$	Algorithm 1
Distributed	$(k, j) \in \mathcal{S}$	$\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$, $0 \leq j \leq K$	Algorithm 2

In this section, we present the non-distributed version of our DC programming and ℓ_0-projection procedure in Algorithm 3. It builds on the framework of Algorithm 1. A corresponding distributed version, analogous to Algorithm 2, can be derived with straightforward extensions; hence, we omit the distributed counterpart of Algorithm 3.

3.4.2 Computational Algorithm

Consider the constrained optimization problem regression for H_0 in (3.7)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^n -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) \\ \text{subj to:} \quad & \sum_{i \in [p] \setminus B} I(\beta_i \neq 0) \leq \kappa, \beta_B = 0 \end{aligned} \tag{3.8}$$

and for H_1 in (3.7)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) \\ \text{subj to:} \quad & \sum_{i \in [p] \setminus B} I(\beta_i \neq 0) \leq \kappa, \end{aligned} \tag{3.9}$$

where $\kappa > 0$ is an integer-valued tuning parameter. Set

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})).$$

The oracle estimators corresponding to H_0 and H_1 are given by

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_0} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_0})^c}=\mathbf{0}} L(\boldsymbol{\beta}) \text{ with } \kappa^0_{H_0} = |A^0_{H_0}|, \text{ and} \tag{3.10}$$

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_1} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_1})^c}=\mathbf{0}} L(\boldsymbol{\beta}), \text{ respectively,} \tag{3.11}$$

where $A^0_{H_0} = \{i \in [p] \setminus B; \beta^0_i \neq 0\}$ and $A^0_{H_1} = \{i \in [p] \setminus B; \beta^0_i \neq 0\} \cup B$.

For the given hypothesis set B, at $(t+1)$-th iteration, we solve the following weighted Lasso problems, corresponding to H_0 and H_1 respectively:

$$\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_B=\mathbf{0}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \text{ and} \tag{3.12}$$

$$\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \tag{3.13}$$

where $\lambda > 0$ is a tuning parameter,

$$S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}) = \frac{1}{n} L(\boldsymbol{\beta}) + \lambda\tau \sum_{i \in [p] \backslash B} I\left(\left|\widetilde{\Gamma}_i^{[t]}\right| \leq \tau\right) |\beta_i|,$$

and $\widetilde{\boldsymbol{\Gamma}}^{[t]}$ is the solution of (3.12) or (3.13), respectively, at the t-th iteration. The DC algorithm terminates at $\widetilde{\boldsymbol{\Gamma}} = \widetilde{\boldsymbol{\Gamma}}^{[t]}$ such that $S(\widetilde{\boldsymbol{\Gamma}}^{[t]}; \widetilde{\boldsymbol{\Gamma}}^{[t]}) \leq S(\widetilde{\boldsymbol{\Gamma}}^{[t+1]}; \widetilde{\boldsymbol{\Gamma}}^{[t]})+$ Machine epsilon, or if t reaches a large pre-specified maximum number of iterations (or $\text{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t]}\} \backslash B = \text{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t+1]}\} \backslash B$). We then obtain the approximated solution $\widehat{\boldsymbol{\Gamma}}$ to (3.8) or (3.9), respectively, by projecting $\widetilde{\boldsymbol{\Gamma}}_{B^c}$ onto the ℓ_0-constrained set $\{\|\boldsymbol{\Gamma}_{B^c}\|_0 \leq \kappa\}$.

Algorithm 3 Constrained minimization via DC programming & ℓ_0 projection

---

1: Specify $\lambda > 0$, $\tau > 0$, and $\kappa \geq 1$. Set $t = 0$. Initialize $\widetilde{\boldsymbol{\Gamma}}^{[0]} = \left\{\widetilde{\Gamma}_i^{[0]}\right\}_{i \in [p]}$.

2: Use a weighted Lasso solver to solve (3.12) for H_0 or (3.13) for H_1 .

3: If $S(\widetilde{\boldsymbol{\Gamma}}^{[t]}; \widetilde{\boldsymbol{\Gamma}}^{[t]}) - S(\widetilde{\boldsymbol{\Gamma}}^{[t+1]}; \widetilde{\boldsymbol{\Gamma}}^{[t]})$ has not converged, set $t \leftarrow t + 1$ and return to line 2.

4: (ℓ_0-projection) Let

$$C = \left\{ i' \in [p] \backslash B; \sum_{i \in [p] \backslash B} I(|\widetilde{\Gamma}_i^{[t]}| \geq |\widetilde{\Gamma}_{i'}^{[t]}|) \leq \kappa, i' \in [p] \backslash B \right\}.$$

WLOG, assume $|C| = \kappa$. Otherwise, if $|C| < \kappa$, then select $\kappa - |C|$ more elements from $\arg\max_{i \in [p] \backslash (B \cup C)} \left|\widetilde{\Gamma}_i^{[t]}\right|$ into C.

5: Compute the ℓ_0 projection estimators $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_{H_0}$ or $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_{H_1}$, respectively, according to:

$$H_0 : \widehat{\boldsymbol{\Gamma}}_{H_0} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p] \backslash C, \text{ or} \tag{3.14}$$

$$H_1 : \widehat{\boldsymbol{\Gamma}}_{H_1} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p] \backslash (B \cup C). \tag{3.15}$$

3.4.3 Assumptions

To derive the convergence and consistency results of Algorithm 3, we will focus exclusively on the least squares regression setting from this point forward in the section. We begin by considering the linear model:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{3.16}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, and σ^2 might depend on n. Consider $B \subset [p]$ such that

$$\frac{\sqrt{|B|}(|A^0| + |B|)}{n} \to 0. \tag{3.17}$$

Without loss of generality, we can set $S(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) = \frac{1}{2}\left\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2$.

Let $\kappa \geq |A^0 \backslash B|$, and $\kappa_{max} = \max\left\{\kappa, |\{i \in [p]\backslash B; |\widetilde{\Gamma}_i^{[0]}| \geq \tau\}|\right\}$, where $\widetilde{\boldsymbol{\Gamma}}^{[0]}$ denotes the initial estimator used in Algorithm 3. Without loss of generality, we can assume that $\widetilde{\boldsymbol{\Gamma}}_B^{[0]} = \mathbf{0}$.

To derive the statistical and computational properties of Algorithm 3 in least squares regression setting, we introduce the following technical assumptions which generalized the convergence and consistency of structure learning assumptions from Li et al. (2023).

Assumption 3.1 (Restricted eigenvalues)**.** For a constant $c_1 > 0$,

$$\min_{A:|A\backslash B|\leq 2\kappa_{max}} \min_{\boldsymbol{\xi}:\|\boldsymbol{\xi}_{A^c}\|_1 \leq 3\|\boldsymbol{\xi}_A\|_1} \frac{\|\mathbf{X}\boldsymbol{\xi}\|_2^2}{n\,\|\boldsymbol{\xi}\|_2^2} \geq c_1, \tag{3.18}$$

where $\boldsymbol{\xi}_A \in \mathbb{R}^{|A|}$ is the projection of $\boldsymbol{\xi} \in \mathbb{R}^p$ onto coordinates in A.

Assumption 3.2. For constants $c_2, c_3 > 0$,

$$\begin{aligned} &\max_{1\le i\le p} \frac{1}{n}(\mathbf{X}^T(I-P_A)\mathbf{X})_{ii} \le c_2^2, \\ &\max_{1\le i\le p} n((\mathbf{X}_A^T\mathbf{X}_A)^\dagger)_{ii} \le c_3^2, \end{aligned} \tag{3.19}$$

where $P_A = \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^\dagger\mathbf{X}_A^T$, and $A \in \{A^0_{H_0}, A^0_{H_1}\}$.

Assumption 3.3 (Nuisance signals)**.**

$$\min_{\beta_i^0\ne 0, i\notin B} \frac{|\beta_i^0|}{\sigma} \ge \frac{50c_3}{3}\sqrt{\frac{\log p}{n} + \frac{\log n}{n}}. \tag{3.20}$$

Assumption 3.4 (Degree of separation)**.**

$$\begin{aligned} C_{\min} = C_{\min}(\boldsymbol{\beta}^0, \mathbf{X}) := &\min_{A:|A|\le|A^0| \text{ and } A\ne A^0} \inf_{\boldsymbol{\beta}} \frac{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_{A\cup B}\boldsymbol{\beta}_{A\cup B}\|_2^2}{n\,|A^0\backslash A|} \\ &\ge 72\sigma^2\frac{\log p + \log n}{n}. \end{aligned} \tag{3.21}$$

Assumption 3.1 is a common condition related to restricted eigenvalues, as discussed in Bickel et al. (2009) and Wainwright (2019). Assumption 3.2 generalizes from the lower eigenvalue and mutual incoherence conditions found in Section 7.5.1 of Wainwright (2019). Assumption 3.3 specifies the minimal signal strength across the support, which is used to establish high-dimensional variable selection consistency, as seen in Fan et al. (2014) and Loh and Wainwright (2017). Finally, Assumption 3.4 is a commonly recognized condition for the degree of separation in feature selection, according to Shen et al. (2013) and Zhu et al. (2020).

3.4.4 Correct Identification

The theory presented extends the correct identification result for structure learning, as found in Theorem 14 of Li et al. (2023), to include selection consistency.

Theorem 3.1. *Under Assumptions 3.1, 3.2, 3.3, and 3.4, if the tuning parameters* (κ, τ, λ) *of Algorithm 3 in the least squares regression setting satisfy:*

1. $\sqrt{32\sigma^2 c_3^2\big(\frac{\log p}{n} + \frac{\log n}{n}\big)} \ \leq \ \tau \ \leq \ \min_{i\in B^c, \beta_i^0\neq 0}|\beta_i^0|,$
2. $\kappa = |A_{H_0}^0|,$
3. $\frac{1}{\tau}\sqrt{32\sigma^2 c_2^2\big(\frac{\log p}{n} + \frac{\log n}{n}\big)} \ \leq \ \lambda \ \leq \ c_1/6,$

then the following statements hold:

- *under both H_0 and H_1, $\widehat{\mathbf{\Gamma}}$ in Algorithm 3 yields the oracle estimators (3.10) and (3.11), as well as the global minimizer of (3.8) and (3.9), respectively;*
- *the DC algorithm almost surely converges in at most $\lceil \log\big(2\,\kappa_{\max}\big)/\log 4\rceil$ iterations, where $\kappa_{\max} = \max\{\kappa, \kappa_1\}$ and $\kappa_1 = \Big|\{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[0]}| \ \geq \ \tau\}\Big|$.*

Moreover, by replacing condition 2 with $\kappa \geq |A_{H_0}^0|$, Algorithm 3 ensures:

- *under both H_0 and H_1, the supports of the oracle estimators (3.10) and (3.11) are subsets of $\mathrm{supp}\big(\widehat{\mathbf{\Gamma}}\big)\backslash B$ and $\mathrm{supp}\big(\widehat{\mathbf{\Gamma}}\big) \cup B$, respectively, almost surely;*
- *the DC algorithm almost surely converges in at most $\lceil \log\big(2\,\kappa_{\max}\big)/\log 4\rceil$ iterations.*

The first part of Theorem 3.1 establishes the result of almost sure subset recovery, while the second part confirms the almost sure selection consistency for Algorithm 3.

Remark 3.2. Building on the foundation established by Algorithm 3 and Theorem 3.1, our constrained DC algorithm, incorporating the ℓ_0 projection, is capable of reaching a global minimum within polynomial time, with the probability approaching 1 as $n, p \to \infty$. This outcome starkly contrasts with previous findings, such as those reported by Chen et al. (2017, 2019), which state that no algorithm can consistently solve such nonconvex minimization problems in polynomial time under worst-case conditions.

3.5 Sampling Distribution and Hypothesis Testing

In this section, we establish the theoretical properties of our proposed estimator, including its sampling distribution under various conditions.

For a hypothesized parameter subset $B \subset \mathcal{S}$, we consider the hypothesis testing

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0, \tag{3.22}$$

where $\boldsymbol{\beta}_B = 0$ if and only if $\beta_{kj} = 0$ for all $(k, j) \in B$.

3.5.1 Constrained likelihood ratio testing

The problem of constructing a constrained likelihood ratio with a sparsity constraint on nuisance parameters has been discussed in Zhu et al. (2020) and Shi et al. (2019). In this section, we illustrate our approach using a simple heterogeneous linear regression setting as an example. In Section 5.1, we derived a heterogeneous linear regression model, which can be summarized as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

with the log-likelihood

$$\mathcal{L}_n(\boldsymbol{\beta}, \sigma) = -\frac{1}{2\sigma^2} \sum_{j=1}^{K} \left\| \boldsymbol{Y}_j - \mathbf{X}_j \boldsymbol{\beta}_0 - \mathbf{W}_j \boldsymbol{\beta}_j \right\|_2^2 - \frac{n}{2} \log(2\pi\sigma^2),$$

where $\|\cdot\|_2$ denotes the Euclidean norm and the relationship between $\{\mathbf{X}_j, \mathbf{W}_j\}_{j=1}^{K}$ and $\mathbf{X}$ is given by (5.3). The constrained log-likelihood ratio, corresponding to the test (3.22), is defined as

$$2\left(\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^1, \widehat{\sigma}^1) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^0, \widehat{\sigma}^0)\right),$$

where $(\widehat{\boldsymbol{\beta}}^0, \widehat{\sigma}^0)$ and $(\widehat{\boldsymbol{\beta}}^1, \widehat{\sigma}^1)$ are the constrained maximum likelihood estimators (CMLE) based on the null and full spaces of the hypothesis test, respectively, that is,

$$\widehat{\boldsymbol{\beta}}^0 = \arg \min_{\|\boldsymbol{\beta}\|_0 \leq \kappa, \boldsymbol{\beta}_B = \mathbf{0}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \text{ and} \tag{3.23}$$

$$\widehat{\boldsymbol{\beta}}^1 = \arg \min_{\|\boldsymbol{\beta}\|_{0,B} \leq \kappa} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \tag{3.24}$$

where $\|\boldsymbol{\beta}\|_{0,B} = \sum_{(k,j)\in\mathcal{S}} I(\beta_{kj} \neq 0) I((k,j) \notin B)$ and

$$L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = \left\|\boldsymbol{Y}_j - \mathbf{X}_j\boldsymbol{\beta}_0 - \mathbf{W}_j\boldsymbol{\beta}_j\right\|_2^2.$$

To conduct the hypothesis test (3.22) in this heterogeneous linear regression setting, we replace the non-convex MLEs in (3.23)–(3.24) by the solutions $\widehat{\boldsymbol{\Gamma}}_{H_0}$ and $\widehat{\boldsymbol{\Gamma}}_{H_1}$ from our DC programming and ℓ_0-projection Algorithm 3.

The resulting statistic

$$\Lambda_n(B) := 2\left(\mathcal{L}_n(\widehat{\boldsymbol{\Gamma}}_{H_1}, \widehat{\sigma}^1) - \mathcal{L}_n(\widehat{\boldsymbol{\Gamma}}_{H_0}, \widehat{\sigma}^0)\right),$$

compares these constrained estimators in a manner analogous to a traditional log-likelihood ratio test, where $(\widehat{\sigma}^l)^2 = \frac{1}{n}\left\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\Gamma}}_{H_l}\right\|^2$ for $l \in \{0, 1\}$.

Under suitable conditions on κ, τ, λ and the set $|B|$, Theorem 3.2 establishes Wilks' phenomenon in both the fixed-dimensional and increasing-dimensional regimes for $|B|$. Specifically, the theorem shows that the distribution of $\Lambda_n(B)$ converges to a chi-square distribution for fixed $|B|$ and to a normal distribution (after appropriate centering and scaling) for $|B| \to \infty$.

Theorem 3.2. *Suppose* $\frac{\sqrt{|B|}\left(|A^0|+|B|\right)}{n} \longrightarrow 0$. *Under Assumptions 3.1–3.4, if there* *exist tuning parameters (κ, τ, λ) satisfying the three conditions in Theorem 3.1, with $\kappa = |A^0_{H_0}|$, then, under the null hypothesis $H_0 : \boldsymbol{\beta}_B = 0$ (i.e., $|A^0| = |A^0_{H_0}|$), the following hold:*

1. ***Wilks' phenomenon.*** *If $\beta_{k,j} = 0$ for all $(k, j) \in B$ and $|B|$ is fixed, then*

$$\Lambda_n(B) \stackrel{d}{\to} \chi^2_{|B|} \quad \text{as } n \to \infty.$$

2. ***Generalized Wilks' phenomenon.*** *If $\beta_{k,j} = 0$ for all $(k, j) \in B$ and $|B| \to \infty$, then*

$$(2|B|)^{-\frac{1}{2}}\big(\Lambda_n(B) - |B|\big) \stackrel{d}{\to} N(0, 1) \quad \text{as } n \to \infty.$$

In the context of linear regression, a straightforward asymptotic result can be derived.

Theorem 3.3. *Under the same setting as Theorem 3.2, let B be fixed. Assume further that the Moore–Penrose inverse*

$$\sigma^2 \left(\frac{1}{n} \mathbf{X}^T_{A^0 \cup B} \mathbf{X}_{A^0 \cup B} \right)^{\dagger}_{B,B}$$

converges in distribution to a positive semidefinite matrix Σ. Also assume that

$$\big\{ \boldsymbol{\xi} \in \mathbb{R}^{A^0 \cup B} : \boldsymbol{\xi}_i = 0 \text{ for all } i \notin B \big\} \ \subset \ \mathcal{R}\big(\mathbf{X}^T_{A^0 \cup B}\big),$$

where $\mathcal{R}(\mathbf{X}^T_{A^0 \cup B})$ denotes the column space of $\mathbf{X}^T_{A^0 \cup B}$. Then

$$\sqrt{n}(\widehat{\boldsymbol{\Gamma}}^{(1)}_B - \boldsymbol{\beta}^0_B) \xrightarrow{d} N(0, \Sigma). \tag{3.25}$$

Since the literature Shen et al. (2012); Zhu et al. (2020); Shen et al. (2013); Kim et al. (2013); Pan et al. (2013); Wu et al. (2020, 2016); Austin et al. (2020) has already systematically studied TLP inference under a wide range of settings, we omit the simulation study on inference performance.

3.6 Real Data Analysis

In this section, we evaluate the predictive performance of our distributed high dimensional framework for heterogeneous data using the METABRIC breast cancer dataset from Kaggle. The dataset contains RNA expression profiles, mutation data, and detailed clinical annotations, making it a rich resource for modeling high-dimensional survival outcomes with potential heterogeneity.

3.6.1 Data Preprocessing

The raw METABRIC data were processed in R, following the pipeline available on our GitHub repository. A detailed dataset description is available on Kaggle dataset webpage. The primary response variable, denoted by `overall_survival_months`, represents the duration from the intervention to the occurrence of death. Among clinical covariates, the `cohort` variable (ranging from 1 to 5) is notable because it naturally partitions subjects into distinct subpopulations, analogous to multiple sites in a multicenter study.

We separate the predictor variables into two categories: (1) clinical attributes, which we further label as 'global' (e.g., baseline demographics, tumor characteristics) or 'site-specific' (e.g., type of surgical intervention, post-chemotherapy cellularity, and treatment indicators) that may differ by cohort, and (2) genetic attributes (RNA expression features), treated as global. In total, these variables yield a high-dimensional design matrix with both global and site-specific components.

To evaluate predictive performance, we consider two modeling strategies. First, a heterogeneous linear model is fitted, integrating both global and site-specific predictors within a unified likelihood framework. Second, independent linear models are estimated separately for each cohort. The results demonstrate that the heterogeneous model, which jointly models global and site-specific effects, exhibits superior predictive accuracy compared to the disjoint cohort-specific models.

3.6.2 Model Estimation and Evaluation

We implement a heterogeneous linear model that jointly fits global and site-specific predictors via a truncated Lasso penalty (TLP), using the `glmtlp` R package. Given the skewness of the survival response, we consider the log-transformed outcome variable defined as

$$\log(\texttt{overall\_survival\_months} + 1).$$

A 70/30 train-test split is performed, and hyperparameter tuning is conducted via 5-fold cross-validation. To ensure stability in evaluation, the process is repeated over 1000 iterations, and the mean squared error (MSE) on the test set is computed for each repetition. The average test MSE and standard deviation for both the heterogeneous model and the cohort-specific model are summarized in Table 3.1.

For comparison, we also fit separate TLP-regularized linear models for each cohort individually, treating them as if they were independent analyses. In that case, we compute the test MSE for each cohort, then take a weighted average (weighted by cohort size). Table 3.1 shows the average test MSE and standard deviation across the 1000 repetitions.

Table 3.1: Comparison of Test MSE Across Model Frameworks

Model	Average Test MSE	Standard Deviation
Heterogeneous Model	0.6368	0.0508
Cohort-Specific Model	0.6650	0.0527

Table 3.1 shows that the heterogeneous linear model—which leverages both global and cohort-specific effects in a unified framework—achieves a lower average test MSE of 0.6368 than the cohort-specific approach (0.6650), with both standard errors no greater than 0.00167. This difference suggests that jointly modeling global features and site-specific terms can yield more accurate predictions, presumably because the global parameters borrow information across cohorts.

3.6.3 Discussion

Overall, the METABRIC example demonstrates that our distributed, high dimensional framework can effectively combine global and site-specific predictors to enhance predictive performance. Even when cohorts are naturally heterogeneous, joint estimation methods improve accuracy by pooling information across sites. This advantage highlights the promise of distributed approaches that incorporate nonconvex regularization.

3.7 Summary

In this paper, we have proposed a novel approach for handling heterogeneous data in high-dimensional statistical inference and structure learning problems. The proposed framework utilizes a parametric likelihood setting and introduces a truncated lasso penalty (TLP) for variable selection and parameter estimation.

For hypothesis testing, we have developed a procedure that leaves the parameters of interest unregularized while imposing an ℓ_0-constraint on the nuisance parameters to control their sparsity. Under a degree of separation condition and suitable choices of the tuning parameters, we have established the asymptotic properties of the constrained likelihood ratio statistic.

In terms of parameter estimation, we have proposed a constrained optimization approach using DC programming and ℓ_0 projection. We have established the theoretical properties of the resulting estimator, including its selection consistency and support recovery when the tuning parameter for the ℓ_0-constraint equals the true sparsity level. Moreover, we have shown that the estimator attains the oracle property and global minimizer of the constrained optimization problem within a logarithmic number of iterations.

The proposed methodology offers several advantages in the context of distributed learning with heterogeneous data. By allowing for site-specific nuisance parameters, our approach can effectively account for the inherent heterogeneity across different data sources. The use of the truncated lasso penalty enables simultaneous variable selection and parameter estimation, leading to more interpretable models.

Chapter 4

Supplementary Material for Subspace Decompositions for Association Structure Learning in Multivariate Categorical Response Regression

4.1 Linear Algebra Definitions and Related Discussions

The following definition of internal direct sum, tensor product and orthogonal direct sum originated from Section 1, 9 and 14 of Roman et al. (2005).

Internal Direct Sum. Let V be a vector space over $\mathbb{R}$. Suppose $V_1, V_2, \ldots, V_q$ are subspaces of V. The vector space V is said to be the *internal direct sum* of $V_1, V_2, \ldots, V_q$, denoted by

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_q,$$

if every element $v \in V$ can be uniquely written as

$$v = v_1 + v_2 + \cdots + v_q,$$

where $v_i \in V_i$ for $i = 1, 2, \ldots, q$.

Tensor Product Space. Let $V_1, V_2, \ldots, V_q$ be vector spaces over $\mathbb{R}$. Suppose $\{v_1^1, v_2^1, \ldots, v_{J_1}^1\}$ is a basis for V_1, $\{v_1^2, v_2^2, \ldots, v_{J_2}^2\}$ is a basis for V_2, and so on up to $\{v_1^q, v_2^q, \ldots, v_{J_q}^q\}$ as a basis for V_q. The *tensor product* of $V_1, V_2, \ldots, V_q$, denoted by $V_1 \otimes V_2 \otimes \cdots \otimes V_q$, is a vector space over $\mathbb{R}$. The basis for the tensor product space $V_1 \otimes V_2 \otimes \cdots \otimes V_q$ is

$$\{\otimes_t(v_{j_1}^1, v_{j_2}^2, \cdots, v_{j_q}^q) \mid 1 \leq j_l \leq J_l, l = 1, 2, \ldots, q\}.$$

Explicitly, this means that the tensor product space $V_1 \otimes V_2 \otimes \cdots \otimes V_q$ has dimension $|\boldsymbol{J}|$, and its basis elements are formed by taking the tensor product of each combination of basis elements from $V_1, V_2, \ldots, V_q$.

Orthogonal Direct Sum. Let V be a vector space over $\mathbb{R}$ equipped with an inner product. Suppose $V_1, V_2, \ldots, V_n$ are subspaces of V. The vector space V is said to be the *orthogonal direct sum* of $V_1, V_2, \ldots, V_n$, denoted by

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_n,$$

if every element $v \in V$ can be uniquely written as

$$v = v_1 + v_2 + \cdots + v_n,$$

where $v_i \in V_i$ for $i = 1, 2, \ldots, n$ and $V_i \perp V_j$ for all $i \neq j$.

In this paper, if we set $V_l = \mathbb{R}^{J_l}$ and $v_{j_l}^l = \boldsymbol{e}_{j_l}^{J_l}$ for any $l \in [q]$, then without loss of any linear algebra property, we can treat the basis $\otimes_t(v^1_{j_1}, v^2_{j_2}, \cdots, v^q_{j_q})$ the same as $\boldsymbol{E}^{\boldsymbol{J}}_{j_1,\cdots,j_q}$, where $\boldsymbol{E}^{\boldsymbol{J}}_{j_1,\cdots,j_q}$ denotes the standard basis q-way array for $\mathbb{R}^{\boldsymbol{J}}$, which is defined as the array in $\mathbb{R}^{\boldsymbol{J}}$, whose $(j_1, \cdots, j_q)$-th entry is 1 and all other entries are 0. Thus, the tensor product space $\mathbb{R}^{J_1} \otimes \cdots \otimes \mathbb{R}^{J_q}$ and $\mathbb{R}^{\boldsymbol{J}}$ are isomorphic as vector spaces.

Using the standard inner product on $\mathbb{R}^{\boldsymbol{J}}$, i.e.,

$$\left\langle \boldsymbol{y}^J, \boldsymbol{w}^J \right\rangle_{\boldsymbol{J}} = \sum_{\boldsymbol{j}\in\boldsymbol{J}} y_{\boldsymbol{j}} w_{\boldsymbol{j}}, \text{ where } \boldsymbol{y}^J = \{y_{\boldsymbol{j}}\}_{\boldsymbol{j}\in\boldsymbol{J}} \text{ and } \boldsymbol{w}^J = \{w_{\boldsymbol{j}}\}_{\boldsymbol{j}\in\boldsymbol{J}},$$

we can further treat the basis $\boldsymbol{E}^{\boldsymbol{J}}_{\boldsymbol{j}} \cong \otimes_t(v^1_{j_1}, v^2_{j_2}, \cdots, v^q_{j_q})$ as an orthonormal basis. Thus, the tensor product space $\mathbb{R}^{J_1} \otimes \cdots \otimes \mathbb{R}^{J_q}$ and $\mathbb{R}^{\boldsymbol{J}}$ are isomorphic as the inner product space.

4.2 Unique Parametrization of Multinomial Loss

Lemma 4.1. *Suppose* $\ell_{\text{Mult}}(\boldsymbol{\theta x}, \boldsymbol{y})$ *is the multinomial negative log-likelihood corresponding to category probabilities* $\frac{e^{\boldsymbol{\theta x}}}{\left\langle \boldsymbol{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta x}}\right\rangle}$ *with parameters* $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|\times p}$.

Then the following are equivalent:

1. $\ell_{\text{Mult}}(\boldsymbol{\theta x}, \boldsymbol{y}) = \ell_{\text{Mult}}(\boldsymbol{\theta}'\boldsymbol{x}, \boldsymbol{y})$ *for all feasible multinomial outcomes* $\boldsymbol{y}$ *and for all* $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p$ *such that* $\operatorname{span}(\mathcal{X}) = \mathbb{R}^p$*;*

2. $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta} = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}'$.

Proof of Lemma 4.1. Notice that the multinomial negative log-likelihood $\ell_{\text{Mult}}(\boldsymbol{\theta x}, \boldsymbol{y})$ is uniquely determined by the category probabilities $\frac{e^{\boldsymbol{\theta x}}}{\left\langle \boldsymbol{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta x}}\right\rangle}$. Hence, for all feasible multinomial outcomes $\boldsymbol{y}$,

$$\ell_{\text{Mult}}(\boldsymbol{\theta x}, \boldsymbol{y}) = \ell_{\text{Mult}}(\boldsymbol{\theta}'\boldsymbol{x}, \boldsymbol{y}) \quad \text{if and only if} \quad \frac{e^{\boldsymbol{\theta x}}}{\left\langle \boldsymbol{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta x}}\right\rangle} = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{\left\langle \boldsymbol{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta}'\boldsymbol{x}}\right\rangle}.$$

Because the category probabilities sum to 1, the above equality holds if and only if $e^{\boldsymbol{\theta}\boldsymbol{x}-\boldsymbol{\theta}'\boldsymbol{x}} = C(x)\mathbf{1}_{|\boldsymbol{J}|}$, i.e.

$$\boldsymbol{\theta}\boldsymbol{x} - \boldsymbol{\theta}'\boldsymbol{x} = \log(C(x))\mathbf{1}_{|\boldsymbol{J}|},$$

which is equivalent to $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}\boldsymbol{x} = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}'\boldsymbol{x}$.

Thus, the first statement in Lemma 4.1 holds if and only if, for any $\boldsymbol{x} \in \mathcal{X}$ such that $\operatorname{span}(\mathcal{X}) = \mathbb{R}^p$,

$$(\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta} - \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}')\boldsymbol{x} = \mathbf{0},$$

and this is equivalent to the second statement in Lemma 4.1. □

4.3 Connection Between Multivariate Categorical Response Regression and Hypergraph Model

Let $H = (V, E)$ be a hypergraph with vertex set $V = [q]$ and hyperedge set E, where each hyperedge $e \in E$ is a subset of V (containing at least two vertices). Assume each vertex corresponds to a discrete state variable $j_i \in [J_i]$. The generalized Hamiltonian is then defined as

$$H(\boldsymbol{j}) = \sum_{e\in E} J_e\big(\{j_i\}_{i\in e}\big) + \sum_{i\in V} h_i(j_i),$$

where $J_e\big(\{j_i\}_{i\in e}\big)$ is an interaction function that depends on the configuration of state variables within hyperedge e, and $h_i(j_i)$ represents the local field acting on vertex i.

Now, we will connect the generalized Hamiltonian above with our reparameterization. For the reparameterization

$$\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in E} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} + \sum_{i\in[q]} \boldsymbol{H}_{\{i\}}\boldsymbol{\beta}_{\{i\}} \in \mathbb{R}^{|\boldsymbol{J}|},$$

where for each $\boldsymbol{k} = \{k_1, \cdots, k_s\}$ with $k_1 < \cdots < k_s$ we have $\boldsymbol{\beta}_{\boldsymbol{k}} \in \mathbb{R}^{\prod_{i=1}^{s}(J_{k_i}-1)}$, the vector $\boldsymbol{\theta}$ is related to the Hamiltonian $H(\boldsymbol{j})$. The identity is given by

$$\{\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})\}_{\boldsymbol{j}} = H(\boldsymbol{j}),$$

because

$$\begin{aligned}
&\{\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})\}_{\boldsymbol{j}} \\
=&\sum_{\boldsymbol{k}\in E} \frac{1}{\sqrt{|\boldsymbol{J}_{-\boldsymbol{k}}|}}\, (\boldsymbol{U}_{J_{k_s}})_{j_{k_s}} \otimes \cdots \otimes (\boldsymbol{U}_{J_{k_1}})_{j_{k_1}}\, \boldsymbol{\beta}_{\boldsymbol{k}} + \sum_{i\in[q]} \frac{1}{\sqrt{|\boldsymbol{J}_{-\{i\}}|}}\, (\boldsymbol{U}_{J_i})_{j_i}\, \boldsymbol{\beta}_{\{i\}}, \\
&J_e\big(\{j_i\}_{i\in e}\big) = \tfrac{1}{\sqrt{|\boldsymbol{J}_{-\boldsymbol{k}}|}}(\boldsymbol{U}_{J_{k_s}})_{j_{k_s}} \otimes \cdots \otimes (\boldsymbol{U}_{J_{k_1}})_{j_{k_1}}\boldsymbol{\beta}_{\boldsymbol{k}}, \\
&h_i(j_i) = \tfrac{1}{\sqrt{|\boldsymbol{J}_{-\{i\}}|}}(\boldsymbol{U}_{J_i})_{j_i}\boldsymbol{\beta}_{\{i\}},
\end{aligned}$$

where $\{\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})\}_{\boldsymbol{j}}$ denotes the entry corresponding to $\boldsymbol{j} = (j_1, \cdots, j_q)$ in the tensor $\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})$, $(\boldsymbol{U}_{J_i})_{j_i}$ denotes the j_i-th row of the matrix $\boldsymbol{U}_{J_i}$, and $|\boldsymbol{J}_{-\boldsymbol{k}}| = |\boldsymbol{J}|/\prod_{i=1}^{s} k_i$.

4.4 Computational Details

4.4.1 Tuning parameter selection

The hyperparameters that need to be selected include the regularization parameter schedule $\{\lambda_1, \cdots, \lambda_{n_\lambda}\}$, and the initial backtracking learning rate η.

Let $n_\lambda = 100$, $\mathrm{ratio}_\lambda \in (0, 1)$, $d \in \{1, 2, \cdots, q\}$, $\gamma \in (0, 1)$. Set the initial step size $\eta = n/\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X})$ for the Poisson categorical response model and $\eta =$

$n|\boldsymbol{J}|/\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X})$ for the multinomial categorical response model. The initial step size η is suggested by (4.27) and (4.28) in Lemma 4.7.

Let λ_1 be a value such that $\widehat{\boldsymbol{\theta}}_{\lambda_1} = 0$. Let $\lambda_{i+1} = \text{ratio}_\lambda \lambda_i, 1 \leq i \leq n_\lambda - 1$. When $\lambda = \lambda_1$, we require that $\widehat{\boldsymbol{\theta}}_{\lambda_1} = 0$.

Recall that $\mathcal{G} = \{(\boldsymbol{k}, j) : \boldsymbol{k} \in \mathcal{K}, j \in [t]\}$. Define the overlapping group $\mathcal{D}(\mathcal{G})$, such that $g \in \mathcal{D}(\mathcal{G})$ if and only if there exists $(\boldsymbol{k}, j) \in \mathcal{G}$ such that $g = \{(\boldsymbol{k}', j) \in \mathcal{G}; \boldsymbol{k} \subset \boldsymbol{k}'\}$. Then, we can rewrite the hierarchical group lasso as $\Omega^H_{\mathcal{G}}(\boldsymbol{\beta}) = \Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta})$.

For both group lasso penalty $\Omega_{\mathcal{G}}(\boldsymbol{\beta})$ and overlapping group lasso penalty $\Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta})$, we can set the maximum regularization parameter $\lambda_1 = \max_{(\boldsymbol{k},j)\in\mathcal{G}} \left\| \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \right\| / w_{\boldsymbol{k},j}$, which is suggested by Lemma 4.8.

4.4.2 Algorithm formulation

The accelerated proximal gradient descent algorithm with backtracking is presented in Algorithm 1.

Algorithm 1: Accelerated proximal gradient descent algorithm

1 **for** $i \in \{2, 3, \ldots, n_\lambda\}$ **do**

2 Set $k = 0$, $\widehat{\boldsymbol{\beta}}_\eta^0 = \widehat{\boldsymbol{\beta}}_\eta^{-1} = \widehat{\boldsymbol{\beta}}_{\lambda_{i-1}}$;

3 $\boldsymbol{z}^{k+1} = \widehat{\boldsymbol{\beta}}_\eta^k + \frac{k}{k+3}\left(\widehat{\boldsymbol{\beta}}_\eta^k - \widehat{\boldsymbol{\beta}}_\eta^{k-1}\right)$;

4 $\widehat{\boldsymbol{\beta}}_\eta^{k+1} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{z}^{k+1} - \eta\nabla\mathcal{L}_n(\boldsymbol{z}^{k+1})\right)\right\|^2 + \lambda_i\eta\cdot\Omega(\boldsymbol{\beta})$;

5 **while** $\mathcal{L}_n(\widehat{\boldsymbol{\beta}}_\eta^{k+1}) > \mathcal{L}_n(\boldsymbol{z}^{k+1}) + \left\langle \nabla\mathcal{L}_n(\boldsymbol{z}^{k+1}), \widehat{\boldsymbol{\beta}}_\eta^{k+1} - \boldsymbol{z}^{k+1}\right\rangle + \frac{1}{2\eta}\left\|\widehat{\boldsymbol{\beta}}_\eta^{k+1} - \boldsymbol{z}^{k+1}\right\|^2$ **do**

6 Shrink $\eta \leftarrow \gamma\eta$;

7 **if** $\widehat{\boldsymbol{\beta}}_\eta^{k+1}$ *has not converged* **then**

8 Set $k \leftarrow k + 1$ and return to step 3 ;

9 $\widehat{\boldsymbol{\beta}}_{\lambda_i} \leftarrow \widehat{\boldsymbol{\beta}}_\eta^{k+1}$, $\quad \widehat{\boldsymbol{\theta}}_{\lambda_i} \leftarrow \boldsymbol{H}\widehat{\boldsymbol{\beta}}_{\lambda_i}$;

10 **Output:** $\widehat{\boldsymbol{\beta}}_{\lambda_i}$ and $\widehat{\boldsymbol{\theta}}_{\lambda_i}$ for any $1 \le i \le n_\lambda$.

For the group lasso penalty (2.14), we replace line 4 in Algorithm 1 with the following line.

Algorithm 2: Proximal operator for nonoverlapping $\mathcal{G}$ (Step 4 of Algorithm 1)

for $g \in \mathcal{G}$ **do**

$\widehat{\boldsymbol{\beta}}_{g,\eta}^{k+1} \leftarrow \max\left(1 - \frac{\lambda_i\eta w_g}{\left\|\widehat{\boldsymbol{z}}_g^{k+1} - \eta\frac{\partial}{\partial\boldsymbol{\beta}_g}\mathcal{L}_n(\widehat{\boldsymbol{z}}^{k+1})\right\|}, 0\right)\left(\widehat{\boldsymbol{z}}_g^{k+1} - \eta\frac{\partial}{\partial\boldsymbol{\beta}_g}\mathcal{L}_n(\widehat{\boldsymbol{z}}^{k+1})\right)$.

To solve the proximal problem for the overlapping group lasso penalty, as defined in (2.17), we adopt the methodology outlined in Jenatton et al. (2011), using block coordinate descent. We modify line 4 of Algorithm 1 accordingly.

Algorithm 3: Proximal operator for overlapping $\mathcal{G}$ (Step 4 of Algorithm 1)

1 Initialize: $\boldsymbol{\zeta}^0 = \widehat{\boldsymbol{z}}^{k+1} - \eta \nabla \mathcal{L}_n(\widehat{\boldsymbol{z}}^{k+1})$, $j = 0$, $\boldsymbol{\xi}^{d(g)} = \mathbf{0}, \forall g \in \mathcal{G}$;

2 **repeat**

3   Set $\boldsymbol{\zeta}^{j+1} = \boldsymbol{\zeta}^{j}$;

4   **for** $g \in \mathcal{G}$ **do**

5         $\boldsymbol{\zeta}^{j+1} \leftarrow \boldsymbol{\zeta}^{j+1} + \boldsymbol{\xi}^{d(g)}$;

6         $\boldsymbol{\xi}^{d(g)} \leftarrow \mathbf{\Pi}_{\lambda w_g}(\boldsymbol{\zeta}^{j+1})$;

7         $\boldsymbol{\zeta}^{j+1} \leftarrow \boldsymbol{\zeta}^{j+1} - \boldsymbol{\xi}^{d(g)}$;

8   **if** $\boldsymbol{\zeta}^{j+1}$ *has not converged* **then**

9   Set $j \leftarrow j + 1$ and return to step 2 ;

10 **until** *convergence of* $\boldsymbol{\zeta}^{j+1}$;

11 Set $\widehat{\boldsymbol{\beta}}_{\eta}^{k+1} = \boldsymbol{\zeta}^{j+1}$.

Here, $\mathbf{\Pi}_{\rho}(\cdot)$ denotes the orthogonal projection onto the ball of radius ρ, descendants of $g = (\boldsymbol{k}, j) \in \mathcal{G}$ is given by $d(g) = \{(\boldsymbol{k}', j) \in \mathcal{G}; \boldsymbol{k} \subset \boldsymbol{k}'\}$, and $\{\boldsymbol{\xi}^{d(g)}\}_{g \in \mathcal{G}}$ denote matrices of compatible size.

The value of η in Algorithms 1, 2, and 3 is suggested by Lemma 4.7. Similarly, the value of λ_1 in these algorithms is suggested by Lemma 4.8.

Applying Theorem 4.4 in Beck and Teboulle (2009), we obtain that the sequence of objective function values at the iterates, generated by an accelerated version of Algorithms 1 and 2, converges to the optimal value at a rate of $O(1/t^2)$, when using the backtracking line search step size η.

4.5 Detailed Setup for Simulation

We consider three distinct structures for $\boldsymbol{\beta}^*$, each corresponding to one of the three generation schemes. Let $m_0 = 1$. We first randomly select two additional elements

$\{m_1, m_2\}$ from $\{2, \cdots, p\}$.

Scheme 1: (Mutual independence) Each element of $\boldsymbol{\beta}^*_{\boldsymbol{k},m_l}$, for any $\boldsymbol{k}$ such that $\|\boldsymbol{k}\|_0 = 1$ and $l \in \{0, 1, 2\}$, is sampled from Uniform$(-2, -1) \cup (1, 2)$. For other cases, set $\boldsymbol{\beta}^*_{\boldsymbol{k},j} = \mathbf{0}$. This model coincides with the mutual independence model in Example 2.2, with

$$\pi_{j_1,j_2,j_3,j_4}(\boldsymbol{x}) = \pi_{j_1,+,+,+}(\boldsymbol{x})\pi_{+,j_2,+,+}(\boldsymbol{x})\pi_{+,+,j_3,+}(\boldsymbol{x})\pi_{+,+,+,j_4}(\boldsymbol{x}).$$

Scheme 2: (Joint independence) Each element of $\boldsymbol{\beta}^*_{\boldsymbol{k},m_l}$ is sampled from Uniform$(-2, -1) \cup (1, 2)$ for all $\boldsymbol{k} \in \Big\{\{1\}, \{2\}, \{3\}, \{4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{2, 3, 4\}\Big\}$ and $l \in \{0, 1, 2\}$. For other cases, set $\boldsymbol{\beta}^*_{\boldsymbol{k},j} = \mathbf{0}$. This model coincides with the joint independence model in Example 2.2, with

$$\pi_{j_1,j_2,j_3,j_4}(\boldsymbol{x}) = \pi_{j_1,+,+,+}(\boldsymbol{x}) \cdot \pi_{+,j_2,j_3,j_4}(\boldsymbol{x}), \text{ for all } (j_1, j_2, j_3, j_4) \in \boldsymbol{J}.$$

Scheme 3: (Conditional independence) Each element of $\boldsymbol{\beta}^*_{\boldsymbol{k},m_l}$ is sampled from Uniform $(-2, -1) \cup (1, 2)$ for all

$$\boldsymbol{k} \in \Big\{\{1\}, \{2\}, \{3\}, \{4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{2, 3, 4\}, \{1, 4\}\Big\}$$

and $l \in \{0, 1, 2\}$. For other cases, set $\boldsymbol{\beta}^*_{\boldsymbol{k},j} = \mathbf{0}$. This model coincides with the conditional independence model in Example 2.2, with

$$\pi_{j_1,j_2,j_3|j_4}(\boldsymbol{x}) = \pi_{j_1,+,+|j_4}(\boldsymbol{x}) \cdot \pi_{+,j_2,j_3|j_4}(\boldsymbol{x}), \text{ for all } (j_1, j_2, j_3, j_4) \in \boldsymbol{J}.$$

Candidate estimators. We consider the following 9 estimators.

1. Overlapping group-penalized multinomial categorical response model (O-Mult):

The overlapping group-penalized multinomial estimator is given by

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{D}(\mathcal{G}_{\text{local}})}(\boldsymbol{\beta}).$$

2. Overlapping group-penalized Poisson categorical response model (O-Pois): The overlapping group-penalized Poisson estimator is given by

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\text{Pois}}(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}_{\mathcal{D}(\mathcal{G}_{\text{local}})}}(\boldsymbol{\beta}).$$

3. Lasso-penalized multinomial categorical response model (L-Mult): The (largest) group-penalized multinomial estimator is given by

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}_{\text{local}}}(\boldsymbol{\beta}).$$

4. Lasso-penalized Poisson categorical response model (L-Pois): The (largest) group-penalized Poisson estimator is given by

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\text{Pois}}(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}_{\text{local}}}(\boldsymbol{\beta}).$$

5. Group-penalized multinomial categorical response model (G-Mult): The (smallest) group-penalized multinomial estimator is given by

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}_{\text{global}}}(\boldsymbol{\beta}).$$

6. Group-penalized Poisson categorical response model (G-Pois): The (smallest) group-penalized Poisson estimator is given by

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\text{Pois}}(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}_{\text{global}}}(\boldsymbol{\beta}).$$

For the above 6 estimators, set $\widehat{\boldsymbol{\theta}} = \boldsymbol{H}\widehat{\boldsymbol{\beta}}$ and assign $w_{\{0\},0} = 0, w_{\{1\},0} = 0, \ldots, w_{\{4\},0} = 0$, and set $w_{\boldsymbol{k},j} = 1$ otherwise, for all group lasso penalties described above.

7. Group-penalized multinomial categorical response model in $\boldsymbol{\theta}$ (G-Mult-$\boldsymbol{\theta}$): The Group-penalized multinomial estimator is given by

$$\arg\min_{\boldsymbol{\theta}} L_n^{\mathrm{Mult}}(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{|\boldsymbol{J}|} \|\boldsymbol{\theta}_i\|_2\,, \text{ where } \boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{\boldsymbol{J}}]^\top.$$

For the above 7 estimators, we define

$$\mathrm{vec}_{\boldsymbol{J}}\left(\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{j_1,\cdots,j_q}(\boldsymbol{x})\right) = \frac{e^{\widehat{\boldsymbol{\theta}}\boldsymbol{x}}}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\widehat{\boldsymbol{\theta}}\boldsymbol{x}}\right\rangle},$$

for multinomial categorical response model, and define

$$\mathrm{vec}_{\boldsymbol{J}}\left(\widehat{\boldsymbol{\mu}}^{\boldsymbol{J}}_{j_1,\cdots,j_q}(\boldsymbol{x})\right) = e^{\widehat{\boldsymbol{\theta}}\boldsymbol{x}}, \text{ and } \mathrm{vec}_{\boldsymbol{J}}\left(\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{j_1,\cdots,j_q}(\boldsymbol{x})\right) = \frac{e^{\widehat{\boldsymbol{\theta}}\boldsymbol{x}}}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\widehat{\boldsymbol{\theta}}\boldsymbol{x}}\right\rangle},$$

for Poisson categorical response model.

8. Separate multinomial (Sep-Mult): We fit the multinomial categorical response model

$$\widehat{\boldsymbol{\eta}}^{(m)} = \arg\min_{\boldsymbol{\eta}^{(m)}} \sum_{i=1}^{n} \frac{1}{n}\Big\{-\big\langle \boldsymbol{y}_i^{J_m}, \boldsymbol{\eta}^{(m)}\boldsymbol{x}_i\big\rangle + n_i \log\Big(\Big\langle \mathbf{1}_{J_m}, e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}_i}\Big\rangle\Big)\Big\} + \lambda^{(m)} \sum_{j=1}^{J_m} \left\|\boldsymbol{\eta}^{(m)}_{j,:}\right\|,$$

for the $m-$th response with $m \in [q]$, where
$\boldsymbol{y}^{J_m} = \{\sum_{j_1=1}^{J_1} \cdots \sum_{j_{m-1}=1}^{J_{m-1}} \sum_{j_{m+1}=1}^{J_{m+1}} \cdots \sum_{j_q=1}^{J_q} y^{\boldsymbol{J}}_{\boldsymbol{j}}\}_{j_m \in J_m}$.
Set $\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{j_1,\cdots,j_q}(\boldsymbol{x}) = \prod_{m=1}^{q} [e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}}]_{j_m} \Big(\sum_{\boldsymbol{j}\in\boldsymbol{J}} \prod_{m=1}^{q} [e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}}]_{j_m}\Big)^{-1}$, where $[e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}}]_{j_m}$ denotes the j_m's element of $e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}}$.

9. True generating model (Oracle): We consider

$$\mathrm{vec}_{\boldsymbol{J}}\left(\boldsymbol{\pi}^{\boldsymbol{J}}_{j_1,\cdots,j_q}(\boldsymbol{x})\right) = \frac{e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}}}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}}\right\rangle}.$$

Tuning criteria. We will tune the parameters based on cross-entropy loss on validation set. Let $\widehat{\boldsymbol{\theta}}_{\lambda} = \boldsymbol{H}\widehat{\boldsymbol{\beta}}_{\lambda}$ denote the first six estimators corresponding to turning parameter λ. Let $n_{\mathrm{val}} = 1000$ represent the size of the validation set, and let Λ denote a finite set of candidate tuning parameters for grid search implementation. Taking the multinomial categorical response model as an example, we will tune the parameters based on cross-entropy loss on validation set, that is

$$\arg\min_{\lambda\in\Lambda} \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} \ell_{\mathrm{Mult}}(\widehat{\boldsymbol{\theta}}_{\lambda}\boldsymbol{x}_i, \boldsymbol{y}_i).$$

Model evaluation. The Hellinger distance and misclassification rate, considered in Section 2.6, are defined as follows:

1. **Hellinger distance** (↓): The Hellinger distance between the true probability mass function $\boldsymbol{\pi}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x})$ and the estimated probability mass function $\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x})$, averaged over $\boldsymbol{x} \sim N_p(\mathbf{0}, \Sigma_X)$, is given by

$$\mathbb{E}_{\boldsymbol{x}\sim N_p(\mathbf{0},\Sigma_X)} H(\boldsymbol{\pi}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x}), \widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x})),$$

where $H(\boldsymbol{\pi}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x}), \widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x})) = \frac{1}{\sqrt{2}} \left\| \sqrt{\boldsymbol{\pi}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x})} - \sqrt{\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}_{\boldsymbol{j}}(\boldsymbol{x})} \right\|$. Generate samples

$$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N_{\mathrm{test}}} \sim N_p(\mathbf{0}, \Sigma_X),$$

and the Monte Carlo estimate of the averaged Hellinger distance is given by

$$\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} H\Big(\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{J}}(\boldsymbol{x}_i), \widehat{\boldsymbol{\pi}}_{\boldsymbol{j}}^{\boldsymbol{J}}(\boldsymbol{x}_i)\Big).$$

2. **Misclassification rate** (↓): Generate i.i.d. samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N_{\text{test}}} \sim N_p(\mathbf{0}, \Sigma_X)$, and generate $\{\boldsymbol{y}_i\}_{i=1}^{N_{\text{test}}}$ according to (2.27). The misclassification rate is given by

$$\frac{\sum_{i=1}^{N_{\text{test}}} I\Big(\arg\max(\boldsymbol{y}_i) \neq \arg\max\big(\text{vec}_{\boldsymbol{J}}(\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}(\boldsymbol{x}_i))\big)\Big)}{N_{\text{test}}},$$

where $\arg\min \boldsymbol{x} = \arg\min_j x_j, \boldsymbol{x} = (x_1, \cdots, x_{|\boldsymbol{J}|})$.

4.6 Proofs of the Isomorphism and Subspace Decomposition Lemmas

In this section, we provide the proofs for Lemma 2.1 and Lemma 2.2. We will use the following simple result from linear algebra.

Lemma 4.2. *For subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$, if (i) $\boldsymbol{U} \subset \boldsymbol{V}$ and (ii) $\dim(\boldsymbol{U}) = \dim(\boldsymbol{V}) < \infty$, then $\boldsymbol{U} = \boldsymbol{V}$.*

4.6.1 Proof of Lemma 2.1

Proof of Lemma 2.1. Let $\boldsymbol{v}_i = (v_i(1), \cdots, v_i(J_i))^\top \in \mathbb{R}^{J_i}$ for any $1 \leq i \leq q$. Due to the bilinearity of tensor product and Kronecker product, we know that

$$\otimes_t(\boldsymbol{v}_q, \cdots, \boldsymbol{v}_1) = \sum_{\boldsymbol{j} \in \boldsymbol{J}} \left\{ \prod_{i=1}^{q} v_i(j_i) \right\} \otimes_t (\boldsymbol{e}_{j_q}^{J_q}, \cdots, \boldsymbol{e}_{j_1}^{J_1})$$

and

$$\boldsymbol{v}_q \otimes \cdots \otimes \boldsymbol{v}_1 = \sum_{\boldsymbol{j} \in \boldsymbol{J}} \left\{ \prod_{i=1}^{q} v_i(j_i) \right\} \boldsymbol{e}_{j_q}^{J_q} \otimes \cdots \otimes \boldsymbol{e}_{j_1}^{J_1}.$$

Because $T_{\boldsymbol{J}}$ is linear, we have

$$\begin{aligned} T_{\boldsymbol{J}}\Big(\otimes_t (\boldsymbol{v}_q, \cdots, \boldsymbol{v}_1)\Big) &= \sum_{\boldsymbol{j} \in \boldsymbol{J}} \left\{ \prod_{i=1}^{q} v_i(j_i) \right\} \cdot T_{\boldsymbol{J}}\Big(\otimes_t (\boldsymbol{e}_{j_q}^{J_q}, \cdots, \boldsymbol{e}_{j_1}^{J_1})\Big) \\ &= \sum_{\boldsymbol{j} \in \boldsymbol{J}} \left\{ \prod_{i=1}^{q} v_i(j_i) \right\} \boldsymbol{e}_{j_q}^{J_q} \otimes \cdots \otimes \boldsymbol{e}_{j_1}^{J_1} = \boldsymbol{v}_q \otimes \cdots \otimes \boldsymbol{v}_1, \end{aligned}$$

which completes the proof of (2.6). Recall that $\mathcal{R}(\boldsymbol{U})$ denotes the column space of the matrix $\boldsymbol{U}$. For any vectors $\boldsymbol{v}_i \in \mathcal{R}(\boldsymbol{V}_i), 1 \leq i \leq q$, there exists $\boldsymbol{\alpha}_i$ such that $\boldsymbol{v}_i = \boldsymbol{V}_i \boldsymbol{\alpha}_i$. Combining identity (2.6), we obtain

$$\begin{aligned} T_{\boldsymbol{J}}\Big(\otimes_t (\boldsymbol{v}_q, \cdots, \boldsymbol{v}_1)\Big) &= \boldsymbol{v}_q \otimes \cdots \otimes \boldsymbol{v}_1 = \boldsymbol{V}_q \boldsymbol{\alpha}_q \otimes \cdots \otimes \boldsymbol{V}_1 \boldsymbol{\alpha}_1 \\ &= (\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1)(\boldsymbol{\alpha}_q \otimes \cdots \otimes \boldsymbol{\alpha}_1) \in \mathcal{R}\Big(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1\Big), \end{aligned}$$

which implies that the basis of $T_{\boldsymbol{J}}(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1))$ is entirely contained in $\mathcal{R}(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1)$. This proves that

$$T_{\boldsymbol{J}}\Big(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)\Big) \subset \mathcal{R}\Big(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1\Big).$$

By Lemma 4.2, to show that $T_{\boldsymbol{J}}(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)) = \mathcal{R}(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1)$, it remains only to show that $\dim T_{\boldsymbol{J}}(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)) = \dim \mathcal{R}(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1)$. To that end, note

$$\dim T_{\boldsymbol{J}}\Big(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)\Big) = \dim\Big(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)\Big) = \prod_{i=1}^{q} \dim\Big(\mathcal{R}(\boldsymbol{V}_i)\Big),$$

and

$$\dim \mathcal{R}\Big(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1\Big) = \operatorname{rank}\Big(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1\Big) = \prod_{i=1}^{q} \operatorname{rank}(\boldsymbol{V}_i) = \prod_{i=1}^{q} \dim\Big(\mathcal{R}(\boldsymbol{V}_i)\Big),$$

which verifies that $\dim T_{\boldsymbol{J}}(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)) = \dim \mathcal{R}(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1)$. □

4.6.2 Proof of Lemma 2.2

Proof of Lemma 2.2. Let $\mathcal{K}' = \cup_{s=0}^{q} \mathcal{K}_s$. Let $\boldsymbol{P_k} = \boldsymbol{H_k}\boldsymbol{H_k}^\top$ for any $\boldsymbol{k} \in \mathcal{K}'$.

To establish the claim that $\mathbb{R}^{|\boldsymbol{J}|} = \otimes_{\boldsymbol{k}\in\mathcal{K}} \mathcal{R}(\boldsymbol{H}_k)$, it suffices to show the following three statements:

1. For any $\boldsymbol{k} \in \mathcal{K}'$, $\boldsymbol{P_k}$ is an orthogonal projection matrix onto $\mathcal{R}(\boldsymbol{H_k})$.

2. For any $\boldsymbol{k}, \boldsymbol{k}' \in \mathcal{K}'$ such that $\boldsymbol{k} \neq \boldsymbol{k}'$, it holds that $\boldsymbol{P_k}\boldsymbol{P_{k'}} = 0$.

3. $\sum_{\boldsymbol{k}\in\mathcal{K}'} \boldsymbol{P_k} = I$.

Because $\boldsymbol{H_k}^\top \boldsymbol{H_k}$ is an identity matrix of order $|\boldsymbol{k}|_{\boldsymbol{J}}$, we know that the columns of $\boldsymbol{H_k}$ are orthonormal. This completes the proof of statement 1.

Due to $\boldsymbol{k} \neq \boldsymbol{k}'$, we can assume without loss of generality that there exists an $l \in \boldsymbol{k}$ and $l \notin \boldsymbol{k}'$, enabling us to rewrite $\boldsymbol{H_k}$ and $\boldsymbol{H_{k'}}$ as

$$\boldsymbol{H_k} = \boldsymbol{A} \otimes \boldsymbol{U}_{J_l} \otimes \boldsymbol{B} \text{ and } \boldsymbol{H_{k'}} = \boldsymbol{A}' \otimes \frac{\boldsymbol{1}_{J_l}}{\sqrt{J_l}} \otimes \boldsymbol{B}',$$

where the number of rows in $\boldsymbol{A}$ and $\boldsymbol{A}'$, as well as in $\boldsymbol{B}$ and $\boldsymbol{B}'$, are the same. Thus,

$$\boldsymbol{H_k}^\top \boldsymbol{H_{k'}} = (\boldsymbol{A}^\top \boldsymbol{A}') \otimes (\frac{1}{\sqrt{J_l}} \boldsymbol{U}_{J_l}^\top \boldsymbol{1}_{J_l}) \otimes (\boldsymbol{B}^\top \boldsymbol{B}') = (\boldsymbol{A}^\top \boldsymbol{A}') \otimes \boldsymbol{0} \otimes (\boldsymbol{B}^\top \boldsymbol{B}') = \boldsymbol{0}, \quad (4.1)$$

which completes the proof of statement 2.

Furthermore, we also know that the columns of $\boldsymbol{H}_{\boldsymbol{k}}$ for any $\boldsymbol{k} \in \mathcal{K}'$ are also orthonormal. By expanding $|\boldsymbol{J}| = \prod_{l=1}^{q}((J_l - 1) + 1)$, we obtain that

$$\begin{aligned}&\sum_{\boldsymbol{k}\in\mathcal{K}'} \dim\Big(\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})\Big)\\ &= \sum_{\boldsymbol{k}\in\mathcal{K}'} |\boldsymbol{k}|_{\boldsymbol{J}} = 1 + \sum_{s=1}^{q} \sum_{(k_1,\cdots,k_s)\in\mathcal{K}'} (J_{k_1} - 1)\cdots(J_{k_s} - 1) = \prod_{l=1}^{q}\Big((J_l - 1) + 1\Big).\end{aligned}$$

Combined with (4.1), we further obtain that

$$\dim\Big(\sum_{\boldsymbol{k}\in\mathcal{K}'} \mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})\Big) = \sum_{\boldsymbol{k}\in\mathcal{K}'} \dim\Big(\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})\Big) = |\boldsymbol{J}|.$$

In summary, the columns of $\boldsymbol{H}_{\boldsymbol{k}}$ for any $\boldsymbol{k} \in \mathcal{K}'$ form an orthonormal basis of $\mathbb{R}^{|\boldsymbol{J}|}$, so by Lemma 4.2 which completes the proof of statement 3.

As a consequence of statements 1–3, for any $\boldsymbol{y} \in \mathbb{R}^{|\boldsymbol{J}|}$, we have

(i) $\boldsymbol{y} = I\boldsymbol{y} = \sum_{\boldsymbol{k}\in\mathcal{K}'} \boldsymbol{P}_{\boldsymbol{k}}\boldsymbol{y}$,

(ii) $\boldsymbol{P}_{\boldsymbol{k}}\boldsymbol{y} \in \mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})$ for any $\boldsymbol{k} \in \mathcal{K}'$,

(iii) $\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}}) \cap \Big(\sum_{\boldsymbol{k}'\in\mathcal{K}', \boldsymbol{k}'\neq\boldsymbol{k}} \mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}'})\Big) = \{\boldsymbol{0}\}$ for any $\boldsymbol{k} \in \mathcal{K}'$, and

(iv) $\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}}) \perp \mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}'})$ for any $\boldsymbol{k}, \boldsymbol{k}' \in \mathcal{K}'$ such that $\boldsymbol{k} \neq \boldsymbol{k}'$.

By Theorem 1.5 in Roman et al. (2005), we know that $\mathbb{R}^{|\boldsymbol{J}|}$ is the internal direct sum of the family $\{\mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}})\}_{\boldsymbol{k}\in\mathcal{K}'}$ of subspaces of $\mathbb{R}^{|\boldsymbol{J}|}$. By integrating the definition of orthogonal direct sum (refer to page 194 in Roman et al. (2005)), we finalize the proof of the Lemma 2.2. $\square$

4.7 Proof of Theorem 2.1

In this section, we first prove the inverse vectorization identities (4.2) and (4.3) for tensor reshaping, which plays a crucial role in the proof of Theorem 2.1.

Lemma 4.3. *(Inverse Vectorization Identity for Nonentangled Tensor) Let $\mathcal{I}_1, \cdots, \mathcal{I}_m$ be a disjoint partition of $[q]$. For any $\mathcal{I} = \{i_1, \cdots, i_s\} \subset [q]$ such that $1 \le i_1 < \cdots < i_s \le q$, define $\boldsymbol{j}_{\mathcal{I}} = (j_{i_1}, \cdots j_{i_s})$ in $\boldsymbol{J}^{\mathcal{I}} = [J_{i_1}] \times \cdots \times [J_{i_s}]$ and partitioned tensor product $\otimes_\pi$ with*

$$\boldsymbol{v} = \otimes_\pi(\boldsymbol{v}_{\mathcal{I}_1}, \cdots, \boldsymbol{v}_{\mathcal{I}_m}; \mathcal{I}_1, \cdots, \mathcal{I}_m) = \sum_{\boldsymbol{j} \in \boldsymbol{J}} \left\{ \prod_{l=1}^{m} v_{\mathcal{I}_l}(\boldsymbol{j}_{\mathcal{I}_l}) \right\} \boldsymbol{e}_{j_q}^{J_q} \otimes \cdots \otimes \boldsymbol{e}_{j_1}^{J_1},$$

where

$$\boldsymbol{v}_{\mathcal{I}} = \sum_{\boldsymbol{j}_{\mathcal{I}} \in \boldsymbol{J}^{\mathcal{I}}} v_{\mathcal{I}}(\boldsymbol{j}_{\mathcal{I}}) \cdot \boldsymbol{e}_{j_{i_s}}^{J_{i_s}} \otimes \cdots \otimes \boldsymbol{e}_{j_{i_1}}^{J_{i_1}} \in \mathbb{R}^{|\boldsymbol{J}^{\mathcal{I}}|}.$$

Then

$$\left(\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{v}) \right)_{\boldsymbol{j}} = \prod_{l=1}^{m} \left(\mathrm{vec}_{\boldsymbol{J}^{\mathcal{I}^l}}^{-1}(\boldsymbol{v}_{\mathcal{I}_l}) \right)_{\boldsymbol{j}_{\mathcal{I}_l}}, \tag{4.2}$$

where $(\cdot)_{\boldsymbol{j}}$ denotes the $\boldsymbol{j}$-element of tensor.

Proof of Lemma 4.3. By the definition of $\mathrm{vec}_{\boldsymbol{J}}(\cdot)$, we know that

$$\left(\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{v}) \right)_{\boldsymbol{j}} = \prod_{l=1}^{m} v_{\mathcal{I}_l}(\boldsymbol{j}_{\mathcal{I}_l}).$$

Similarly, we obtain that

$$\left(\mathrm{vec}_{\boldsymbol{J}^{\mathcal{I}}}^{-1}(\boldsymbol{v}_{\mathcal{I}}) \right)_{\boldsymbol{j}_{\mathcal{I}}} = v_{\mathcal{I}}(\boldsymbol{j}_{\mathcal{I}}).$$

Thus, we obtain

$$\left(\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{v})\right)_{\boldsymbol{j}} = \prod_{l=1}^{m} v_{\mathcal{I}_l}(\boldsymbol{j}_{\mathcal{I}_l}) = \prod_{l=1}^{m} \left(\operatorname{vec}_{\boldsymbol{J}^{\mathcal{I}^l}}^{-1}(\boldsymbol{v}_{\mathcal{I}_l})\right)_{\boldsymbol{j}_{\mathcal{I}_l}}.$$

□

Lemma 4.4. *(Inverse Vectorization Identity) Let* $\boldsymbol{J_k} = [J_{i_1}] \times [J_{i_2}] \times \cdots [J_{i_s}]$ *where* $\boldsymbol{k} = \{i_1, i_2, \cdots, i_s\} \subset [q]$ *and* $i_1 < i_2 < \cdots < i_s$*, and let* $\boldsymbol{H\beta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H_k\beta_k}$*, where* $\mathcal{K} = \cup_{s=0}^{d}\mathcal{K}_s$ *and* $1 \le d \le q$*. If we define* $\operatorname{vec}_{\boldsymbol{J_k}}^{-1}$ *such that*

$$\operatorname{vec}_{\boldsymbol{J_k}}\left(\operatorname{vec}_{\boldsymbol{J_k}}^{-1}(\boldsymbol{z})\right) = \boldsymbol{z} \text{ for any } \boldsymbol{z} \in \mathbb{R}^{|\boldsymbol{J_k}|},$$

then the following inverse vectorization identity holds,

$$\left(\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H\beta x})\right)_{j_1,\cdots,j_q} = \frac{1}{\sqrt{|\boldsymbol{J}|}}\boldsymbol{\beta}_0\boldsymbol{x} + \frac{1}{\sqrt{|\boldsymbol{J}|}} \sum_{1\le s\le d}\sum_{\boldsymbol{k}\in\mathcal{K}_s} \left(\operatorname{vec}_{\boldsymbol{J_k}}^{-1}\left(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta_k x}\right)\right)_{j_{i_1},\cdots,j_{i_s}}, \quad (4.3)$$

where $\mathcal{U}_{\boldsymbol{k}} = [\sqrt{J_{i_s}}\boldsymbol{U}_{i_s} \otimes \cdots \otimes \sqrt{J_{i_1}}\boldsymbol{U}_{i_1}]$ *for* $\boldsymbol{k} = \{i_1, \cdots, i_s\} \subset [q]$ *with* $i_1 < \cdots < i_s$.

Proof of Lemma 4.4. Recall the definition of $\boldsymbol{H}$ in (2.13) and $\boldsymbol{H_k}$ in (2.8). Combined with the linearity of $\operatorname{vec}_{\boldsymbol{J}}^{-1}$, we know that

$$\left(\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H\beta x})\right)_{j_1,\cdots,j_q} = \frac{1}{\sqrt{|\boldsymbol{J}|}}\boldsymbol{\beta}_0\boldsymbol{x} + \sum_{1\le s\le d}\sum_{\boldsymbol{k}\in\mathcal{K}_s} \left(\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H_k\beta_k x})\right)_{j_1,\cdots,j_q}.$$

To show the inverse vectorization identity (4.3), it suffices to show that

$$\left(\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H_k\beta_k x})\right)_{j_1,\cdots,j_q} = \frac{1}{\sqrt{|\boldsymbol{J}|}}\left(\operatorname{vec}_{\boldsymbol{J_k}}^{-1}\left(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta_k x}\right)\right)_{j_{i_1},\cdots,j_{i_s}}, (j_1, \cdots, j_q) \in \boldsymbol{J}, \quad (4.4)$$

for any $\boldsymbol{k} = \{i_1, \cdots, i_s\}$, $1 \le i_1 < \cdots < i_s \le q$ and $1 \le s \le q$.

First, suppose $\boldsymbol{k} = \{1, 2, \cdots, s\}$. By (2.8), we know that

$$\begin{aligned}
\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x} &= \frac{1}{\sqrt{|\boldsymbol{J}|}}\Big(\mathbf{1}_{J_q} \otimes \mathbf{1}_{J_{q-1}} \otimes \cdots \otimes \mathbf{1}_{J_{s+1}} \otimes \mathcal{U}_{\boldsymbol{k}}\Big)\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x} \\
&= \frac{1}{\sqrt{|\boldsymbol{J}|}}\Big(\mathbf{1}_{\prod_{i=s+1}^{q} J_i} \otimes \mathcal{U}_{\boldsymbol{k}}\Big)(1 \otimes \boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{J}|}}\mathbf{1}_{\prod_{i=s+1}^{q} J_i} \otimes \Big(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x}\Big).
\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}
\Big(\mathrm{vec}_{\boldsymbol{J}}^{-1}\left(\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x}\right)\Big)_{j_1,\cdots,j_q} &= \frac{1}{\sqrt{|\boldsymbol{J}|}}\Big(\mathrm{vec}_{\boldsymbol{J}}^{-1}\left(\mathbf{1}_{\prod_{i=s+1}^{q} J_i} \otimes (\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)\Big)_{j_1,\cdots,j_q} \\
&= \frac{1}{\sqrt{|\boldsymbol{J}|}}\Big(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}}^{-1}\left(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x}\right)\Big)_{j_1,\cdots,j_s},
\end{aligned}$$

where the last equation holds due to identity (4.2) in Lemma 4.3.

The same result can be established for any other $\boldsymbol{k}$ using nearly identical arguments, which we omit for brevity. This completes the proof of (4.4), thereby completing the proof of Lemma 4.4 as well. □

Utilizing Lemma 4.4, we proceed to prove Theorem 2.1.

Proof of Theorem 2.1. Consider the representation

$$\boldsymbol{\theta} = \boldsymbol{H}\boldsymbol{\beta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} = \sum_{0\le s\le q}\sum_{\boldsymbol{k}\in\mathcal{K}_s} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}, \qquad \boldsymbol{\beta}_{\boldsymbol{k}} \in \mathbb{R}^{|\boldsymbol{k}|_{\boldsymbol{J}}\times p}.$$

Applying the inverse vectorization identity (4.3) in Lemma 4.4, we know that there exists some $B(\boldsymbol{x})$ and $B'(\boldsymbol{x})$ not depending on $\boldsymbol{j}$, such that

$$\begin{aligned}
\pi_{\boldsymbol{j}}(\boldsymbol{x}) &= \exp\{B(\boldsymbol{x})\}\left[\exp\{\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x})\}\right]_{j_1,\cdots,j_q} \\
&= \exp\{B'(\boldsymbol{x})\}\underbrace{\exp\left[\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{1\le s\le q}\sum_{\boldsymbol{k}\in\mathcal{K}_s}\{\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}}^{-1}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\}_{j_{i_1},\cdots,j_{i_s}}\right]}_{=:\xi_{\boldsymbol{j}}(\boldsymbol{x})}.
\end{aligned} \tag{4.5}$$

where $\boldsymbol{k} = \{i_1, \cdots, i_s\}$.

For any $\mathcal{I} = \{i_1, \cdots, i_s\} \subset [q]$ such that $1 \leq i_1 < \cdots < i_s \leq q$, define $\boldsymbol{J}^{\mathcal{I}} = [J_{i_1}] \times \cdots \times [J_{i_s}]$, and $\boldsymbol{J}^{-\mathcal{I}} = \boldsymbol{J}^{[q]\setminus\mathcal{I}}$. Given $\boldsymbol{j}' \in \boldsymbol{J}^{\mathcal{I}}$, $\boldsymbol{j}'' \in \boldsymbol{J}^{-\mathcal{I}}$, define $\sigma(\boldsymbol{j}', \boldsymbol{j}''; \mathcal{I}) \in \boldsymbol{J}$ where σ is the operator such that $[\sigma(\boldsymbol{j}', \boldsymbol{j}''; \mathcal{I})]_{\mathcal{I}} = \boldsymbol{j}'$ and $[\sigma(\boldsymbol{j}', \boldsymbol{j}''; \mathcal{I})]_{[q]\setminus\mathcal{I}} = \boldsymbol{j}''$, where here, $[a]_{\mathcal{S}}$ denotes the subvector of a corresponding to components indexed by $\mathcal{S}$.

Thus, to show (2.9), it suffices to show that

$$\log \xi_{\boldsymbol{j}}(\boldsymbol{x}) = \sum_{l=1}^{m} \log \xi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}) + C(\boldsymbol{x}), \text{ for all } \boldsymbol{j} \in \boldsymbol{J}, \tag{4.6}$$

where

$$\xi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}) = \sum_{\boldsymbol{j}' \in \boldsymbol{J}^{-\mathcal{I}_l}} \xi_{\sigma(\boldsymbol{j}_{\mathcal{I}_l}, \boldsymbol{j}'; \mathcal{I}_l)}(\boldsymbol{x}),$$

and $C(\boldsymbol{x})$ does not depend on $\boldsymbol{j}$.

Mutual and joint independence: Based on the definition of $\mathcal{S}_{\text{joint}}$, we know that

$$\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{S}_{\text{joint}}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}} = \boldsymbol{H}_{\{0\}} \boldsymbol{\beta}_{\{0\}} + \sum_{l=1}^{m} \sum_{\boldsymbol{k} \subset \mathcal{I}_l, \|\boldsymbol{k}\|_0 \geq 1} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}.$$

Similar to (4.5), we obtain that

$$\xi_{\boldsymbol{j}}(\boldsymbol{x}) = \prod_{l=1}^{m} \exp\Big(\frac{1}{\sqrt{|\boldsymbol{J}|}} \sum_{\boldsymbol{k} \subset \mathcal{I}_l, \|\boldsymbol{k}\|_0 \geq 1} \Big(\text{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}} \boldsymbol{x}) \Big)_{j_{i_1}, \cdots, j_{i_s}} \Big).$$

Based on the above equation and the definition of $\xi_{\boldsymbol{j}_{\mathcal{I}_l},+}$, we know that

$$\log \xi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{J}|}} \Big(\sum_{\boldsymbol{k} \subset \mathcal{I}_l, \|\boldsymbol{k}\|_0 \geq 1} \Big(\text{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}} \boldsymbol{x}) \Big)_{j_{i_1}, \cdots, j_{i_s}} \Big) + C_l(\boldsymbol{x}),$$

where

$$C_l(\boldsymbol{x}) = \log\left(\sum_{\boldsymbol{j}'\in\boldsymbol{J}^{-\mathcal{I}_l}}\prod_{l'\in[m]:l'\neq l}\exp\left(\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\boldsymbol{k}\subset\mathcal{I}_{l'},\|\boldsymbol{k}\|_0\geq 1}\left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j'_{i_1},\cdots,j'_{i_s}}\right)\right)$$

does not depend on $\boldsymbol{j}$. In conclusion, we obtain that

$$\begin{aligned}&\log\xi_{\boldsymbol{j}}(\boldsymbol{x})\\ =&\sum_{l=1}^{m}\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\boldsymbol{k}\subset\mathcal{I}_l,\|\boldsymbol{k}\|_0\geq 1}\left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j_{i_1},\cdots,j_{i_s}} = \sum_{l=1}^{m}\log\xi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}) - \sum_{l=1}^{m}C_l(\boldsymbol{x}),\end{aligned}$$

which implies (4.6), by setting $C(\boldsymbol{x}) = -\sum_{l=1}^{m}C_l(\boldsymbol{x})$. Thus, we complete the proof of (2.9).

Conditional independence: Based on the definition of $\mathcal{S}_{\mathrm{joint}|\mathcal{I}_{\mathrm{m}}}$, we know that

$$\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{S}_{\mathrm{joint}|\mathcal{I}_{\mathrm{m}}}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} = \boldsymbol{H}_{\{0\}}\boldsymbol{\beta}_{\{0\}} + \sum_{l=1}^{m-1}\sum_{\substack{|\boldsymbol{k}\cap\mathcal{I}_l|>0\\ \boldsymbol{k}\subset\mathcal{I}_l\cup\mathcal{I}_m}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} + \sum_{\boldsymbol{k}\subset\mathcal{I}_m,\|\boldsymbol{k}\|_0\geq 1}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}.$$

By (4.5), we obtain that

$$\begin{aligned}\xi_{\boldsymbol{j}}(\boldsymbol{x}) = \prod_{l=1}^{m-1}&\exp\left(\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\substack{|\boldsymbol{k}\cap\mathcal{I}_l|>0\\ \boldsymbol{k}\subset\mathcal{I}_l\cup\mathcal{I}_m}}\left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j_{i_1},\cdots,j_{i_s}}\right)\\ &\cdot\exp\left(\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\boldsymbol{k}\subset\mathcal{I}_m,\|\boldsymbol{k}\|_0\geq 1}\left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j_{i_1},\cdots,j_{i_s}}\right).\end{aligned} \tag{4.7}$$

By definition, $\xi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},+}(\boldsymbol{x}) = \sum_{\boldsymbol{j}'\in\boldsymbol{J}^{-\mathcal{I}_l\cup\mathcal{I}_m}}\xi_{\sigma(\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},\boldsymbol{j}';\mathcal{I}_l\cup\mathcal{I}_m)}(\boldsymbol{x})$ so that based on the above equation, we have

$$\log\xi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},+}(\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\substack{|\boldsymbol{k}\cap\mathcal{I}_l|>0\\ \boldsymbol{k}\subset\mathcal{I}_l\cup\mathcal{I}_m}}\left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j_{i_1},\cdots,j_{i_s}} + C_{l,\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}), \tag{4.8}$$

where

$$
\begin{aligned}
& C_{l,\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) \\
&= \log\left(\sum_{\boldsymbol{j}'\in \boldsymbol{J}_{\mathcal{I}_m}^{-\mathcal{I}_l\cup\mathcal{I}_m}} \prod_{1\le l'\le m-1, l'\neq l} \exp\left(\frac{1}{\sqrt{|\boldsymbol{J}|}} \sum_{\substack{|\boldsymbol{k}\cap\mathcal{I}_{l'}|>0\\ \boldsymbol{k}\subset\mathcal{I}_{l'}\cup\mathcal{I}_m}} \left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j'_{i_1},\cdots,j'_{i_s}}\right)\right) \\
&\quad + \frac{1}{\sqrt{|\boldsymbol{J}|}} \sum_{\boldsymbol{k}\subset\mathcal{I}_m, \|\boldsymbol{k}\|_0\ge 1} \left(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j_{i_1},\cdots,j_{i_s}},
\end{aligned}
$$

with

$$
\boldsymbol{J}_{\mathcal{I}_m}^{-\mathcal{I}_l\cup\mathcal{I}_m} = \times_{i\in[q]\setminus\mathcal{I}_l}\boldsymbol{W}_i, \qquad \boldsymbol{W}_i = \begin{cases} [J_i] & : i\in[q]\setminus(\mathcal{I}_l\cup\mathcal{I}_m) \\ \{j_i\} & : i\in\mathcal{I}_m \end{cases} .
$$

Evidently, $C_{l,\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x})$ does not depend on $\boldsymbol{j}_{\mathcal{I}_1},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}$. Note that

$$
\log\pi_{\boldsymbol{j}_{\mathcal{I}_1},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \log\pi_{\boldsymbol{j}}(\boldsymbol{x}) - \log\pi_{\boldsymbol{j}_{\mathcal{I}_m}+}(\boldsymbol{x}) = \log\xi_{\boldsymbol{j}}(\boldsymbol{x}) + C_{\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}), \tag{4.9}
$$

where $C_{\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x})$ only depends on $\boldsymbol{j}_{\mathcal{I}_m}$ and $\boldsymbol{x}$.

We can also show that

$$
\begin{aligned}
\log\pi_{\boldsymbol{j}_{\mathcal{I}_l}+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) &= \log\pi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m}+}(\boldsymbol{x}) - \log\pi_{\boldsymbol{j}_{\mathcal{I}_m}+}(\boldsymbol{x}) \\
&= \log\xi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},+}(\boldsymbol{x}) + C'_{\mathcal{I}_l}(\boldsymbol{x}),
\end{aligned} \tag{4.10}
$$

where for any $1\le l\le m-1$, $C'_{\mathcal{I}_l}(\boldsymbol{x})$ does not depend on $\boldsymbol{j}_{\mathcal{I}_1},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}$.

Combining (4.7), (4.8), (4.9) and (4.10), we obtain that

$$
\log\pi_{\boldsymbol{j}_{\mathcal{I}_1},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) - \sum_{l=1}^{m-1}\log\pi_{\boldsymbol{j}_{\mathcal{I}_l}+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \log\xi_{\boldsymbol{j}}(\boldsymbol{x}) - \sum_{l=1}^{m-1}\log\xi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},+}(\boldsymbol{x}) + C''_{\mathcal{I}_m}(\boldsymbol{x})
$$

does not depend on $\boldsymbol{j}_{\mathcal{I}_1},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}$.

Note that

$$\pi_{\boldsymbol{j}_{\mathcal{I}_1},\boldsymbol{j}_{\mathcal{I}_2},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \prod_{l=1}^{m-1} \pi_{\boldsymbol{j}_{\mathcal{I}_l},+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}), \boldsymbol{j} \in \boldsymbol{J}$$

if and only if

$$\pi_{\boldsymbol{j}_{\mathcal{I}_1},\boldsymbol{j}_{\mathcal{I}_2},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = A(\boldsymbol{x}, \boldsymbol{j}_{\mathcal{I}_m}) \prod_{l=1}^{m-1} \pi_{\boldsymbol{j}_{\mathcal{I}_l},+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}), \boldsymbol{j} \in \boldsymbol{J},$$

where $A(\boldsymbol{x}, \boldsymbol{j}_{\mathcal{I}_m})$ is a function that does not depend on $\boldsymbol{j}_{\mathcal{I}_1}, \cdots, \boldsymbol{j}_{\mathcal{I}_{m-1}}$. This completes the proof of (2.10). □

4.8 Proof of Theorem 2.2

To complete the proof of Theorem 2.2, we will show the following error bound for the group lasso without assuming $w_g \equiv 1$.

Theorem 4.1. *Let B, B_1, B_2 and B' be some positive absolute constants. Let $\lambda_\delta = BC\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\big), \delta \geq 0$ for multinomial categorical response model, and let $\lambda_\delta = BC\sqrt{\Lambda}\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\big), 0 \leq \delta \leq B_2$ for Poisson categorical response model. Assume $d^* \leq d$. Under Assumptions 1-4, if $\frac{m}{n} + \frac{\log|\mathcal{G}|}{n} \leq B_1 \cdot \min\big(1, \frac{\kappa}{AC^2\Psi^2(\mathcal{S})}\big)$ and $\lambda_\delta \leq \frac{R\cdot\kappa}{6\Psi(\mathcal{S})}$, then*

$$\left\|\boldsymbol{\theta}^\dagger - \widehat{\boldsymbol{\theta}}\right\| = \left\|\boldsymbol{\beta}^\dagger - \widehat{\boldsymbol{\beta}}_0\right\| \leq \frac{6\lambda_\delta\Psi(\mathcal{S})}{\kappa},$$

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

Consequently, Theorem 2.2 can be viewed as a specific case of Theorem 4.1.

Proof of Theorem 2.2. Applying Theorem 4.1 and assuming $w_g \equiv 1$, we complete the proof of Theorem 2.2. □

4.9 Justification of the Restricted Strong Convexity Condition

In this section, we provide theoretical justification for Assumption 4 by assuming a multivariate GLM model and some mild regularity conditions on the predictors. The justification is based on the generalization of the results from Section 9 of Wainwright (2019).

We first state the following multivariate GLM assumption, which includes multinomial and Poisson categorical response models as special cases.

Assumption 4.1. Consider the multivariate GLM model: conditionally on $\boldsymbol{x}_i$, each response $\boldsymbol{y}_i$ is independent according to a conditional distribution of the following form:

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = h(\boldsymbol{x}, \boldsymbol{y}) \exp\Big\{ \langle \boldsymbol{y}, \boldsymbol{\theta}\boldsymbol{x} \rangle - \psi(\boldsymbol{\theta}\boldsymbol{x}) \Big\}, \quad \boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}| \times p}.$$

Assume there exists $\mathcal{V} \subset \mathbb{R}^{|\boldsymbol{J}| \times p}$ such that $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \mathbb{P}_{\boldsymbol{\theta}'}(\boldsymbol{y}|\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y}$ if and only if $\boldsymbol{P}_{\mathcal{V}}(\boldsymbol{\theta}) = \boldsymbol{P}_{\mathcal{V}}(\boldsymbol{\theta}')$. Assume the hypothesis space $\mathcal{M}_H \subset \mathbb{R}^{|\boldsymbol{J}| \times p}$, such that for any $M > 0$,

$$\gamma_M = \inf_{\substack{\|\boldsymbol{\theta}\boldsymbol{x}\| \leq M \\ \boldsymbol{\theta} \in \mathcal{M}_H}} \inf_{\substack{\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x} \neq 0 \\ \Delta\boldsymbol{\theta} \in \mathcal{M}_H}} \frac{\langle \nabla^2 \psi(\boldsymbol{\theta}\boldsymbol{x}) \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x} \rangle}{\|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}\|^2} > 0. \tag{4.11}$$

Define $L_n(\boldsymbol{\theta}) = -\sum_{i=1}^n \langle \boldsymbol{y}_i, \boldsymbol{\theta}\boldsymbol{x}_i \rangle - \psi(\boldsymbol{\theta}\boldsymbol{x}_i)$ and $\mathcal{E}_n(\Delta\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}^\dagger + \Delta\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}^\dagger) - \left\langle \nabla L_n(\boldsymbol{\theta}^\dagger), \Delta\boldsymbol{\theta} \right\rangle$.

The next theorem generalizes Theorem 9.36 in Wainwright (2019) to the multivariate GLM model.

Theorem 4.2. *Let $\mathcal{M}_H \subset \mathbb{R}^{|\boldsymbol{J}| \times p}$ denote the largest hypothesis space considered for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\dagger$. Assume the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn i.i.d. from a zero-mean distribution* *such that, for some positive constants (α, β), we have*

$$\mathbb{E}\left\|\Delta\boldsymbol{\theta}\boldsymbol{x}_i\right\|^2 \geq \alpha \text{ and } \mathbb{E}\left\|\Delta\boldsymbol{\theta}\boldsymbol{x}_i\right\|^4 \leq \beta, \tag{4.12}$$

for all vector $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$ such that $\|\Delta\boldsymbol{\theta}\| = 1$. Assume that a norm Φ defined in $\mathcal{M}_H$ satisfies $\Phi(\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta})$, for any $\boldsymbol{\theta} \in \mathcal{M}_H$. Under Assumption 4.1, for any $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$, we have

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) \geq \frac{\kappa}{2}\left\|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\right\|^2 - c_0 \cdot \mu_n^2(\Phi, \mathcal{M}_H) \cdot \Phi^2(\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}), \|\Delta\boldsymbol{\theta}\| \leq R, R > 0, \tag{4.13}$$

with probability at least $1 - c_1 e^{-c_2 n}$, where

$$\begin{aligned}\mu_n(\Phi, \mathcal{M}_H) &= \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i 1 \leq i \leq n} \sup_{\substack{\Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\leq 1\\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \left(\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\varepsilon}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i\rangle\right)\\ &= \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i 1 \leq i \leq n}\Phi^*\Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{P}_{\mathcal{M}_H}\boldsymbol{P}_{\mathcal{V}}\big(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^\top\big)\Big),\end{aligned} \tag{4.14}$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i|\boldsymbol{J}|})^\top \in \{-1, 1\}^{|\boldsymbol{J}|}, 1 \leq i \leq n$, ε_{ij} is a sequence of independent doubly indexed Rademacher variables, Φ^* denotes the dual norm of Φ, and $\boldsymbol{P}_{\mathcal{M}_H}$ denotes the orthogonal projection operator onto subspace $\mathcal{M}_H$.

Here, the constants (κ, c_0, c_1, c_2) can depend on the GLM, the fixed vector $\boldsymbol{\theta}^\dagger$, and (α, β), but are independent of the dimension, sample size, and regularizer Φ. Furthermore, for pre-specified large value M, we know that $\kappa, c_0 \propto \gamma_M$, where γ_M is defined in (4.11).

Remark 4.1. The definition (4.14) of $\mu_n(\Phi, \mathcal{M}_H)$ does not exactly match that in Theorem 9.36 as presented in Wainwright (2019), because in our setting $\boldsymbol{y}_i \in \mathbb{R}^{|\boldsymbol{J}|}$ for $1 \leq i \leq n$ are multivariate responses. $\mu_n(\Phi, \mathcal{M}_H)$ is simply the Rademacher complexity of the class of linear operators $\boldsymbol{x} \mapsto \boldsymbol{\theta}\boldsymbol{x}$ as $\boldsymbol{\theta} \in \mathcal{M}_H$ ranges over the unit ball of norm Φ.

Lemma 4.5. *Under Assumptions 1-4, for both multinomial and Poisson categorical response models, there exists absolute constant A such that*

$$\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) = \mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{\varepsilon}_i,1\le i\le n}\Omega^*_{\mathcal{G}}\Big(\frac{1}{n}\boldsymbol{H}^\top\sum_{i=1}^n \boldsymbol{P}_{\mathcal{V}}\big(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^\top\big)\Big) \le A\cdot C\Big(\sqrt{\frac{\log|\mathcal{G}|}{n}}+\sqrt{\frac{m}{n}}\Big), \tag{4.15}$$

where $\Omega^*_{\mathcal{G}}$ denotes the dual norm of $\Omega_{\mathcal{G}}$, $|\mathcal{G}|$ denotes the number of groups and $m = \max_{(\boldsymbol{k},j)\in\mathcal{G}}|\boldsymbol{k}|_{\boldsymbol{J}}\cdot p_j$ denotes the maximum group size.

4.9.1 Proof of Theorem 4.2

In this subsection, we provide the proofs for Theorem 4.2, Lemma 4.5 and Lemma 5.

Proof of Theorem 4.2. The proof follows closely the approach used in the proof of Theorem 9.36 as presented in Wainwright (2019).

Recall that $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}')$ if and only if $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \mathbb{P}_{\boldsymbol{\theta}'}(\boldsymbol{y}|\boldsymbol{x}), \forall\boldsymbol{x}\in\mathbb{R}^p$ and $\boldsymbol{y}$, if and only if $\boldsymbol{P}_{\mathcal{V}}(\boldsymbol{\theta}-\boldsymbol{\theta}') = 0$. Because

$$\boldsymbol{P}_{\mathcal{V}}\Big((\boldsymbol{\theta}+\Delta\boldsymbol{\theta}) - (\boldsymbol{\theta}+\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\Big) = \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}) - \boldsymbol{P}^2_{\mathcal{V}}(\Delta\boldsymbol{\theta}) = \mathbf{0},$$

we know that $L_n(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}+\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))$. Notice that

$$\begin{aligned}\big\langle\nabla L_n(\boldsymbol{\theta}^\dagger), \Delta\boldsymbol{\theta}\big\rangle &= \lim_{\varepsilon\to 0}\frac{L_n(\boldsymbol{\theta}^\dagger+\varepsilon\Delta\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}^\dagger)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0}\frac{L_n(\boldsymbol{\theta}^\dagger+\varepsilon\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})) - L_n(\boldsymbol{\theta}^\dagger)}{\varepsilon} = \big\langle\nabla L_n(\boldsymbol{\theta}^\dagger), \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\big\rangle.\end{aligned}$$

In conclusion, $\mathcal{E}_n(\Delta\boldsymbol{\theta}) = \mathcal{E}_n(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))$. Without loss of generality, we can assume $R = 1$.

Under Assumption 4.1, to demonstrate that (4.13) holds for any $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$, it suffices to show that

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) = \mathcal{E}_n(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})) \geq \frac{\kappa}{2} \|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\|^2 - c_0 \cdot \mu_n^2 \cdot \Phi^2(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})), \|\Delta\boldsymbol{\theta}\| \leq 1. \quad (4.16)$$

We emulate the proof of Theorem 9.36 as presented in Wainwright (2019). Using the Lagrange remainder theorem for the Taylor series, we obtain

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left\langle \nabla^2\psi(\boldsymbol{\theta}^\dagger \boldsymbol{x}_i + t\Delta\boldsymbol{\theta}\boldsymbol{x}_i)\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i \right\rangle,$$

for some scalar $t \in [0,1]$. Similar to the argument in the proof of Theorem 9.36 in Wainwright (2019), we define $\|\Delta\boldsymbol{\theta}\| = \delta \in (0,1]$, and set $\tau = K\delta$ for a constant $K > 0$ to be chosen later. Define function $\varphi_\tau(\boldsymbol{u}) = \|\boldsymbol{u}\|^2 I[\|\boldsymbol{u}\| \leq 2\tau]$.

Now, we can replace (9.96) in Wainwright (2019) by

$$\begin{aligned} &\mathcal{E}_n(\Delta\boldsymbol{\theta}) \\ \geq &\frac{1}{n}\sum_{i=1}^{n} \left\langle \nabla^2\varphi\big(\boldsymbol{\theta}^\dagger \boldsymbol{x}_i + t\Delta\boldsymbol{\theta}\boldsymbol{x}_i\big)\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i \right\rangle \varphi_\tau(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\big\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\big\| \leq T], \end{aligned}$$

where T is the second truncation parameter to be chosen. Let $\gamma := \gamma_{T+2K}$. By Assumption 4.1, we know that $\gamma > 0$.

Thus, we obtain that

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) \geq \frac{\gamma}{n}\sum_{i=1}^{n} \varphi_\tau(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\big\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\big\| \leq T]. \quad (4.17)$$

Applying (4.17) and inequality $xy \leq \frac{x^2+y^2}{2}$, to show (4.13), it suffices to show that for all $\delta \in (0, 1]$ and for $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$ with $\|\Delta\boldsymbol{\theta}\| = \delta$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \varphi_\tau(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\right\| \leq T] \geq c_3\delta^2 - c_4\mu_n \Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\delta. \tag{4.18}$$

If the lower bound in (4.18) holds, then inequality (4.13) holds with constants (κ, c_0) depending on (c_3, c_4, γ). To be more specific, $\frac{\kappa}{\gamma}$ and $\frac{c_0}{\gamma}$ depending on (c_3, c_4).

Based on the argument presented in the proof of Theorem 9.36 in Wainwright (2019), it suffices to demonstrate that the bound (4.18) holds for $\|\Delta\boldsymbol{\theta}\| = \delta = 1$. Define a new truncation function

$$\widetilde{\varphi}_\tau(\boldsymbol{u}) = \|\boldsymbol{u}\|^2 I[\|\boldsymbol{u}\| \leq \tau] + (\|\boldsymbol{u}\| - 2\tau)^2 I[\tau < \|\boldsymbol{u}\| \leq 2\tau],$$

which is Lipschitz with parameter $2r$. Since $\widetilde{\varphi}_\tau(\boldsymbol{u}) \leq \varphi_\tau(\boldsymbol{u})$, it suffices to show that if $\|\Delta\boldsymbol{\theta}\| = 1$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\varphi_\tau}(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\right\| \leq T] \geq c_3 - c_4 \cdot \mu_n \cdot \Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})). \tag{4.19}$$

For a given radius $r \geq 1$, define the random variables

$$Z_n(r) := \sup_{\substack{\|\Delta\boldsymbol{\theta}\|=1 \\ \Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\leq r \\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \mathbb{E}\left\{\widetilde{\varphi}_\tau\left(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}\right) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}\right\| \leq T] - \frac{1}{n}\sum_{i=1}^{n} \widetilde{\varphi}_\tau\left(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i\right) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\right\| \leq T]\right\}. \tag{4.20}$$

Suppose that we can prove that

$$\mathbb{E}\left[\widetilde{\varphi_\tau}(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\right\| \leq T]\right] \geq \frac{3}{4}\alpha, \tag{4.21}$$

and

$$\mathbb{P}\big[Z_n(r) > \frac{\alpha}{2} + c_4 r \mu_n\big] \le \exp\big(-c_2 \frac{n r^2 \mu_n^2}{\sigma^2} - c_2 n\big). \tag{4.22}$$

Then with $c_3 = \alpha/4$, we obtain that for any $\|\Delta\boldsymbol{\theta}\| = 1$ and $\Phi(\Delta\boldsymbol{\theta}) \le r$, the bound (4.19) hold. Similar to the argument in the proof of Theorem 9.34 in Wainwright (2019), we obtain that (4.13) holds with probability at least $1 - c_1 e^{-c_2 n}$.

Now, we turn to the proof of inequalities (4.21) and (4.22). The proof of (4.21) closely follows the argument for the expectation bound (9.99a) presented in Wainwright (2019); therefore, we omit its detailed proof here. However, for (4.22), adapting the approach used for expectation bound (9.99b) in Wainwright (2019) requires the following justification.

Applying Lemma 26.2 in Shalev-Shwartz and Ben-David (2014) for Rademacher complexity, we know that

$$\mathbb{E}[Z_n(r)] \le 2 \cdot \mathbb{E}_{\boldsymbol{x}_i, \varepsilon_i} \sup_{\substack{\|\Delta\boldsymbol{\theta}\|=1 \\ \Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})) \le r \\ \Delta\boldsymbol{\theta} \in \mathcal{M}_H}} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \widetilde{\varphi_r}(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\left\|\boldsymbol{\theta}^{\dagger}\boldsymbol{x}_i\right\| \le T] \right).$$

Because $I[\left\|\boldsymbol{\theta}^{\dagger}\boldsymbol{x}_i\right\| \le T] \le 1$, $\widetilde{\varphi_r}(\boldsymbol{u})$ is Lipschitz with parameter $2K$, applying vector-contraction inequality for Rademacher complexities (see Corollary 1 in Maurer (2016)), we obtain

$$\begin{aligned}
\mathbb{E}[Z_n(r)] &\le 8K \cdot \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i} \sup_{\substack{\|\Delta\boldsymbol{\theta}\|=1 \\ \Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})) \le r \\ \Delta\boldsymbol{\theta} \in \mathcal{M}_H}} \left(\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\varepsilon}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i \rangle \right) \\
&\le 8Kr \cdot \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i} \sup_{\substack{\Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})) \le 1 \\ \Delta\boldsymbol{\theta} \in \mathcal{M}_H}} \left(\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\varepsilon}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i \rangle \right),
\end{aligned}$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i|\boldsymbol{J}|})^{\top} \in \{-1, 1\}^{|\boldsymbol{J}|}, 1 \le i \le n$, and $\{\varepsilon_{ij}\}_{1 \le i \le n, 1 \le j \le |\boldsymbol{J}|}$ is a sequence of independent doubly indexed Rademacher variables.

A direct calculation shows that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{\varepsilon}_i,1\le i\le n} \sup_{\substack{\Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\le 1\\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \left(\frac{1}{n}\sum_{i=1}^{n}\langle\boldsymbol{\varepsilon}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i\rangle\right)\\
&\quad\le \mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{\varepsilon}_i,1\le i\le n} \sup_{\substack{\Phi(\Delta\boldsymbol{\theta})\le 1\\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \left(\frac{1}{n}\sum_{i=1}^{n}\left\langle\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^{\top}, \boldsymbol{P}_{\mathcal{V}}\boldsymbol{P}_{\mathcal{M}_H}(\Delta\boldsymbol{\theta})\right\rangle\right)\\
&\quad= \mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{\varepsilon}_i,1\le i\le n} \sup_{\substack{\Phi(\Delta\boldsymbol{\theta})\le 1\\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \left(\frac{1}{n}\sum_{i=1}^{n}\left\langle\boldsymbol{P}_{\mathcal{M}_H}\boldsymbol{P}_{\mathcal{V}}(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^{\top}), \Delta\boldsymbol{\theta}\right\rangle\right)\\
&\quad= \mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{\varepsilon}_i,1\le i\le n}\Phi^*\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{P}_{\mathcal{M}_H}\boldsymbol{P}_{\mathcal{V}}\big(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^{\top}\big)\right).
\end{aligned}$$

With this modification of the proof of (9.99b) in Wainwright (2019), we complete the proof of (4.22). In conclusion, we complete the proof of (4.13), and obtain the expression (4.14).

Hence, the proof is concluded. □

4.9.2 Proof of Lemma 4.5

Proof of Lemma 4.5.

Recall that $\mathcal{M}_H = \{\boldsymbol{H}\boldsymbol{\beta}; \boldsymbol{\beta}\in\mathbb{R}^{\sum_{s=0}^{d} L_s\times p}\}$ and $\Phi_{\mathcal{G}}(\boldsymbol{\theta}) = \Omega_{\mathcal{G}}(\boldsymbol{H}^{\top}\boldsymbol{\theta})$. Thus, $\boldsymbol{P}_{\mathcal{M}_H}(\boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{H}^{\top}\boldsymbol{\theta}$.

Let $\Omega_{\mathcal{G}}^*$ and $\Phi_{\mathcal{G}}^*$ denote the dual norms of $\Omega_{\mathcal{G}}$ and $\Phi_{\mathcal{G}}$, respectively. Because

$$\Omega_{\mathcal{G}}^*(\boldsymbol{\alpha}) = \max_{\Omega_{\mathcal{G}}(\boldsymbol{\beta})\le 1}\langle\boldsymbol{\alpha},\boldsymbol{\beta}\rangle = \max_{\Omega_{\mathcal{G}}(\boldsymbol{\beta})\le 1}\langle\boldsymbol{H}\boldsymbol{\alpha},\boldsymbol{H}\boldsymbol{\beta}\rangle = \max_{\Phi_{\mathcal{G}}(\boldsymbol{\theta})\le 1,\boldsymbol{\theta}\in\mathcal{M}_H}\langle\boldsymbol{H}\boldsymbol{\alpha},\boldsymbol{\theta}\rangle = \Phi_{\mathcal{G}}^*(\boldsymbol{H}\boldsymbol{\alpha}), \tag{4.23}$$

we know that

$$\begin{aligned}\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) &= \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i, 1\le i\le n} \Phi_{\mathcal{G}}^* \Big(\frac{1}{n} \sum_{i=1}^n \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{P}_{\mathcal{V}} \big(\boldsymbol{\varepsilon}_i \boldsymbol{x}_i^\top \big) \Big) \\ &= \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i, 1\le i\le n} \Omega_{\mathcal{G}}^* \Big(\frac{1}{n} \boldsymbol{H}^\top \sum_{i=1}^n \boldsymbol{P}_{\mathcal{V}} \big(\boldsymbol{\varepsilon}_i \boldsymbol{x}_i^\top \big) \Big).\end{aligned}$$

By the definition of dual norm $\Phi_{\mathcal{G}}^*$ and Lemma 4.13, we obtain

$$\begin{aligned}\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) &= \mathbb{E}_{\boldsymbol{x}_i, 1\le i\le n} \mathbb{E}_{\boldsymbol{\varepsilon}_i, 1\le i\le n} \Phi_{\mathcal{G}}^* \Big(\frac{1}{n} \sum_{i=1}^n \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{P}_{\mathcal{V}} \big(\boldsymbol{\varepsilon}_i \boldsymbol{x}_i^\top \big) \Big) \\ &= \mathbb{E}_{\boldsymbol{x}_i, 1\le i\le n} \mathbb{E}_{\boldsymbol{\varepsilon}_i, 1\le i\le n} \max_{\|\boldsymbol{k}\|_0 \le d, 1\le j\le t} \frac{1}{w_{\boldsymbol{k},j} \cdot n} \left\| \sum_{i=1}^n \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{P}_{\mathcal{V}} \boldsymbol{\varepsilon}_i \{\boldsymbol{x}_{i(j)}\}^\top \right\|.\end{aligned}$$

Let $\boldsymbol{b}_{\boldsymbol{k},j} = \frac{1}{w_{\boldsymbol{k},j} \cdot n} \sum_{i=1}^n \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{P}_{\mathcal{V}} \boldsymbol{\varepsilon}_i \{\boldsymbol{x}_{i(j)}\}^\top$.

For non-negative random variable, we have

$$\begin{aligned}\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) &= \mathbb{E} \max_{\|\boldsymbol{k}\|_0 \le d, 1\le j\le t} \|\boldsymbol{b}_{\boldsymbol{k},j}\| \\ &= \int_0^\infty \mathbb{P}\big(\max_{\|\boldsymbol{k}\|_0 \le d, 1\le j\le t} \|\boldsymbol{b}_{\boldsymbol{k},j}\| > u \big) du \\ &\le u_* + \sum_{\|\boldsymbol{k}\|_0 \le d, 1\le j\le t} \int_0^\infty \mathbb{P}\big(\|\boldsymbol{b}_{\boldsymbol{k},j}\| > u + u_* \big) du.\end{aligned}$$

Note that

$$\|\boldsymbol{b}_{\boldsymbol{k},j}\| = \sup_{\substack{\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{k}|_{\boldsymbol{J}} \times p_j} \\ \|\boldsymbol{v}\| \le 1}} \langle \boldsymbol{v}, \boldsymbol{b}_{\boldsymbol{k},j} \rangle .$$

The supremum is taken over a $|\boldsymbol{k}|_{\boldsymbol{J}} \cdot p_j$-dimensional unit ball. Thus, there exists

$\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N$, $N \leq 5^{|\boldsymbol{k}|_J \times p_j}$, such that $\|\boldsymbol{v}_k\| \leq 1$ and

$$\|\boldsymbol{b}_{\boldsymbol{k},j}\| \leq 2 \max_{1 \leq l \leq N} \langle \boldsymbol{v}_l, \boldsymbol{b}_{\boldsymbol{k},j} \rangle .$$

Thus,

$$\begin{aligned}
&\mathbb{P}\big(\|\boldsymbol{b}_{\boldsymbol{k},j}\| > u + u_* \big) \\
=&\mathbb{P}\big(\max_{1 \leq l \leq N} \langle \boldsymbol{v}_l, \boldsymbol{b}_{\boldsymbol{k},j} \rangle > u/2 + u_*/2 \big) \leq \sum_{l=1}^{N} \mathbb{P}\big(\langle \boldsymbol{v}_l, \boldsymbol{b}_{\boldsymbol{k},j} \rangle > u/2 + u_*/2 \big).
\end{aligned}$$

Notice that for Rademacher variable ε,

$$\mathbb{E} e^{\varepsilon u} = \frac{e^u + e^{-u}}{2} \leq e^{u^2/2}, \forall u \in \mathbb{R}.$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}_i, 1 \leq i \leq n} e^{u \left\langle \boldsymbol{v}_l, \boldsymbol{b}_{\boldsymbol{k},j} \right\rangle} &= \mathbb{E} e^{u \left\langle \boldsymbol{v}_l, \frac{1}{w_{\boldsymbol{k},j} \cdot n} \sum_{i=1}^{n} \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{P}_{\mathcal{V}} \boldsymbol{\varepsilon}_i \{\boldsymbol{x}_{i(j)}\}^\top \right\rangle} \\
&= \prod_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}_i} \exp\left(\frac{u}{w_{\boldsymbol{k},j} \cdot n} \left\langle \boldsymbol{P}_{\mathcal{V}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{v}_l \boldsymbol{x}_{i(j)}, \boldsymbol{\varepsilon}_i \right\rangle \right) \\
&\leq \prod_{i=1}^{n} \exp\left(\frac{u^2}{2 \cdot w_{\boldsymbol{k},j}^2 \cdot n^2} \left\| \boldsymbol{P}_{\mathcal{V}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{v}_l \boldsymbol{x}_{i(j)} \right\|^2 \right) \\
&\leq \exp\left(\frac{u^2}{2 \cdot w_{\boldsymbol{k},j}^2 \cdot n^2} \sum_{i=1}^{n} \left\| \boldsymbol{x}_{i(j)} \right\|^2 \right) \leq \exp\left(\frac{u^2 \cdot C^2}{2 \cdot n} \right).
\end{aligned}$$

Thus, $\langle \boldsymbol{v}_l, \boldsymbol{b}_{\boldsymbol{k},j} \rangle$ is a $C/\sqrt{n}$-sub-Gaussian random variable (see Definition 2.2 in Wainwright (2019)). By Hoeffding bound for sub-Gaussian random variable (see Proposition 2.5 in Wainwright (2019)), we know that for any $u > 0$,

$$\mathbb{P}\big(\langle \boldsymbol{v}_l, \boldsymbol{b}_{\boldsymbol{k},j} \rangle \geq u \big) \leq \exp\left(-\frac{n \cdot u^2}{2 \cdot C^2} \right).$$

Let $m = \max_{(\boldsymbol{k},j)\in\mathcal{G}} |\boldsymbol{k}|_{\boldsymbol{J}} \cdot p_j$. In conclusion,

$$\begin{aligned}\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) &\le u_* + \sum_{\|\boldsymbol{k}\|_0\le d, 1\le j\le t} \int_0^\infty \mathbb{P}\big(\|\boldsymbol{b}_{\boldsymbol{k},j}\| > u + u_*\big) du \\ &\le u_* + |\mathcal{G}| \int_0^\infty 5^m \exp\Big(- \frac{n(u+u_*)^2}{8\cdot C^2}\Big) du.\end{aligned}$$

There exists absolute constants $A_1, A_2, A > 0$ such that if $u_* = A_1 \cdot C\big(\sqrt{\frac{\log|\mathcal{G}|}{n}} + \sqrt{\frac{m}{n}}\big)$, then

$$\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) \le u_* + \int_0^\infty \exp\Big(-\frac{nu^2}{8\cdot C^2}\Big) du = u_* + \frac{A_2\ C}{\sqrt{n}} = A\cdot C\Big(\sqrt{\frac{\log|\mathcal{G}|}{n}} + \sqrt{\frac{m}{n}}\Big).$$

□

4.9.3 Proof of Lemma 5

Proof of Lemma 5. By applying Lemma 4.5, Lemma 4.11 and Theorem 4.2, we complete the proof of Lemma 5. □

4.10 Auxiliary Lemmas

4.10.1 Subspace Lipschitz Constant

We first establish the subspace Lipschitz constant $\Psi(\mathcal{S})$.

Lemma 4.6.

$$\Psi(\mathcal{S}) = \sqrt{\sum_{(\boldsymbol{k},j)\in\mathcal{S}} w_{\boldsymbol{k},j}^2}\ .$$

Proof of Lemma 4.6. Assume $\boldsymbol{\theta}\boldsymbol{x} = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}\boldsymbol{x}^j$ for all vectors $\boldsymbol{x}$. Because columns of $\boldsymbol{H}$ are orthonormal, we know that

$$\|\boldsymbol{\theta}\|^2 = \|\boldsymbol{H}\boldsymbol{\beta}\|^2 = \|\boldsymbol{\beta}\|^2 = \sum_{(\boldsymbol{k},j)\in\mathcal{S}} \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\|^2.$$

By the definition of $\Phi_{\mathcal{G}}(\boldsymbol{\theta})$, we know that

$$\Phi_{\mathcal{G}}(\boldsymbol{\theta}) = \sum_{(\boldsymbol{k},j)\in\mathcal{S}} w_{\boldsymbol{k},j} \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\|.$$

By the Cauchy–Schwarz inequality, we know that

$$\Phi_{\mathcal{G}}(\boldsymbol{\theta})^2 \leq \sum_{(\boldsymbol{k},j)\in\mathcal{S}} w_{\boldsymbol{k},j}^2 \sum_{(\boldsymbol{k},j)\in\mathcal{S}} \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\|^2 = \sum_{(\boldsymbol{k},j)\in\mathcal{S}} w_{\boldsymbol{k},j}^2 \left\|\boldsymbol{\theta}\right\|^2$$

and the equation is achievable. Thus,

$$\Psi(\mathcal{S}) = \sqrt{\sum_{(\boldsymbol{k},j)\in\mathcal{S}} w_{\boldsymbol{k},j}^2}.$$

□

4.10.2 Gradient and Hessian Expressions

For completeness, we rederive the gradient and Hessian for negative log-likelihood of Poisson and multinomial categorical response models.

Poisson case:

The negative log-likelihood of Poisson categorical response model is given by

$$\begin{aligned} n\cdot\mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta}) &= \sum_{i=1}^n -\left\langle \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}}), \mathrm{vec}_{\boldsymbol{J}}(\log\boldsymbol{\mu}^{\boldsymbol{J}}(\boldsymbol{x}_i))\right\rangle + \left\langle \mathbf{1}_{|\boldsymbol{J}|}, \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{\mu}^{\boldsymbol{J}}(\boldsymbol{x}_i))\right\rangle \\ &= \sum_{i=1}^n -\left\langle \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}}), \boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i\right\rangle + \left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i}\right\rangle. \end{aligned}$$

By Taylor expansion,

$$\begin{aligned} &\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}(\boldsymbol{\beta}+\varepsilon\Delta\boldsymbol{\beta})\boldsymbol{x}_i}\right\rangle - \left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i}\right\rangle \\ =&\left\langle e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i}, e^{\varepsilon\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i} - \mathbf{1}_{|\boldsymbol{J}|}\right\rangle \\ =&\varepsilon\left\langle e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i}, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\right\rangle + \frac{\varepsilon^2}{2}\left\langle \mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\right\rangle + o(\varepsilon^2). \end{aligned}$$

By the formula for vectorization of matrix multiplication as Kronecker product

$$\mathrm{vec}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \left(\boldsymbol{C}^\top\otimes\boldsymbol{A}\right)\mathrm{vec}(\boldsymbol{B}),$$

we have

$$\begin{aligned} &\left\langle \mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\right\rangle \\ =&\left\langle \mathrm{vec}\left(\mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\right), \mathrm{vec}(\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i)\right\rangle \\ =&\left\langle \boldsymbol{x}_i^\top\otimes\mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i})\boldsymbol{H}\,\mathrm{vec}(\Delta\boldsymbol{\beta}), \boldsymbol{x}_i^\top\otimes\boldsymbol{H}\,\mathrm{vec}(\Delta\boldsymbol{\beta})\right\rangle \\ =&\left\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top\otimes\boldsymbol{H}^\top\mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i})\boldsymbol{H}\,\mathrm{vec}(\Delta\boldsymbol{\beta}), \mathrm{vec}(\Delta\boldsymbol{\beta})\right\rangle. \end{aligned}$$

Then, because

$$\begin{aligned} &\mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta}+\varepsilon\Delta\boldsymbol{\beta}) - \mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta}) \\ =&\varepsilon\left\langle \frac{\partial\mathcal{L}_n^{Posi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \Delta\boldsymbol{\beta}\right\rangle + \frac{\varepsilon^2}{2}\left\langle \frac{\partial^2\mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta})}{\partial\,\mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top}\mathrm{vec}(\Delta\boldsymbol{\beta}), \mathrm{vec}(\Delta\boldsymbol{\beta})\right\rangle + o(\varepsilon^2), \end{aligned}$$

we obtain

$$\frac{\partial \mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}^\top \left[e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i} - \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}})\right] \boldsymbol{x}_i^\top, \text{ and}$$

$$\frac{\partial^2 \mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta})}{\partial\, \mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \otimes \boldsymbol{H}^\top \mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i})\boldsymbol{H}.$$

Multinomial case: The negative log-likelihood of multinomial categorical response model is given by

$$n \cdot \mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\left\langle \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}}), \boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i\right\rangle + n_i \log\left(\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i}\right\rangle\right).$$

Let $\mathrm{vec}_{\boldsymbol{J}}\left(p^{\boldsymbol{J}}(\boldsymbol{x})\right) = \frac{e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\right\rangle}$. Note that

$$\begin{aligned}
&\log\left(\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}(\boldsymbol{\beta}+\varepsilon\Delta\boldsymbol{\beta})\boldsymbol{x}}\right\rangle\right) - \log\left(\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\right\rangle\right)\\
=&\log\left(1 + \frac{\left\langle e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}, e^{\varepsilon\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}} - \mathbf{1}_{|\boldsymbol{J}|}\right\rangle}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\right\rangle}\right)\\
=&\log\left(1 + \varepsilon\frac{\left\langle e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}\right\rangle}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\right\rangle} + \frac{\varepsilon^2}{2}\frac{\left\langle \mathrm{diag}(e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}\right\rangle}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\right\rangle} + o(\varepsilon^2)\right)\\
=&\log\left(1 + \varepsilon\left\langle \mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x})), \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}\right\rangle + \frac{\varepsilon^2}{2}\left\langle \mathrm{diag}\{\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}))\}\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}\right\rangle + o(\varepsilon^2)\right)\\
=&\varepsilon\left\langle \mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x})), \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}\right\rangle\\
&\quad + \frac{\varepsilon^2}{2}\left\langle \left\{\mathrm{diag}[\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}))] - \mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}))\{\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}))\}^\top\right\} \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}\right\rangle + o(\varepsilon^2).
\end{aligned}$$

Thus, we have

$$
\begin{aligned}
&\frac{\partial \mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}^\top \left[\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i)) - \mathrm{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}})\right] \boldsymbol{x}_i^\top, \\
&\frac{\partial^2 \mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta})}{\partial \,\mathrm{vec}(\boldsymbol{\beta}) \partial \{\mathrm{vec}(\boldsymbol{\beta})\}^\top} \\
=&\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \otimes \boldsymbol{H}^\top \Big\{\mathrm{diag}[\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i))] - \mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i))\{\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i))\}^\top\Big\}\boldsymbol{H}.
\end{aligned} \tag{4.24}
$$

4.10.3 Hyperparameters for Proximal Methods

We will study the global Lipschitz constant, for multinomial categorical response model, and local Lipschitz constant for Poisson categorical response model, L' such that

$$
\|\nabla \mathcal{L}_n(\boldsymbol{\beta}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}')\| \le L' \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|.
$$

Lemma 4.7. *For any twice differentiable function $\mathcal{L}_n(\boldsymbol{\beta})$ and any $\forall \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^{\sum_{s=0}^{d} L_s \times p}$, we have*

$$
\|\nabla \mathcal{L}_n(\boldsymbol{\beta}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}')\| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}'\| \cdot \max_{\substack{0 \le t \le 1; \\ \boldsymbol{\beta}_t = t\boldsymbol{\beta} + (1-t)\boldsymbol{\beta}'}} \left\| \frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta}_t)}{\partial \,\mathrm{vec}(\boldsymbol{\beta}) \partial \{\mathrm{vec}(\boldsymbol{\beta})\}^\top} \right\|_{\mathrm{op}}. \tag{4.25}
$$

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$. For multinomial categorical response model, for any $\boldsymbol{\beta} \in \mathbb{R}^{\sum_{s=0}^{d} L_s \times p}$,

$$
\left\| \frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta})}{\partial \,\mathrm{vec}(\boldsymbol{\beta}) \partial \{\mathrm{vec}(\boldsymbol{\beta})\}^\top} \right\|_{\mathrm{op}} \le \frac{1}{2n} \lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) \text{ and} \tag{4.26}
$$

$$\left\|\frac{\partial^2 \mathcal{L}_n(\mathbf{0})}{\partial \operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}} = \frac{1}{n\cdot|\boldsymbol{J}|}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}). \tag{4.27}$$

For Poisson categorical response model, we have

$$\left\|\frac{\partial^2 \mathcal{L}_n(\mathbf{0})}{\partial \operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}} = \frac{1}{n}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}). \tag{4.28}$$

Proof of Lemma 4.7. Because

$$\operatorname{vec}\big(\nabla\mathcal{L}_n(\boldsymbol{\beta}') - \nabla\mathcal{L}_n(\boldsymbol{\beta})\big) = \int_0^1 \frac{\partial^2\mathcal{L}_n(\boldsymbol{\beta}+u(\boldsymbol{\beta}'-\boldsymbol{\beta}))}{\partial \operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\operatorname{vec}(\boldsymbol{\beta}'-\boldsymbol{\beta})du,$$

we obtain inequality (4.25).

Let $p(\boldsymbol{x}) = \frac{1}{\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\rangle}e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}$. By (4.24), we obtain

$$\begin{aligned}
&\left\|\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta})}{\partial \operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}} \\
&= \lambda_{\max}\Big(\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta})}{\partial \operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\Big) \\
&= \max_{\|\Delta\boldsymbol{\beta}\|=1}\frac{1}{n}\sum_{i=1}^n\Big\langle\Big(\operatorname{diag}(p(\boldsymbol{x}_i)) - p(\boldsymbol{x}_i)p^\top(\boldsymbol{x}_i)\Big)\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\Big\rangle \\
&\le \frac{1}{n}\max_{\|\Delta\boldsymbol{\beta}\|=1}\sum_{i=1}^n\lambda_{\max}\big(\operatorname{diag}(p(\boldsymbol{x}_i)) - p(\boldsymbol{x}_i)p^\top(\boldsymbol{x}_i)\big)\,\|\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\|^2 \\
&\le \frac{1}{n}\max_{\boldsymbol{x}}\lambda_{\max}\big(\operatorname{diag}(p(\boldsymbol{x})) - p(\boldsymbol{x})p^\top(\boldsymbol{x})\big)\cdot\max_{\|\Delta\boldsymbol{\beta}\|=1}\sum_{i=1}^n\|\Delta\boldsymbol{\beta}\boldsymbol{x}_i\|^2 \\
&= \frac{1}{n}\max_{\boldsymbol{x}}\lambda_{\max}\big(\operatorname{diag}(p(\boldsymbol{x})) - p(\boldsymbol{x})p^\top(\boldsymbol{x})\big)\cdot\max_{\|\Delta\boldsymbol{\beta}\|=1}\|\Delta\boldsymbol{\beta}\boldsymbol{X}\|^2 \\
&\le \frac{1}{n}\max_{\boldsymbol{x}}\lambda_{\max}\big(\operatorname{diag}(p(\boldsymbol{x})) - p(\boldsymbol{x})p^\top(\boldsymbol{x})\big)\cdot\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}).
\end{aligned}$$

By Geršgorin circle theorem (see Theorem 1.1 in Varga (2010)), we know that

$$\begin{aligned}\lambda_{\max}\big(\operatorname{diag}(p(\boldsymbol{x})) - p(\boldsymbol{x})p^\top(\boldsymbol{x})\big) \leq & \max_{1\leq i\leq n}\Big(p_i(\boldsymbol{x})(1-p_i(\boldsymbol{x})) + p_i(\boldsymbol{x})\sum_{j\neq i} p_j(\boldsymbol{x})\Big)\\ =& 2\max_{1\leq i\leq n}\Big(p_i(\boldsymbol{x})\big(1-p_i(\boldsymbol{x})\big)\Big) \leq \frac{1}{2}.\end{aligned}$$

In conclusion, we complete the proof of inequality (4.26). Then, because for multinomial categorical response model $p(\mathbf{0}) = \frac{1}{|\boldsymbol{J}|}\mathbf{1}_{|\boldsymbol{J}|}$, we know that

$$\begin{aligned}&\left\|\frac{\partial^2\mathcal{L}_n(\mathbf{0})}{\partial\operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\right\|\\ =& \lambda_{\max}\Big(\frac{\partial^2\mathcal{L}_n(\mathbf{0})}{\partial\operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\Big)\\ =& \max_{\|\Delta\boldsymbol{\beta}\|=1}\frac{1}{n}\sum_{i=1}^n\Big\langle\big(\operatorname{diag}(p(\mathbf{0})) - p(\mathbf{0})p^\top(\mathbf{0})\big)\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\Big\rangle\\ =& \max_{\|\Delta\boldsymbol{\beta}\|=1}\frac{1}{n\cdot|\boldsymbol{J}|}\sum_{i=1}^n\langle\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\rangle - \frac{1}{n}\sum_{i=1}^n\Big\langle p^\top(\mathbf{0})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, p^\top(\mathbf{0})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\Big\rangle\\ =& \max_{\|\Delta\boldsymbol{\beta}\|=1}\frac{1}{n\cdot|\boldsymbol{J}|}\|\Delta\boldsymbol{\beta}\boldsymbol{X}\|^2 - \frac{1}{n}\left\|p^\top(\mathbf{0})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{X}\right\|^2\\ =& \frac{1}{n\cdot|\boldsymbol{J}|}\max_{\|\Delta\boldsymbol{\beta}\|=1}\left\|\big(\boldsymbol{X}^\top\otimes I\big)\operatorname{vec}(\Delta\boldsymbol{\beta})\right\|^2 - \left\|\boldsymbol{e}_1^\top\Delta\boldsymbol{\beta}\boldsymbol{X}\right\|^2\\ =& \frac{1}{n\cdot|\boldsymbol{J}|}\max_{\|\Delta\boldsymbol{\beta}\|=1}\Big\langle\Big(\boldsymbol{X}\boldsymbol{X}^\top\otimes(I-\boldsymbol{e}_1\boldsymbol{e}_1^\top)\Big)\operatorname{vec}(\Delta\boldsymbol{\beta}), \operatorname{vec}(\Delta\boldsymbol{\beta})\Big\rangle\\ &\frac{1}{n\cdot|\boldsymbol{J}|}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X})\cdot\lambda_{\max}\big(I-\boldsymbol{e}_1\boldsymbol{e}_1^\top\big) = \frac{1}{n\cdot|\boldsymbol{J}|}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}).\end{aligned}$$

Similarly, for Poisson categorical response model,

$$\begin{aligned}\left\|\frac{\partial^2\mathcal{L}_n(\mathbf{0})}{\partial\operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\right\| =& \lambda_{\max}\Big(\frac{\partial^2\mathcal{L}_n(\mathbf{0})}{\partial\operatorname{vec}(\boldsymbol{\beta})\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^\top}\Big)\\ =& \max_{\|\Delta\boldsymbol{\beta}\|=1}\frac{1}{n}\sum_{i=1}^n\Big\langle\operatorname{diag}(e^{\boldsymbol{H}\mathbf{0}\boldsymbol{x}_i})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{x}_i\Big\rangle\\ =& \max_{\|\Delta\boldsymbol{\beta}\|=1}\frac{1}{n}\|\Delta\boldsymbol{\beta}\boldsymbol{X}\|^2 = \frac{1}{n}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}).\end{aligned}$$

□

The next lemma establishes the critical value of the regularization parameter, λ_1 defined in Section 4.4, for the group lasso penalty $\Omega_{\mathcal{G}}(\boldsymbol{\beta})$.

Lemma 4.8. *If $\Omega_{\mathcal{G}}(\boldsymbol{\beta})$ denotes the group lasso penalty defined in (2.14), then*

$$\mathbf{0} \in \arg\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}) \text{ if and only if } \lambda \geq \max_{(\boldsymbol{k},j)\in\mathcal{G}} \frac{1}{w_{\boldsymbol{k},j}} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\mathbf{0}) \right\|.$$

Moreover, if $\Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta})$ denotes the overlapping group lasso penalty defined in (2.17), then

$$\lambda \geq \max_{(\boldsymbol{k},j)\in\mathcal{G}} \frac{1}{w_{\boldsymbol{k},j}} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\mathbf{0}) \right\| \implies \mathbf{0} \in \arg\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta}).$$

Proof of Lemma 4.8. Assume

$$\mathbf{0} \in \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}_n(\mathbf{0}) + \lambda \partial \Omega_{\mathcal{G}}(\mathbf{0}), \tag{4.29}$$

where ∂ denotes the subgradient.

We know that (4.29) holds if and only if for any $(\boldsymbol{k}, j) \in \mathcal{G}$,

$$\left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\mathbf{0}) \right\| \leq \lambda w_{\boldsymbol{k},j},$$

which means

$$\lambda \geq \max_{(\boldsymbol{k},j)\in\mathcal{G}} \frac{1}{w_{\boldsymbol{k},j}} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\mathbf{0}) \right\|.$$

Note that for any $\boldsymbol{\beta}$

$$\Omega_{\mathcal{G}}(\boldsymbol{\beta}) \leq \Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta}).$$

If $\lambda \geq \max_{(\boldsymbol{k},j)\in\mathcal{G}} \frac{1}{w_{\boldsymbol{k},j}} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\mathbf{0}) \right\|$, we know that $\mathbf{0}$ is a global minimizer of $\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta})$. Thus,

$$\begin{aligned}
&\mathcal{L}_n(\mathbf{0}) + \lambda\Omega_{\mathcal{D}(\mathcal{G})}(\mathbf{0}) = \mathcal{L}_n(\mathbf{0}) = \mathcal{L}_n(\mathbf{0}) + \lambda\Omega_{\mathcal{G}}(\mathbf{0}) \\
\leq &\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}) \leq \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta}),
\end{aligned}$$

which implies that

$$\mathbf{0} \in \arg\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{D}(\mathcal{G})}(\boldsymbol{\beta}).$$

□

4.10.4 Monotone property of reparameterized group lasso penalty

We provide a sufficient condition for the monotone property of reparameterized grouped lasso penalty in the following lemma. The following lemma remains valid for a general convex loss function $L_n(\boldsymbol{\theta})$.

Lemma 4.9. *Assume that there exists an orthogonal projection matrix $\boldsymbol{P}_{\mathcal{V}}$ such that $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \mathcal{M}_H$ and $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top = \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}$ for any $\boldsymbol{k} \in \mathcal{K}$. Define the linear operator $\boldsymbol{Q}_{\mathcal{V}}$ such that $\left(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})\right)_{\boldsymbol{k},j} = \boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}$ for any $(\boldsymbol{k},j) \in \mathcal{G}$. The following statements hold:*

1. *$\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{H}\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})$ and $\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})) \leq \Omega_{\mathcal{G}}(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}$.*

2. *If $w_g > 0$ for any $g \in \mathcal{G}$, then for any $\boldsymbol{\beta}^\dagger$ and $\widehat{\boldsymbol{\beta}}$ given by*

$$\boldsymbol{\beta}^\dagger \in \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \Omega_{\mathcal{G}}(\boldsymbol{\beta}) \text{ and } \widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}),$$

we have $\boldsymbol{\beta}^\dagger = \boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger)$ and $\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})$.

3. *If* $\boldsymbol{\theta}^*$ *is the true model generation parameter and* $\boldsymbol{H}_{\text{full}} = \{\boldsymbol{H}_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\cup_{s=0}^{q}\mathcal{K}_s} \in \mathbb{R}^{|\boldsymbol{J}|\times|\boldsymbol{J}|}$, *then*

$$\boldsymbol{\beta}^\dagger = \boldsymbol{H}_{\text{full}}^\top \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \Omega_{\mathcal{G}}(\boldsymbol{\beta}).$$

Proof of Lemma 4.9. Because $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top = \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}$, we know that $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}} = \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}$. It is easy to check that $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{H}\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})$ and for any $\boldsymbol{k} \in \mathcal{K}$, $\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}$ is an orthogonal projection matrix.

Because

$$\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}\right\| \le \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\| = \Omega_{\mathcal{G}}(\boldsymbol{\beta}),$$

the proof of statement 1 has been completed.

Assume $\boldsymbol{\beta}^\dagger \in \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \Omega_{\mathcal{G}}(\boldsymbol{\beta})$. Because

$$\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger) = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\boldsymbol{\beta}^\dagger = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\boldsymbol{\beta}^\dagger = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}^*,$$

and

$$\begin{aligned}\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger)) &= \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}^\dagger_{\boldsymbol{k},j}\right\| \\ &\le \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{\beta}^\dagger_{\boldsymbol{k},j}\right\| = \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger) \le \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger)),\end{aligned}$$

we obtain that if $w_{\boldsymbol{k},j} > 0$, then $\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}^\dagger_{\boldsymbol{k},j}\right\| = \left\|\boldsymbol{\beta}^\dagger_{\boldsymbol{k},j}\right\|$, and thus

$$\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}^\dagger_{\boldsymbol{k},j} = \boldsymbol{\beta}^\dagger_{\boldsymbol{k},j}$$

for any $(\boldsymbol{k},j) \in \mathcal{G}$. We obtain $\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger) = \boldsymbol{\beta}^\dagger$ for any $\boldsymbol{\beta}^\dagger \in \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \Omega_{\mathcal{G}}(\boldsymbol{\beta})$.

Assume $\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta})$. We can show that

$$\mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) = L_n(\boldsymbol{H}\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) = L_n(\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\widehat{\boldsymbol{\beta}}) = L_n(\boldsymbol{H}\widehat{\boldsymbol{\beta}}) = \mathcal{L}_n(\widehat{\boldsymbol{\beta}}).$$

We can also show that

$$\begin{aligned}
\mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) &= L_n(\boldsymbol{H}\widehat{\boldsymbol{\beta}}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) \\
&\le L_n(\boldsymbol{H}\widehat{\boldsymbol{\beta}}) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) \\
&= \mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) \\
&= \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) \le \mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})).
\end{aligned}$$

Thus, we obtain $\mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) = \mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}}))$, which implies $\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) = \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}}))$. If $w_{\boldsymbol{k},j} > 0$, then $\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k},j}\right\| = \left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{k},j}\right\|$, and thus $\boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k},j} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{k},j}$ for any $(\boldsymbol{k}, j) \in \mathcal{G}$. We obtain $\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})$ for any $\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta})$. This completes the proof of statement 2.

By the definition of $\mathcal{F}_{\boldsymbol{\theta}^*}$ in (2.24), we have $\boldsymbol{\beta}^\dagger \in \mathcal{F}_{\boldsymbol{\theta}^*}$,

$$\underset{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}}{\arg\min}\, \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}}{\arg\min}\, \Omega_{\mathcal{G}}(\boldsymbol{\beta})$$

and

$$\mathcal{F}_{\boldsymbol{\theta}^*} = \left\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{\sum_{i=0}^q L_i \times p}; \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}^* = \boldsymbol{H}_{\text{full}}\boldsymbol{Q}_{\mathcal{V}}(\widetilde{\boldsymbol{\beta}})\right\} = \left\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{\sum_{i=0}^q L_i \times p}; \boldsymbol{\beta}^\dagger = \boldsymbol{Q}_{\mathcal{V}}(\widetilde{\boldsymbol{\beta}})\right\}.$$

Applying statement 1 of Lemma 4.9, for any $\widetilde{\boldsymbol{\beta}} \in \mathcal{F}_{\boldsymbol{\theta}^*}$, we have

$$\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger) = \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widetilde{\boldsymbol{\beta}})) \le \Omega_{\mathcal{G}}(\widetilde{\boldsymbol{\beta}}).$$

The proof of statement 3 has been completed.

□

4.10.5 Proof of Lemma 4.11

To show that the γ_M in (4.11) for multinomial and Poisson categorical response model both satisfy Assumptions 4.1, we only need to show Lemma 4.11. To show Lemma 4.11 for multinomial categorical response model, we introduce the following lemma.

Lemma 4.10. *Let* $\psi(\boldsymbol{u}) = \log(\sum_{i=1}^m e^{u_i})$, $\boldsymbol{u} = (u_1, \cdots, u_m)^\top \in \mathbb{R}^m$. *Let* $u_{\max} = \max_{1\leq i\leq m} u_i$ *and* $u_{\min} = \min_{1\leq i\leq m} u_i$. *Then*

$$\min_{\substack{\boldsymbol{x}\in\mathbb{R}^m\\ \mathbf{1}^\top\boldsymbol{x}=0\\ \boldsymbol{x}\neq 0}} \frac{\langle \nabla^2\psi(\boldsymbol{u})\boldsymbol{x}, \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} \geq \frac{1}{m\cdot e^{2(u_{\max}-u_{\min})}}.$$

Proof of Lemma 4.10. First of all, we know that for any $\boldsymbol{u}$,

$$\nabla^2\psi(\boldsymbol{u}) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top,$$

where $\boldsymbol{p} = (p_1, \cdots, p_m)^\top$ and $p_j = \frac{e^{u_j}}{\sum_{i=1}^m e^{u_i}}, 1 \leq j \leq m$. Note that

$$1 = \sum_{i=1}^m p_i \leq m\cdot p_{\max} = m\cdot e^{u_{\max}-u_{\min}}\cdot p_{\min}, \tag{4.30}$$

where $p_{\max} = \max_{1\leq i\leq m} p_i$, $p_{\min} = \min_{1\leq i\leq m} p_i$, and u_{min} and u_{max} can be defined accordingly.

If $\boldsymbol{x} \in \mathbb{R}^m$ such that $\mathbf{1}^\top \boldsymbol{x} = 0$, we know that

$$\begin{aligned}
\left\langle \nabla^2\psi(\boldsymbol{u})\boldsymbol{x}, \boldsymbol{x} \right\rangle &= \sum_{i=1}^m p_i x_i^2 - (\sum_{i=1}^m p_i x_i)^2 \\
&= \big(\sum_{i=1}^m p_i x_i^2\big)\big(\sum_{j=1}^m p_j\big) - \big(\sum_{i=1}^m p_i x_i\big)\big(\sum_{j=1}^m p_j x_j\big) \\
&= \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m p_i p_j (x_i^2 - 2x_i x_j + x_j^2) \\
&= \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m p_i p_j (x_i - x_j)^2 \geq \frac{1}{2} p_{\min}^2 \sum_{i=1}^m \sum_{j=1}^m (x_i - x_j)^2 = m \cdot p_{\min}^2 \sum_{i=1}^m x_i^2.
\end{aligned}$$

Combined with inequality (4.30), we obtain that

$$\left\langle \nabla^2\psi(\boldsymbol{u})\boldsymbol{x}, \boldsymbol{x} \right\rangle \geq \frac{1}{m \cdot e^{2(u_{\max} - u_{\min})}} \left\| \boldsymbol{x} \right\|^2, \mathbf{1}^\top \boldsymbol{x} = 0.$$

$\square$

Lemma 4.11. *For multinomial categorical response model with $n_i = 1$, and $\psi(\boldsymbol{u}) = \log(\sum_{l=1}^{|\boldsymbol{J}|} e^{u_l})$, we have*

$$\gamma_M \geq \frac{1}{|\boldsymbol{J}| \cdot \exp\big(4M\big)}. \tag{4.31}$$

For Poisson categorical response model with $\psi(\boldsymbol{u}) = \sum_{l=1}^{|\boldsymbol{J}|} e^{u_l}$, we have

$$\gamma_M \geq e^{-M}. \tag{4.32}$$

Proof of Lemma 4.11. Let $\boldsymbol{u} = (u_1, \cdots, u_{|\boldsymbol{J}|})^\top = \boldsymbol{\theta}\boldsymbol{x}$. We know that

$$-M \leq -\left\|\boldsymbol{\theta}\boldsymbol{x}\right\| \leq u_{\min} \leq u_{\max} \leq \left\|\boldsymbol{\theta}\boldsymbol{x}\right\| \leq M,$$

where $u_{\max} = \max_{1\le l\le |\boldsymbol{J}|} u_l$ and $u_{\min} = \min_{1\le l\le |\boldsymbol{J}|} u_l$. For multinomial categorical response model, by Lemma 4.10 we obtain (4.31). For Poisson categorical response model, we know that $\lambda_{\min}\big(\nabla^2\psi(\boldsymbol{u})\big) = e^{u_{min}}$. This completes the proof of (4.32). □

4.10.6 Norm and Dual Norm for Group Lasso

Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_t] \in \mathcal{M}_H \subset \mathbb{R}^{|\boldsymbol{J}|\times p}$, $\boldsymbol{\theta}_j \in \mathbb{R}^{|\boldsymbol{J}|\times p_j}, 1 \le j \le t$. Recall that $\mathcal{M}_H := \{\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|\times p}; \boldsymbol{\theta}\boldsymbol{x} = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}\boldsymbol{x}^j, \forall \boldsymbol{x} \in \mathbb{R}^p\}$. Due to mapping $\boldsymbol{\beta} \mapsto \boldsymbol{H}\boldsymbol{\beta}$ being one to one, by (2.15), we know that

$$\Phi_{\mathcal{G}}(\boldsymbol{\theta}) = \sum_{g\in\mathcal{G}} w_g \left\|\boldsymbol{\beta}_g\right\| = \Omega_{\mathcal{G}}(\boldsymbol{\beta}), \boldsymbol{\theta} = \boldsymbol{H}\boldsymbol{\beta}.$$

Lemma 4.12. *If $w_g > 0$ for any $g \in \mathcal{G}$, then $\mathcal{M}_H \ni \boldsymbol{\theta} \mapsto \Phi_{\mathcal{G}}(\boldsymbol{\theta}) \in \mathbb{R}^+ \cup \{0\}$ is a norm.*

Proof of Lemma 4.12. The proof is very similar to the proof of Lemma 2 in Obozinski et al. (2011). Therefore, we omit the proof. □

Next, we define the dual norm for $\boldsymbol{\alpha} \in \mathcal{M}_H$ as

$$\Phi^*_{\mathcal{G}}(\boldsymbol{\alpha}) = \max\Big\{ \langle\boldsymbol{\theta}, \boldsymbol{\alpha}\rangle\,; \Phi_{\mathcal{G}}(\boldsymbol{\theta}) \le 1, \forall\boldsymbol{\theta} \in \mathcal{M}_H\Big\}.$$

The following lemma shows that $\Phi^*_{\mathcal{G}}(\boldsymbol{\alpha})$ has a simple closed form expression.

Lemma 4.13. *The dual norm $\Phi^*_{\mathcal{G}}(\boldsymbol{\alpha})$ of $\Phi_{\mathcal{G}}(\boldsymbol{\theta})$ satisfies:*

$$\Phi^*_{\mathcal{G}}(\boldsymbol{\alpha}) = \max_{g=(\boldsymbol{k},j)\in\mathcal{G}} \quad w_g^{-1}\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{\alpha}_j\right\|.$$

Proof of Lemma 4.13. The proof is similar to the proof of Lemma 3 in Obozinski et al. (2011). Let $\boldsymbol{P}_{\mathcal{M}_H} = \boldsymbol{H}\boldsymbol{H}^\top$. For any $\boldsymbol{\alpha} \in \mathcal{M}_H$, we know that $\boldsymbol{\alpha} = \boldsymbol{P}_{\mathcal{M}_H}\boldsymbol{\alpha}$.

For any $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_t] \in \mathcal{M}_H$, starting from the definition of the dual norm, we obtain that:

$$
\begin{aligned}
&\max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle \, ; \Phi_{\mathcal{G}}(\boldsymbol{\theta}) \le 1, \boldsymbol{\theta} \in \mathcal{M}_H \Big\} \\
&\quad = \max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle \, ; \Phi_{\mathcal{G}}(\boldsymbol{\theta}) \le 1, \boldsymbol{\theta} = \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{H} \boldsymbol{\beta} \Big\} \\
&\quad = \max\Big\{ \langle \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{H} \boldsymbol{\beta}, \boldsymbol{\alpha} \rangle \, ; \sum_{g \in \mathcal{G}} w_g \left\| \boldsymbol{\beta}_g \right\| \le 1 \Big\} \\
&\quad = \max\Big\{ \left\langle \boldsymbol{\beta}, \boldsymbol{H}^\top \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{\alpha} \right\rangle ; \sum_{g \in \mathcal{G}} w_g \left\| \boldsymbol{\beta}_g \right\| \le 1 \Big\} \\
&\quad = \max\Big\{ \sum_{(\boldsymbol{k}, j) \in \mathcal{G}} \left\langle \boldsymbol{\beta}_{\boldsymbol{k}, j}, \boldsymbol{H}_k^\top \boldsymbol{\alpha}_j \right\rangle ; w_g \left\| \boldsymbol{\beta}_g \right\| \le \eta_g, \forall g \in \mathcal{G}, \sum_{g \in \mathcal{G}} \eta_g \le 1 \Big\} \\
&\quad = \max\Big\{ \sum_{g = (\boldsymbol{k}, j) \in \mathcal{G}} \eta_g w_g^{-1} \left\| \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{\alpha}_j \right\| ; \sum_{g \in \mathcal{G}} \eta_g \le 1, \eta_g \ge 0, \forall g \in \mathcal{G} \Big\} \\
&\quad = \max_{g = (\boldsymbol{k}, j) \in \mathcal{G}} w_g^{-1} \left\| \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{\alpha}_j \right\| .
\end{aligned}
$$

□

Because $\Phi_{\mathcal{G}}(\boldsymbol{\theta})$ is define on finite dimensional vector space $\mathcal{M}_H$, the dual norm of the dual norm is the original norm. Thus, for any $\boldsymbol{\theta} \in \mathcal{M}_H$

$$
\begin{aligned}
\Phi_{\mathcal{G}}(\boldsymbol{\theta}) &= \max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle \, ; \Phi_{\mathcal{G}}^*(\boldsymbol{\alpha}) \le 1, \boldsymbol{\alpha} \in \mathcal{M}_H \Big\} \\
&= \max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle \, ; \boldsymbol{\alpha} \in \mathcal{M}_H, \left\| \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{\alpha}_j \right\| \le w_g, \forall g = (\boldsymbol{k}, j) \in \mathcal{G} \Big\} \\
&= \max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{\alpha} \rangle \, ; \boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{J}| \times p}, \left\| \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{P}_{\mathcal{M}_H} \boldsymbol{\alpha}_j \right\| \le w_g, \forall g = (\boldsymbol{k}, j) \in \mathcal{G} \Big\} \\
&= \max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle \, ; \boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{J}| \times p}, \left\| \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{\alpha}_j \right\| \le w_g, \forall g = (\boldsymbol{k}, j) \in \mathcal{G} \Big\}.
\end{aligned}
$$

which leads to the following Lemma.

Lemma 4.14. *For any* $\boldsymbol{\theta} \in \mathcal{M}_H$*,*

$$\Phi_{\mathcal{G}}(\boldsymbol{\theta}) = \max\Big\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle \,;\, \big\|\boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{\alpha}_j\big\| \le w_g, \forall g = (\boldsymbol{k}, j) \in \mathcal{G}, \boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_t] \in \mathbb{R}^{|\boldsymbol{J}| \times p} \Big\}.$$

4.10.7 Dual Norm for Partial Derivatives

Let $\ell(\boldsymbol{u}, \boldsymbol{y})$ be a differentiable function in $\boldsymbol{u} \in \mathbb{R}^{|\boldsymbol{J}|}$. ℓ is not necessarily ℓ_{Mult} or ℓ_{Pois}. For any $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_t] \in \mathcal{M}_H$, define

$$L_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta} \boldsymbol{x}_i, \boldsymbol{y}_i) \text{ and } \mathcal{L}_n(\boldsymbol{\beta}) = L_n(\boldsymbol{H}\boldsymbol{\beta}).$$

Lemma 4.15. *For any* $\boldsymbol{\beta}, \Delta\boldsymbol{\beta} \in \mathbb{R}^{\sum_{s=0}^d L_s \times p}$*,*

$$\big| \langle \nabla \mathcal{L}_n(\boldsymbol{\beta}), \Delta\boldsymbol{\beta} \rangle \big| \le \Omega_{\mathcal{G}}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta})) \cdot \Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}),$$

where the dual norm of $\nabla \mathcal{L}_n(\boldsymbol{\beta})$ *is given by*

$$\Omega_{\mathcal{G}}^*\Big(\nabla \mathcal{L}_n(\boldsymbol{\beta})\Big) = \max_{g \in \mathcal{G}} \; w_g^{-1} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_g} \mathcal{L}_n(\boldsymbol{\beta}) \right\|.$$

Proof of Lemma 4.15. First of all, notice that

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{u}} \ell(\boldsymbol{\theta} \boldsymbol{x}_i, \boldsymbol{y}_i) \boldsymbol{x}_i^\top,$$

$$\nabla \mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{H}^\top \frac{\partial}{\partial \boldsymbol{u}} \ell(\boldsymbol{H}\boldsymbol{\beta} \boldsymbol{x}_i, \boldsymbol{y}_i) \boldsymbol{x}_i^\top = \boldsymbol{H}^\top \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \text{ and}$$

$$\frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g} = \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{H}_{\boldsymbol{k}}^\top \frac{\partial}{\partial \boldsymbol{u}} \ell(\boldsymbol{H}\boldsymbol{\beta} \boldsymbol{x}_i, \boldsymbol{y}_i) \{\boldsymbol{x}_{i(j)}\}^\top, g = (\boldsymbol{k}, j) \in \mathcal{G}.$$

Because

$$\Omega^*_{\mathcal{G}}(\boldsymbol{\alpha}) = \max_{\Omega_{\mathcal{G}}(\boldsymbol{\beta})\le 1} \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle = \max_{\Omega_{\mathcal{G}}(\boldsymbol{\beta})\le 1} \langle \boldsymbol{H}\boldsymbol{\alpha}, \boldsymbol{H}\boldsymbol{\beta}\rangle = \max_{\Phi_{\mathcal{G}}(\boldsymbol{\theta})\le 1, \boldsymbol{\theta}\in\mathcal{M}_H} \langle \boldsymbol{H}\boldsymbol{\alpha}, \boldsymbol{\theta}\rangle = \Phi^*_{\mathcal{G}}(\boldsymbol{H}\boldsymbol{\alpha}),$$

and $\nabla\mathcal{L}_n(\boldsymbol{\beta}) = \boldsymbol{H}^\top \nabla L_n(\boldsymbol{H}\boldsymbol{\beta})$, we know that

$$\begin{aligned}
&\Omega^*_{\mathcal{G}}(\nabla\mathcal{L}_n(\boldsymbol{\beta}))\\
&= \Phi^*_{\mathcal{G}}(\boldsymbol{H}\boldsymbol{H}^\top \nabla L_n(\boldsymbol{H}\boldsymbol{\beta}))\\
&= \max_{(\boldsymbol{k},j)\in\mathcal{G}} w^{-1}_{\boldsymbol{k},j} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{H}_{\boldsymbol{k}}^\top \frac{\partial}{\partial \boldsymbol{u}} \ell(\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{y}_i)\{\boldsymbol{x}_{i(j)}\}^\top \right\|\\
&= \max_{(\boldsymbol{k},j)\in\mathcal{G}} w^{-1}_{\boldsymbol{k},j} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}) \right\|.
\end{aligned}$$

Applying Hölder's inequality twice, we know that

$$|\langle \nabla\mathcal{L}_n(\boldsymbol{\beta}), \Delta\boldsymbol{\beta}\rangle| \le \sum_{g\in\mathcal{G}} \left\| \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g} \right\| \left\| \Delta\boldsymbol{\beta}_g \right\| \le \Omega^*_{\mathcal{G}}(\nabla\mathcal{L}_n(\boldsymbol{\beta})) \cdot \Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}).$$

$\square$

4.11 Tail Bound Estimates

Before we show Theorem 4.1, we first prove the sub-exponential tail bound for $\Omega^*_{\mathcal{G}}(\mathcal{L}_n(\boldsymbol{\beta}^\dagger))$ for the Poisson categorical response model, and the sub-Gaussian tail bound for the multinomial categorical response model.

Lemma 4.16. *For Poisson categorical response model, there exists absolute constants A_1, A_2 such that for any $u \ge 0$,*

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w^{-1}_{\boldsymbol{k},j} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \right\| \ge u \Big) \le \exp\Big\{ -\frac{A_1 \cdot n}{C^2\Lambda} \big(\min\{u, A_2 C\sqrt{\Lambda}\}\big)^2 + \log(5)m + \log|\mathcal{G}| \Big\}.$$

For multinomial categorical response model, for any $u \geq 0$, we have

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \Big\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \Big\| \geq u \Big) \leq \exp\Big(- \frac{nu^2}{4C^2} + \log(5)m + \log|\mathcal{G}| \Big).$$

Corollary 4.1. *For Poisson categorical response model, there exists absolute constants B, B_1, B_2, B' such that if $\frac{m}{n} + \frac{\log|\mathcal{G}|}{n} \leq B_1$ and $\lambda = BC\sqrt{\Lambda}\Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\Big)$, then*

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \Big\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \Big\| \geq \frac{\lambda}{2} \Big) \leq \exp\big(- B'\delta^2 n\big), 0 \leq \delta \leq B_2.$$

For multinomial categorical response model, there exists absolute constants B, B_1, B' such that if $\frac{m}{n} + \frac{\log|\mathcal{G}|}{n} \leq B_1$ and $\lambda = BC\Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\Big)$, then

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \Big\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \Big\| \geq \frac{\lambda}{2} \Big) \leq \exp\big(- B'\delta^2 n\big), \delta \geq 0.$$

Proof of Lemma 4.16. **Case 1: Poisson categorical response model**

First of all, define the Orlicz space for mean zero random variable

$$L^0_{\psi_1} := \{X; \mathbb{E}X = 0, \exists K > 0, \mathbb{E}\exp(|X/K|) \leq 2\}.$$

Consider the following two norm on Orlicz space $L^0_{\psi_1}$

$$\|X\|_{\psi_1} := \inf\{K > 0; \mathbb{E}\exp(|X/K|) \leq 2\}, \text{ and}$$
$$\tau_{\varphi_1}(X) := \inf\{K > 0; \forall t \in [-1,1], \mathbb{E}\exp(tX) \leq \exp(K^2t^2/2)\}.$$

Assume random variable Y has Poisson distribution with mean parameter $\lambda > 0$. Note that the moment generating function of Y is given by $\mathbb{E}e^{\xi(Y-\lambda)} = \exp(\lambda(e^\xi -$

$\xi - 1)$). Because

$$\frac{d}{d\xi}\frac{e^{\xi}-\xi-1}{\xi^2} = \frac{e^{\xi}(\xi-2)+\xi+2}{\xi^3} = \frac{\int_0^{\xi} du_1 \int_0^{u_1} e^{u_2} u_2 du_2}{\xi^3} > 0, \forall \xi \in \mathbb{R}\backslash\{0\},$$

and

$$\begin{aligned} &\inf\{K>0; \forall t \in [-1,1], \exp(\lambda(e^t - t - 1)) \le \exp(K^2 t^2/2)\} \\ &\quad = \inf\{K>0; \forall t \in [-1,1], 2\lambda \frac{e^t - t - 1}{t^2} \le K^2\} = \sqrt{2(e-2)\lambda}, \end{aligned}$$

we know that $\tau_{\varphi_1}(Y-\lambda) = \sqrt{2(e-2)\lambda}$. Theorem 2.7 in Zajkowski (2020) shows that norms $\|Y-\lambda\|_{\psi_1}$ and $\tau_{\varphi_1}(Y-\lambda)$ are equivalent on the space $L^0_{\psi_1}$. Thus, there exists an absolute constant C_0 such that $\|Y-\lambda\|_{\psi_1} \le C_0\sqrt{\lambda}$.

By the union bound, we obtain

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1}\Big\|\frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\Big\| \ge u\Big) \le \sum_{(\boldsymbol{k},j)\in\mathcal{G}} \mathbb{P}\Big(w_{\boldsymbol{k},j}^{-1}\Big\|\frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\Big\| \ge u\Big).$$

Applying Lemma 5.7 in Wainwright (2019), there exists $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N \in \mathbb{R}^{|\boldsymbol{k}|_{\boldsymbol{J}}\cdot p_j}$ such that $N \le 5^{|\boldsymbol{k}|_{\boldsymbol{J}}\cdot p_j}$, $\|\boldsymbol{v}_l\| \le 1, 1 \le l \le N$, for any $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{k}|_{\boldsymbol{J}}\cdot p_j}$ and $\|\boldsymbol{v}\| \le 1$, there exists $1 \le l \le N$ such that $\|\boldsymbol{v} - \boldsymbol{v}_l\| \le 1/2$. Furthermore,

$$\Big\|\frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\Big\| \le \frac{2}{n}\max_{1\le l\le N}\sum_{i=1}^n \Big\langle \boldsymbol{v}_l, \boldsymbol{H}_{\boldsymbol{k}}^{\top}\big(\boldsymbol{y}_i - e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\big)\{\boldsymbol{x}_{i(j)}\}^{\top}\Big\rangle.$$

Note that

$$\frac{2}{n}\sum_{i=1}^n \Big\langle \boldsymbol{v}_l, \boldsymbol{H}_{\boldsymbol{k}}^{\top}\big(\boldsymbol{y}_i - e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\big)\{\boldsymbol{x}_{i(j)}\}^{\top}\Big\rangle = \frac{2}{n}\sum_{i=1}^n \Big\langle \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}, \boldsymbol{y}_i - e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\Big\rangle.$$

Thus,

$$\mathbb{P}\Big(w_{\boldsymbol{k},j}^{-1}\Big\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\Big\|\geq u\Big)\leq\sum_{1\leq l\leq N}\mathbb{P}\Big(\sum_{i=1}^{n}\Big\langle\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)},\boldsymbol{y}_i-e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\Big\rangle\geq\frac{w_{\boldsymbol{k},j}nu}{2}\Big).$$

For fixed l and j, set $\mathrm{vec}_{\boldsymbol{J}}^{-1}\left(\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right)=\{v_{i,\boldsymbol{j}}\}_{\boldsymbol{j}\in\boldsymbol{J}}$ for any $1\leq i\leq n$. Let $\mathrm{vec}_{\boldsymbol{J}}^{-1}\left(\boldsymbol{y}_i\right)=\{y_{i,\boldsymbol{j}}\}_{\boldsymbol{j}\in\boldsymbol{J}}$ be a sequence of independent Poisson random variables with parameters $\lambda_{i,\boldsymbol{j}},1\leq i\leq n,\boldsymbol{j}\in\boldsymbol{J}$.

We aim to use Bernstein's inequality for sub-exponential random variables to estimate the following tail probability

$$\mathbb{P}\Big(\sum_{i=1}^{n}\Big\langle\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)},\boldsymbol{y}_i-e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\Big\rangle\geq\frac{w_{\boldsymbol{k},j}nu}{2}\Big)=\mathbb{P}\Big(\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\boldsymbol{J}}v_{i,\boldsymbol{j}}(y_{i,\boldsymbol{j}}-\lambda_{i,\boldsymbol{j}})\geq\frac{w_{\boldsymbol{k},j}nu}{2}\Big).\tag{4.33}$$

Notice that $\sum_{\boldsymbol{j}\in\boldsymbol{J}}v_{i,\boldsymbol{j}}^2=\left\|\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|^2\leq\left\|\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|^2\leq\left\|\boldsymbol{x}_{i(j)}\right\|^2$, for any $1\leq l\leq N$, $1\leq j\leq t$, and $1\leq i\leq n$.

It is straightforward to show that

$$\max_{\boldsymbol{j}\in\boldsymbol{J}}|v_{i,\boldsymbol{j}}|=\left\|\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|_{\infty}\leq\left\|\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|\leq\left\|\boldsymbol{x}_{i(j)}\right\|.$$

In conclusion, we know that

$$\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\boldsymbol{J}}\|v_{i,\boldsymbol{j}}(y_{i,\boldsymbol{j}}-\lambda_{i,\boldsymbol{j}})\|_{\psi_1}^2\leq C_0^2\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\boldsymbol{J}}v_{i,\boldsymbol{j}}^2\lambda_{i,\boldsymbol{j}}\leq C_0^2\Lambda\sum_{i=1}^{n}\left\|\boldsymbol{x}_{i(j)}\right\|^2\leq C_0^2C^2w_{\boldsymbol{k},j}^2\Lambda n,$$

and

$$\max_{1\leq i\leq N}\max_{\boldsymbol{j}\in\boldsymbol{J}}\|v_{i,\boldsymbol{j}}(y_{i,\boldsymbol{j}}-\lambda_{i,\boldsymbol{j}})\|_{\psi_1}\leq C_0\sqrt{\Lambda}\max_{1\leq i\leq n}\left\|\boldsymbol{x}_{i(j)}\right\|\leq C_0Cw_{\boldsymbol{k},j}\sqrt{\Lambda},$$

where $\Lambda = \max_{1\le i\le n, \boldsymbol{j}\in \boldsymbol{J}} \lambda_{i,\boldsymbol{j}} = \max_{1\le i\le n} \left\| e^{\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i} \right\|_\infty$.

Applying Theorem 2.8.1 in Vershynin (2018), we obtain

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^n \Big\langle \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{v}_l \boldsymbol{x}_{i(j)}, \boldsymbol{y}_i - e^{\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i} \Big\rangle \ge \frac{(w_{\boldsymbol{k},j}\cdot n)u}{2} \Big) \\
&\le \exp\Big\{ -c\cdot \min\Big(\frac{nu^2}{4C_0^2 C^2 \Lambda}, \frac{nu}{2C_0 C\sqrt{\Lambda}} \Big) \Big\} \\
&\le \exp\Big\{ -\frac{c\cdot n}{4C_0^2 C^2 \Lambda} \big(\min\{u, 2C_0 C\sqrt{\Lambda}\} \big)^2 \Big\}.
\end{aligned}$$

By the union bound, we obtain

$$\begin{aligned}
&\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \right\| \ge u \Big) \\
&\le \exp\Big(-\frac{c\cdot n}{4C_0^2 C^2 \Lambda} \big(\min\{u, 2C_0 C\sqrt{\Lambda}\} \big)^2 + \log(5)m + \log|\mathcal{G}| \Big).
\end{aligned}$$

Let $A_1 = \frac{c}{4C_0^2}$ and $A_2 = 2C_0$ be absolute constants. Then we complete the proof of Lemma 4.16.

Case 2: Multinomial categorical response model

Similar to previous discussion, we obtain

$$\left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \right\| \le \frac{2}{n} \max_{1\le l\le N} \sum_{i=1}^n \left\langle \boldsymbol{v}_l, \boldsymbol{H}_{\boldsymbol{k}}^\top \big(\boldsymbol{y}_i - p_{\boldsymbol{J}}(\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i) \big) \{\boldsymbol{x}_{i(j)}\}^\top \right\rangle,$$

where $p_{\boldsymbol{J}}(\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i) = \frac{1}{\left\langle 1, e^{\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i} \right\rangle} e^{\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i}$.

By Lemma 4.17, we know that $\boldsymbol{y}_i - p_{\boldsymbol{J}}(\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i)$ is a $\sqrt{\frac{1}{2}}$-sub-Gaussian random vector. Applying the Chernoff (Hoeffding) bound for sub-Gaussian random vectors, we obtain that

$$\mathbb{P}\Big(\sum_{i=1}^n \big\langle \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{v}_l \boldsymbol{x}_{i(j)}, \boldsymbol{y}_i - p_{\boldsymbol{J}}(\boldsymbol{H}\boldsymbol{\beta}^\dagger \boldsymbol{x}_i) \big\rangle \ge \frac{(w_{\boldsymbol{k},j}\cdot n)u}{2} \Big) \le \exp\Big(-\frac{nu^2}{4C^2} \Big).$$

By the union bound, we obtain

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \Big\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \Big\| \geq u \Big) \leq \exp\Big(-\frac{nu^2}{4C^2} + \log(5)m + \log|\mathcal{G}| \Big).$$

$\square$

Lemma 4.17. *Assume $(Y_1, \cdots, Y_k)$ denote the multinomial distribution random vector associated with number of trials n and probabilities $(p_1, \cdots, p_k)$ such that $\sum_{i=1}^k p_i = 1$. Let $\boldsymbol{P}_{\mathbf{1}} = I - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top$. Then for any $\boldsymbol{t} \in \mathbb{R}^k$, we have*

$$\mathbb{E}\exp\big(\sum_{i=1}^k t_i(Y_i - np_i) \big) \leq \exp\big(\frac{n\left\|\boldsymbol{P}_{\mathbf{1}}\boldsymbol{t}\right\|^2}{4} \big).$$

Proof of Lemma 4.17. Based on the moment generating function of multinomial distribution random vector, we obtain

$$\log\mathbb{E}\exp\big(\sum_{i=1}^k t_i(Y_i - np_i) \big) = n\Big(\log\Big(\sum_{i=1}^k p_i e^{t_i} \Big) - \sum_{i=1}^k p_i t_i \Big) = n\big(f(1) - f(0) - f'(0) \big), \tag{4.34}$$

where $f(x) = \log(\sum_{i=1}^k p_i e^{xt_i})$.

Applying Taylor series with Lagrange remainder, we obtain

$$f(1) - f(0) - f'(0) \leq \frac{1}{2} \max_{0\leq x\leq 1} f''(x). \tag{4.35}$$

Notice that

$$f''(x) = \frac{-(\sum_{i=1}^k p_i t_i e^{xt_i})^2 + (\sum_{i=1}^k p_i t_i^2 e^{xt_i})(\sum_{i=1}^k p_i e^{xt_i})}{(\sum_{i=1}^k p_i e^{xt_i})^2}.$$

Let $\widetilde{p}_j = \frac{p_j e^{xt_j}}{\sum_{i=1}^k p_i e^{xt_i}}$ for any $1 \leq j \leq k$. It is easy to check that $\widetilde{p}_j \leq 1$ and $\sum_{j=1}^k \widetilde{p}_j = 1$.

We can rewrite $f''(x)$ as below,

$$f''(x) = \big(\sum_{i=1}^{k} t_i\widetilde{p}_i\big)^2 - \big(\sum_{i=1}^{k} t_i^2\widetilde{p}_i\big) = \boldsymbol{t}^\top[\operatorname{diag}(\widetilde{\boldsymbol{p}}) - \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^\top]\boldsymbol{t},$$

with $\widetilde{\boldsymbol{p}} = [\widetilde{p}_1, \cdots, \widetilde{p}_k]^\top$.

It is easy to check that 0 is an eigenvalue of the matrix $[\operatorname{diag}(\widetilde{\boldsymbol{p}}) - \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^\top]$, with the corresponding eigenvector $\mathbf{1}_k$. By Geršgorin circle theorem (see Theorem 1.1 in Varga (2010)), we know that the maximum eigenvalue is upper bounded by

$$\max_{1\le i\le n} \widetilde{p}_i(1-\widetilde{p}_i) + \sum_{j:j\neq i} \widetilde{p}_j\widetilde{p}_i = 2\max_{1\le i\le n} \widetilde{p}_i(1-\widetilde{p}_i) \le \frac{1}{2}.$$

Hence, we know that

$$\boldsymbol{t}^\top[\operatorname{diag}(\widetilde{\boldsymbol{p}}) - \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^\top]\boldsymbol{t} \le \frac{1}{2}\left\|\boldsymbol{P}_{\mathbf{1}}\boldsymbol{t}\right\|^2.$$

Combined with (4.34) and (4.35), we complete the proof of Lemma 4.17. □

4.12 Proof of Theorem 4.1

4.12.1 Proof Overview

The proof of Theorem 4.1 for the $\Omega_{\mathcal{G}}$-regularized estimator $\widehat{\boldsymbol{\beta}}$ follows a strategy similar to that of the non-asymptotic bound presented in Corollary 9.28 of Wainwright (2019).

In the following series of lemmas, we first justify that the error $\Delta\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\dagger$ belongs to a star-shaped set $\mathbb{C}(\mathcal{S}, \phi)$ in Lemma 4.18. Next, we provide a general bound for $\Delta\boldsymbol{\beta}$ in Lemma 4.19. Finally, as a consequence of the above lemmas, we obtain the proofs of Theorem 4.1 and its special case, Theorem 2.2.

4.12.2 A Star-Shaped Set

Define orthogonal projection operator

$$\Big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta})\Big)_{\boldsymbol{k},j} = \boldsymbol{\beta}_{\boldsymbol{k},j}, (\boldsymbol{k},j) \in \mathcal{S} \text{ and } \Big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta})\Big)_{\boldsymbol{k},j} = \boldsymbol{0}, (\boldsymbol{k},j) \in \mathcal{G}\backslash\mathcal{S},$$

where we denote the support of $\boldsymbol{\beta}^{\dagger}$ as $\mathcal{S} = \{g \in \mathcal{G}; \boldsymbol{\beta}^{\dagger}_g \neq 0\}$

As originated from Proposition 9.13 in Wainwright (2019), we define a star-shaped set for any $\phi > 1$ as follows:

$$\mathbb{C}(\mathcal{S},\phi) = \Big\{\Delta\boldsymbol{\beta}; \Omega_{\mathcal{G}}\big(\Delta\boldsymbol{\beta} - \boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big) \leq \phi \cdot \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big), \boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta}) = \Delta\boldsymbol{\beta}\Big\},$$

where $\boldsymbol{Q}_{\mathcal{V}}$ is defined in Lemma 4.9.

We know that if $\Delta\boldsymbol{\beta} \in \mathbb{C}(\mathcal{S},\phi)$, then

$$\Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}) \leq (\phi+1)\sum_{g\in\mathcal{S}} w_g \left\|\Delta\boldsymbol{\beta}_g\right\|. \tag{4.36}$$

Similar to Proposition 9.13 in Wainwright (2019), we obtain the following lemma.

Lemma 4.18. *Conditioned on the event* $\mathbb{G}(\lambda) = \{\Omega^*_{\mathcal{G}}(\nabla\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})) \leq \frac{\phi-1}{\phi+1}\lambda\}$*, the error* $\Delta\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\dagger} \in \mathbb{C}(\mathcal{S},\phi)$*. Here,*

$$\Omega^*_{\mathcal{G}}(\nabla\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})) = \max_{(\boldsymbol{k},j)\in\mathcal{G}} \quad w^{-1}_{\boldsymbol{k},j}\left\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\right\|.$$

Proof of Lemma 4.18. Let $\widehat{\boldsymbol{\beta}}$ be a global minimizer of $\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta})$. Let $\Delta\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\dagger}$. It is easy to verify that $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^{\top} = \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^{\top}\boldsymbol{P}_{\mathcal{V}}$ for any $\boldsymbol{k} \in \mathcal{K}$. By Lemma 4.9, we know that $\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta}) = \Delta\boldsymbol{\beta}$. Set $u = \frac{\phi-1}{\phi+1}$. Combined with Lemma 4.15, we know that

$$
\begin{aligned}
0 &\geq \mathcal{L}_n(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - (\mathcal{L}_n(\boldsymbol{\beta}^\dagger) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)) \\
&\geq \big\langle \Delta\boldsymbol{\beta}, \nabla\mathcal{L}_n(\boldsymbol{\beta}^\dagger)\big\rangle + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)) \\
&\geq -\Omega_{\mathcal{G}}^*(\nabla\mathcal{L}_n(\boldsymbol{\beta}^\dagger)) \cdot \Big(\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})) \\
&\quad + \Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big)\Big) + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)).
\end{aligned}
$$

Because $(I - \boldsymbol{Q}_{\mathcal{S}})(\boldsymbol{\beta}^\dagger) = \mathbf{0}$, by applying triangle inequality and exploiting the decomposability of the regularizer, we obtain

$$
\begin{aligned}
&\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger) \\
\geq&\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta}^\dagger) + (I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) - \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big) - \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta}^\dagger)) \\
=&\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta}^\dagger)\big) + \Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) - \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big) - \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta}^\dagger)\big) \\
=&\Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) - \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big).
\end{aligned}
$$

Thus, we have

$$
0 \geq \lambda\Big((1-u)\Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) - (1+u)\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big)\Big),
$$

which implies

$$
\Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) \leq \frac{1+u}{1-u}\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big) = \phi \cdot \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big).
$$

Combined with Lemma 4.9, we obtain that $\Delta\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\dagger \in \mathbb{C}(\mathcal{S}, \phi)$. $\square$

4.12.3 Non-Asymptotic bound

Lemma 4.19. *Under Assumptions 1-4 and over event* $\mathbb{G}(\lambda) = \{\Omega_{\mathcal{G}}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta}^\dagger)) \leq \frac{\lambda}{2}\}$, *if* $A \cdot C^2\left(\frac{\log|\mathcal{G}|}{n} + \frac{m}{n}\right) \leq \frac{\kappa}{64\Psi^2(\mathcal{S})}$ *and* $\lambda \leq \frac{R\cdot\kappa}{6\Psi(\mathcal{S})}$, *then*

$$\left\|\boldsymbol{\beta}^\dagger - \widehat{\boldsymbol{\beta}}\right\| \leq \frac{6\lambda\Psi(\mathcal{S})}{\kappa}.$$

Proof of Lemma 4.19. Let $\boldsymbol{\theta}^\dagger = \boldsymbol{H}\boldsymbol{\beta}^\dagger$ and $\widehat{\boldsymbol{\theta}} = \boldsymbol{H}\widehat{\boldsymbol{\beta}}$. Define

$$\mathcal{F}(\Delta\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}^\dagger + \Delta\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}^\dagger) + \lambda(\Phi_{\mathcal{G}}(\boldsymbol{\theta}^\dagger + \Delta\boldsymbol{\theta}) - \Phi_{\mathcal{G}}(\boldsymbol{\theta}^\dagger)).$$

Define $\mathbb{K}(\delta) = \{\Delta\boldsymbol{\theta} : \Delta\boldsymbol{\theta} \in \mathcal{M}_H, \|\Delta\boldsymbol{\theta}\| = \delta\} \bigcap \{\boldsymbol{H}\Delta\boldsymbol{\beta} : \Delta\boldsymbol{\beta} \in \mathbb{C}(\mathcal{S}, 3)\}$. Due to $\Delta\boldsymbol{\beta} \in \mathbb{C}(\mathcal{S}, 3)$, we know that $\Delta\boldsymbol{\beta} = \boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})$. It is easy to verify that $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top = \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}$ for any $\boldsymbol{k} \in \mathcal{K}$. Applying Lemma 4.9 and Assumption 4, we know that

$$\begin{aligned}
&\mathcal{F}(\Delta\boldsymbol{\theta}) \\
&\geq \left\langle \nabla L_n(\boldsymbol{\theta}^\dagger), \boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta} \right\rangle + \frac{\kappa}{2}\left\|\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)),
\end{aligned}$$

where $\tau_n^2 = A \cdot C^2\left(\frac{\log|\mathcal{G}|}{n} + \frac{m}{n}\right)$.

By Lemma 4.9, we know that $\boldsymbol{\beta}^\dagger = \boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger)$, $\|\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}\| = \|\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})\|$, and for any $\boldsymbol{\beta}$,

$$\Omega_{\mathcal{G}}(\boldsymbol{\beta}) \geq \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})\big).$$

Thus, we obtain that

$$
\begin{aligned}
&\mathcal{F}(\Delta\boldsymbol{\theta}) \\
&\geq \left\langle \nabla L_n(\boldsymbol{\theta}^\dagger), \boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta} \right\rangle + \frac{\kappa}{2}\left\|\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)) \\
&\geq -\Omega_{\mathcal{G}}^*(\nabla\mathcal{L}_n(\boldsymbol{\beta}^\dagger)) \cdot \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \frac{\kappa}{2}\left\|\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) \\
&\quad + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)) \\
&\geq -\frac{\lambda}{2}\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \frac{\kappa}{2}\left\|\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) \\
&\quad + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)) \\
&= -\frac{\lambda}{2}\Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}) + \frac{\kappa}{2}\left\|\Delta\boldsymbol{\beta}\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\Delta\boldsymbol{\beta}) + \lambda\Big(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)\Big) \\
&\geq \frac{\kappa}{2}\left\|\Delta\boldsymbol{\beta}\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\Delta\boldsymbol{\beta}) + \lambda\Big(\frac{1}{2}\Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) - \frac{3}{2}\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big)\Big) \\
&\geq \frac{\kappa}{2}\left\|\Delta\boldsymbol{\beta}\right\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\Delta\boldsymbol{\beta}) - \frac{3\lambda}{2}\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big).
\end{aligned}
$$

Recall that $\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})) \leq \Psi(\mathcal{S}) \cdot \|\Delta\boldsymbol{\beta}\|$. Notice that for any $\Delta\boldsymbol{\beta} \in \mathbb{C}(\mathcal{S}, 3)$, by inequality (4.36) and Lemma 4.6, we have

$$
\Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}) \leq 4 \cdot \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})) \leq 4 \cdot \Psi(\mathcal{S}) \cdot \left\|\Delta\boldsymbol{\beta}\right\|.
$$

This implies that

$$
\mathcal{F}(\Delta\boldsymbol{\theta}) \geq \Big(\frac{\kappa}{2} - 16\tau_n^2\Psi^2(\mathcal{S})\Big) \cdot \left\|\Delta\boldsymbol{\beta}\right\|^2 - \frac{3\lambda}{2}\Psi(\mathcal{S}) \cdot \left\|\Delta\boldsymbol{\beta}\right\|.
$$

Applying the assumed bound $16\tau_n^2\Psi^2(\mathcal{S}) \leq \frac{\kappa}{4}$, we obtain that

$$
\mathcal{F}(\Delta\boldsymbol{\theta}) \geq \Big(\frac{\kappa}{4}\Big)\Big(\left\|\Delta\boldsymbol{\beta}\right\| - \frac{6\lambda}{\kappa}\Psi(\mathcal{S})\Big)\left\|\Delta\boldsymbol{\beta}\right\| > 0,
$$

if $\|\Delta\boldsymbol{\theta}\| = \|\Delta\boldsymbol{\beta}\| > \frac{6\lambda}{\kappa}\Psi(\mathcal{S})$.

In conclusion, we obtain that if $\delta = \frac{6\lambda}{\kappa}\Psi(\mathcal{S}) + \varepsilon$ ($\varepsilon > 0$), then for any $\Delta\boldsymbol{\theta} \in \mathbb{K}(\delta)$,

$$\mathcal{F}(\Delta\boldsymbol{\theta}) > 0.$$

By Lemma 9.21 in Wainwright (2019), we reach the conclusion that

$$\|\Delta\boldsymbol{\beta}\| \leq \frac{6\lambda}{\kappa}\Psi(\mathcal{S}) + \varepsilon, \qquad \forall\varepsilon > 0,$$

that is

$$\|\Delta\boldsymbol{\beta}\| \leq \frac{6\lambda}{\kappa}\Psi(\mathcal{S}).$$

Recall that

$$\|\Delta\boldsymbol{\theta}\|^2 = \operatorname{tr}\left(\{\Delta\boldsymbol{\theta}\}^\top \Delta\boldsymbol{\theta}\right) = \operatorname{tr}\left(\{\Delta\boldsymbol{\beta}\}^\top \boldsymbol{H}^\top \boldsymbol{H} \Delta\boldsymbol{\beta}\right) = \operatorname{tr}\left(\{\Delta\boldsymbol{\beta}\}^\top \Delta\boldsymbol{\beta}\right) = \|\Delta\boldsymbol{\beta}\|^2.$$

All of the above argument is valid as long as $\frac{6\lambda}{\kappa}\Psi(\mathcal{S}) \leq R$ by Assumption 4. $\square$

4.12.4 Proof of Theorem 4.1

Proof of Theorem 4.1. The proof of Theorem 4.1 is a straightforward application of Lemma 4.19 and Corollary 4.1, under Assumptions 1-4.

$\square$

Chapter 5

Supplemental Material for Distributed Algorithms for High-Dimensional Statistical Inference and Structure Learning with Heterogeneous Data

5.1 Design Matrix of Heterogeneous Linear Model

I will explore a simple heterogeneous linear regression setting below. Assume the i-th observation at the j-th site follows

$$Y_{ij} = \boldsymbol{\beta}_0^T \boldsymbol{x}_i^{(j)} + \boldsymbol{\beta}_j^T \boldsymbol{w}_i^{(j)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_j^2); i = 1, \cdots, n_j, \tag{5.1}$$

where vectors $\boldsymbol{\beta}_0, \boldsymbol{x}_i^{(j)} \in \mathbb{R}^{p_0}$, $\boldsymbol{\beta}_j, \boldsymbol{w}_i^{(j)} \in \mathbb{R}^{p_j}, j = 1, 2, \cdots, K$, and

$$\{\boldsymbol{x}_i^{(j)}, \boldsymbol{w}_i^{(j)}\}_{1 \le j \le K, 1 \le i \le n_j}$$

are independent of error $\{\varepsilon_{ij}\}_{1 \le j \le K, 1 \le i \le n_j}$. Recall that $p = p_0 + \sum_{j=1}^K p_j$. Assume that we already know some positive numbers $r_1, \cdots, r_K$, such that $\sigma_j^2 = r_j^2 \sigma^2, j =$

$1, 2 \cdots, K$. Assume $\sigma^2 = \max_{1 \le j \le K} \sigma_j^2$.

Replacing $Y_{ij}, \boldsymbol{x}_i^{(j)}, \boldsymbol{w}_i^{(j)}, \varepsilon_{ij}$ in (5.1) by $Y_{ij} r_j, \boldsymbol{x}_i^{(j)} r_j, \boldsymbol{w}_i^{(j)} r_j, \varepsilon_{ij} r_j$, respectively, we have

$$Y_{ij} = \boldsymbol{\beta}_0^T \boldsymbol{x}_i^{(j)} + \boldsymbol{\beta}_j^T \boldsymbol{w}_i^{(j)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2); i = 1, \cdots, n_j. \tag{5.2}$$

Rewrite linear model (5.2) as $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, that is

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_K \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_K & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_K \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_K \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_K \end{pmatrix}, \tag{5.3}$$

where error terms $\boldsymbol{\varepsilon}_j \sim N_{n_j}(0, \sigma^2 I_{n_j})$, unknown central server parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ and site-specific nuisance parameters $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}, j = 1, \cdots, K$, response vectors $\boldsymbol{Y}_j = [Y_{1j}, Y_{2j}, \cdots, Y_{n_j j}]^T$, design matrices

$$\mathbf{X}_j^T = [\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, \cdots, \boldsymbol{x}_{n_j}^{(j)}], \text{ and } \mathbf{W}_j^T = [\boldsymbol{w}_1^{(j)}, \boldsymbol{w}_2^{(j)}, \cdots, \boldsymbol{w}_{n_j}^{(j)}], j = 1, 2, \cdots, K.$$

The j-th site log-likelihood function is given by

$$L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j, \sigma^2) = -\frac{1}{2\sigma^2} \left\| \boldsymbol{Y}_j - \mathbf{X}_j \boldsymbol{\beta}_0 - \mathbf{W}_j \boldsymbol{\beta}_j \right\|_2^2 - \frac{n_j}{2} \log(2\pi\sigma^2),$$

and $\|\cdot\|_2$ denotes the Euclidean norm.

5.2 Proof of Theorem 3.1

Proof. We will show that if $\kappa \geq |A^0_{H_0}|$, then $\{i \in [p]\backslash B : \beta^0_i \neq 0\} \subset \{i \in [p]\backslash B : \widehat{\Gamma}_i \neq 0\}$ almost surely, where the estimators $\widehat{\mathbf{\Gamma}}_{H_0}$ and $\widehat{\mathbf{\Gamma}}_{H_1}$ are obtained from Algorithm 3. For H_0, let $A^0 = \{i \in [p]\backslash B : \beta^0_i \neq 0\}$ and $A^{[t]} = \{i \in [p]\backslash B : |\widetilde{\Gamma}^{[t]}_i| \geq \tau\}$. For H_1, let $A^0 = \{i \in [p]\backslash B : \beta^0_i \neq 0\} \cup B$ and $A^{[t]} = \{i \in [p]\backslash B : |\widetilde{\Gamma}^{[t]}_i| \geq \tau\} \cup B$. Set $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{ol}$, where $\widehat{\boldsymbol{\beta}}^{ol}$ is the oracle estimate that minimizes $\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ under the constraint $\boldsymbol{\beta}_{(A^0)^c} = 0$. Set $E = \left\{\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n\right\|_\infty \leq 0.5\lambda\tau\right\} \bigcap \left\{\left\|\boldsymbol{\beta}^0 - \widehat{\boldsymbol{\beta}}^{ol}\right\|_\infty \leq 0.5\tau\right\}$.

We will show that $A^0 \triangle A^{[t]}$ is eventually empty set on event E, which has a probability tending to 1, where $\triangle$ denotes the symmetric difference. By the optimality criterion Lee and Lee (2005) for (3.12) and (3.13), we have

$$\left\langle \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}, -\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\widetilde{\mathbf{\Gamma}}^{[t]})/n + \lambda\tau\nabla\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle \geq 0. \tag{5.4}$$

Rearranging the terms, we have

$$\begin{aligned}
&\left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]})\right\|_2^2/n \\
\leq &\left\langle \widetilde{\mathbf{\Gamma}}^{[t]} - \widehat{\boldsymbol{\beta}}^{ol}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle \\
= &\left\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\backslash A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\backslash A^{[t-1]}}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle \\
&+ \left\langle \widetilde{\mathbf{\Gamma}}_{A^{[t-1]}\backslash A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\backslash A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle \\
&+ \left\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle \\
&+ \left\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle.
\end{aligned} \tag{5.5}$$

Let $\mathbf{X}_{A^0}$ denote the matrix, whose A^0 columns are the same as the A^0 columns of $\mathbf{X}$ and all other columns are zeros. Similarly, we can define $\mathbf{X}_{(A^0)^c}$. Notice that $\mathbf{X} = \mathbf{X}_{A^0} + \mathbf{X}_{(A^0)^c}$, $\widehat{\boldsymbol{\beta}}^{ol}_{(A^0)^c} = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}^{ol} = (\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\mathbf{X}^T_{A^0}\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}} = \mathbf{X}_{A^0}(\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\mathbf{X}^T_{A^0}\boldsymbol{Y}$, $P_{A^0} = \mathbf{X}_{A^0}(\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\mathbf{X}^T_{A^0}$, as well as $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (I - P_{A^0})\boldsymbol{Y} = (I - P_{A^0})\boldsymbol{\varepsilon}$. Note that $\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}} = \mathbf{X}^T_{(A^0)^c}\widehat{\boldsymbol{\varepsilon}}$.

Thus, we know that the fourth term in the last equation of (5.5) vanishes, since

$$\begin{aligned}
&\left\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle \\
&= \left\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T_{(A^0)^c}\widehat{\boldsymbol{\varepsilon}}/n \right\rangle \\
&= \left\langle \mathbf{X}_{(A^0)^c}(\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}), \widehat{\boldsymbol{\varepsilon}}/n \right\rangle = 0.
\end{aligned} \tag{5.6}$$

Note that the third term in the last equation of (5.5) satisfies

$$\left\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \right\rangle = \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1, \tag{5.7}$$

because $\widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c} = 0$.

Recalling (5.5), we have

$$\begin{aligned}
0 &\leq \left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]})\right\|_2^2/n \\
&\leq \left\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\setminus A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\setminus A^{[t-1]}}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\setminus A^{[t-1]}}\right\|_1 \right\rangle \\
&\quad + \left\langle \widetilde{\boldsymbol{\Gamma}}_{A^{[t-1]}\setminus A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\setminus A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n \right\rangle \\
&\quad + \left\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \right\rangle \\
&\leq \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n + \lambda\tau) \\
&\quad + \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n - \lambda\tau).
\end{aligned} \tag{5.8}$$

Rearranging inequality (5.8) implies that

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \cdot (\lambda\tau - \left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n) \\
&\leq \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n + \lambda\tau).
\end{aligned}$$

Over event E, we have

$$\begin{aligned}\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c}-\widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 &\le 3\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}}-\widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1\\ &\le 3\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\cup A^{[t-1]}}-\widehat{\boldsymbol{\beta}}^{ol}_{A^0\cup A^{[t-1]}}\right\|_1.\end{aligned}$$

Recall that $\kappa_1=\left|\{i\in[p]\backslash B; |\widetilde{\Gamma}^{[0]}_i|\ge\tau\}\right|$ and $\kappa_{max}=\max\{\kappa,\kappa_1\}$. Without loss of generality, we can assume that $\widetilde{\mathbf{\Gamma}}^{[0]}_B=\mathbf{0}$. For the base case, we know that

$$\left|A^0\triangle A^{[0]}\right|\le|A^0\backslash B|+|A^{[0]}\backslash A^0|\le 2\kappa_{max}.$$

For the induction step, assume $\left|A^0\triangle A^{[t-1]}\right|\le 2\kappa_{max}$ on event E, for $t\ge 1$. We aim to show that $\left|A^0\triangle A^{[t]}\right|\le 2\kappa_{max}$ over event E.

Applying Assumption 3.1 and (5.8), over event E, we have

$$\begin{aligned} & c_1\left\|\widehat{\boldsymbol{\beta}}^{ol}-\widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2^2\le\left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol}-\widetilde{\mathbf{\Gamma}}^{[t]})\right\|_2^2/n\\ \le & \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}}-\widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1\cdot\left(\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n+\lambda\tau\right)\\ & +\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c}-\widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1\cdot\left(\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n-\lambda\tau\right)\\ \le & \frac{3}{2}\lambda\tau\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}}-\widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1.\end{aligned}\tag{5.9}$$

By Cauchy-Schwarz inequality,

$$\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}}-\widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1^2\le\left|A^0\triangle A^{[t-1]}\right|\left\|\widetilde{\mathbf{\Gamma}}^{[t]}-\widehat{\boldsymbol{\beta}}^{ol}\right\|_2^2.\tag{5.10}$$

Combining (5.9), (5.10) and the third condition in Theorem 3.1,

$$\left\|\widehat{\boldsymbol{\beta}}^{ol}-\widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2\le\frac{3\lambda\tau}{2c_1}\sqrt{|A^0\triangle A^{[t-1]}|}\le\frac{\tau}{4}\sqrt{|A^0\triangle A^{[t-1]}|}.\tag{5.11}$$

Applying (5.11) and induction assumption $\left|A^0 \triangle A^{[t-1]}\right| \leq 2\kappa_{max}$, we have

$$\left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]}\right\|_2 / \tau \leq \frac{1}{4}\sqrt{2\kappa_{max}} \leq \frac{1}{2}\sqrt{\kappa_{max}} \leq \sqrt{\kappa_{max}}. \tag{5.12}$$

Because for any $i \in A^{[t]} \backslash A^0$, $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| = |\widetilde{\Gamma}_i^{[t]}| \geq \tau$, we have

$$\sqrt{|A^{[t]} \backslash A^0|} \leq \left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]}\right\|_2 / \tau \leq \sqrt{\kappa_{\max}}.$$

Thus,

$$\left|A^0 \triangle A^{[t]}\right| \leq |A^0 \backslash B| + |A^{[t]} \backslash A^0| \leq 2\kappa_{max}.$$

By induction, we already showed that over event E, $\left|A^0 \triangle A^{[t]}\right| \leq 2\kappa_{max}$ for any $0 \leq t \leq t_{max}$, where t_{max} denotes the total number of the iteration of Algorithm 3.

Note that Algorithm 3 is terminated at t if

$$\mathrm{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t]}\} \backslash B = \mathrm{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t-1]}\} \backslash B.$$

We will show that over event E, for any $t \geq 1$,

$$\sqrt{|A^0 \triangle A^{[t]}|} \leq \frac{1}{2}\sqrt{|A^0 \triangle A^{[t-1]}|}.$$

If $i \in A^0 \backslash A^{[t]}$, then we know that $\beta_i^0 \neq 0$ and $\widetilde{\Gamma}_i^{[t]} = 0$. Over event E, $|\beta_i^0 - \widehat{\beta}_i^{ol}| \leq 0.5\tau$. According to assumption $\kappa \geq |A_{H_0}^0|$ in Theorem 3.1, we know that

$$|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| \geq |\widetilde{\Gamma}_i^{[t]} - \beta_i^0| - |\beta_i^0 - \widehat{\beta}_i^{ol}| \geq \tau - 0.5\tau = 0.5\tau. \tag{5.13}$$

If $i \in A^{[t]} \backslash A^0$, then we know that $\beta_i^0 = \widehat{\beta}_i^{ol} = 0$ and $|\widetilde{\Gamma}_i^{[t]}| \geq \tau$, which implies $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| = |\widetilde{\Gamma}_i^{[t]}| \geq \tau$.

Over event E, we obtain that for any $i \in A^0 \triangle A^{[t]}$, $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| \geq \tau - 0.5\tau = 0.5\tau$. Combining (5.11), we obtain

$$\sqrt{|A^0 \triangle A^{[t]}|} \leq \frac{1}{0.5\tau} \left\| \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]} \right\|_2 \leq \frac{1}{2} \sqrt{|A^0 \triangle A^{[t-1]}|}. \tag{5.14}$$

In conclusion, over event E, we obtain

$$\sqrt{|A^0 \triangle A^{[t]}|} \leq \frac{1}{2^t} \sqrt{2\kappa_{max}}. \tag{5.15}$$

If $t \geq \lceil \frac{\log(2\kappa_{max})}{\log 4} \rceil$, $\sqrt{|A^0 \triangle A^{[t]}|} < 1$, i.e., $A^0 \triangle A^{[t]} = \emptyset$.

Set $t_{max} = \lceil \frac{\log(2\kappa_{max})}{\log 4} \rceil$. We already showed that over event E,

$$\{i \in [p] \backslash B : |\widetilde{\Gamma}_i^{[t_{max}]}| \geq \tau\} = \{i \in [p] \backslash B : \beta_i^0 \neq 0\}.$$

This means that for any $j \notin \{i \in [p] \backslash B : |\widetilde{\Gamma}_i^{[t_{max}]}| \geq \tau\}$, $|\widetilde{\Gamma}_j^{[t_{max}]}| < \tau$, and for any $k \in \{i \in [p] \backslash B : \widetilde{\Gamma}_i^{[t_{max}]} \geq \tau\}$, $|\widetilde{\Gamma}_k^{[t_{max}]}| \geq \tau$. Thus, over the event E

$$\begin{aligned} \operatorname{supp}\{\boldsymbol{\beta}^0\} \backslash B &= \{i \in [p] \backslash B : \beta_i^0 \neq 0\} \\ &= \{i \in [p] \backslash B : \left|\widetilde{\Gamma}_i^{[t_{max}]}\right| \geq \tau\} \subset \operatorname{supp}\{\widehat{\mathbf{\Gamma}}\} \backslash B. \end{aligned} \tag{5.16}$$

Next, we aim to bound the probability of event E. Let $\boldsymbol{a} := \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}} = (a_1, \cdots, a_p)^T$. It is obvious that $\boldsymbol{a} = \mathbf{X}^T (I - P_{A^0}) \boldsymbol{\varepsilon} = \mathbf{X}_{(A^0)^c}^T (I - P_{A^0}) \boldsymbol{\varepsilon}$. Recall the error $\boldsymbol{\varepsilon} \sim N(0, \Sigma)$ with $\Sigma = \sigma^2 I$. By Assumption 3.2, for any $1 \leq l \leq p$,

$$\operatorname{Var}(a_l) = \boldsymbol{e}_l^T \mathbf{X}^T (I - P_{A^0}) \Sigma (I - P_{A^0}) \mathbf{X} \boldsymbol{e}_l \leq \sigma^2 (\mathbf{X}^T (I - P_{A^0}) \mathbf{X})_{ll} \leq n\sigma^2 c_2^2. \tag{5.17}$$

By the upper-tail inequality for sub-gaussian distribution and the third condition in

Theorem 3.1, we have

$$\begin{aligned}
&\mathbb{P}(\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n > 0.5\lambda\tau) = \mathbb{P}(\|\boldsymbol{a}\|_\infty/n > 0.5\lambda\tau)\\
\leq&\sum_{l=1}^{p}\mathbb{P}(|\boldsymbol{a}|_l > 0.5n\lambda\tau) \leq 2p\exp\left(-\frac{1}{8}\frac{n\lambda^2\tau^2}{\sigma^2 c_2^2}\right) \leq \frac{2}{p^3n^4}.
\end{aligned} \tag{5.18}$$

Set $\boldsymbol{b} := \widehat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0$. A direct calculation implies that

$$\boldsymbol{b} = (\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T\boldsymbol{\varepsilon} = (b_1, \cdots, b_p)^T.$$

By Assumption 3.2, for any $1 \leq l \leq p$,

$$\begin{aligned}
\mathrm{Var}(b_l) =&\boldsymbol{e}_l^T(\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T\Sigma\mathbf{X}_{A^0}(\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\boldsymbol{e}_l\\
\leq&\sigma^2\big((\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\big)_{ll} \leq \frac{1}{n}\sigma^2c_3^2 I(l \in A^0).
\end{aligned}$$

By Assumptions 3.2 and (3.17), we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0\right\|_\infty > 0.5\tau\right) \leq \sum_{i\in A^0}\mathbb{P}(|\boldsymbol{b}|_i > 0.5\tau)\\
\leq&2|A^0|\exp\left(-\frac{1}{8}\frac{n\tau^2}{\sigma^2c_3^2}\right) \leq \frac{2(\kappa_{H_0}^0 + |B|)}{p^4n^4} = o\Big(\frac{1}{p^4n^3}\Big).
\end{aligned} \tag{5.19}$$

Case 1: $\kappa = |A_{H_0}^0|$. In this case, (5.16) implies that over the event E

$$\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B = \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B. \tag{5.20}$$

Thus, under the requirement for (τ, λ) in Theorem 3.1,

$$\begin{aligned}
&\mathbb{P}(\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B)\\
=&\mathbb{P}(\{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B\} \cap E) + \mathbb{P}(\{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B\} \cap E^c).
\end{aligned} \tag{5.21}$$

Notice that $E \subset \{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B = A^{[t_{max}]}\backslash B\}$, and $E \subset \{A^{[t_{max}]}\backslash B = A^0\backslash B\}$. Thus,

$$\mathbb{P}(\{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B\} \cap E) = 0, \text{ and}$$

$$\mathbb{P}(\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B) \leq \mathbb{P}(E^c) \leq \frac{C'}{n^3}, \tag{5.22}$$

where C' is a absolute constant.

Note that $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol}\} = \{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B = A^0\backslash B\}$. By Borel-Cantelli lemma, we have $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol}\}$ almost surely as $n \to \infty$.

It remains to show that $\widehat{\boldsymbol{\beta}}^{ol}$ is a global minimizer of (3.8) or (3.9) with high probability.

Applying Theorem 2 of Shen et al. (2013) and its proof, with the degree of separation condition 3.4, we obtain

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{\beta}}^{\ell_0} \neq \widehat{\boldsymbol{\beta}}^{ol}\right) &\leq \frac{e+1}{e-1}\exp\left(-\frac{n}{18\sigma^2}\left(C_{\min} - 36\frac{\log p + \log n}{n}\sigma^2\right)\right)\\ &\leq \frac{e+1}{e-1}\frac{1}{p^2 n^2},\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}^{\ell_0}$ denotes the global minimizer of (3.8) or (3.9). By Borel-Cantelli lemma, we have $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol} = \widehat{\boldsymbol{\beta}}^{\ell_0}\}$ almost surely as $n \to \infty$.

Case 2: $\kappa \geq |A^0_{H_0}|$. Thus, under the requirement for (τ, λ) in Theorem 3.1,

$$\begin{aligned}&\mathbb{P}(\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B)\\ =&\mathbb{P}(\{\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\} \cap E)\\ &+\mathbb{P}(\{\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\} \cap E^c).\end{aligned}$$

By (5.16), we know that $E \subset \{\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$. Thus,

$$\mathbb{P}(\{\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\} \cap E) = 0 \text{ and}$$

$$\mathbb{P}\big(\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\big) \leq \mathbb{P}(E^c) \leq \frac{C'}{n^3}. \tag{5.23}$$

By the Borel-Cantelli lemma, we have $\{\mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B \subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$ almost surely as $n \to \infty$.

Under H_0, we know that $\mathrm{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_0}\} \subset \mathrm{supp}\{\boldsymbol{\beta}^0\}\backslash B$, which implies that

$$\{\mathrm{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_0}\} \subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$$

almost surely as $n \to \infty$.

Under H_1, we know that $\mathrm{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_1}\} \subset \mathrm{supp}\{\boldsymbol{\beta}^0\} \cup B$, which implies that

$$\{\mathrm{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_1}\} \subset \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \cup B\}$$

almost surely as $n \to \infty$. □

5.3 Proof of Theorem 3.2

Proof. By Theorem 3.1, we have

$$\lim_{n\to\infty} \mathbb{P}(\{\widehat{\boldsymbol{\Gamma}}^{(0)} = \widehat{\boldsymbol{\beta}}^{ls}_{H_0}\} \cap \{\widehat{\boldsymbol{\Gamma}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}_{H_1}\}) = 1, \tag{5.24}$$

where $\widehat{\boldsymbol{\Gamma}}^{(0)}$ and $\widehat{\boldsymbol{\Gamma}}^{(1)}$ are obtained from Algorithm 3, and $\widehat{\boldsymbol{\beta}}^{ls}_{H_0}$ and $\widehat{\boldsymbol{\beta}}^{ls}_{H_1}$ are obtained from (3.10) and (3.11), respectively.

The remainder of the proof follows the argument in Theorem 2 of Zhu et al. (2020), which establishes the sampling distribution of $\Lambda_n(B)$.

□

5.4 Proof of Theorem 3.3

Proof. By Theorem 3.1, we have

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\mathbf{\Gamma}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}) = 1, \tag{5.25}$$

where $\widehat{\mathbf{\Gamma}}^{(1)}$ is obtained from Algorithm 3 and $\widehat{\boldsymbol{\beta}}^{ls} = \widehat{\boldsymbol{\beta}}^{ls}_{H_1}$ is obtained from (3.11), supported on $A^0 \cup B$.

Thus, to show (3.25), it suffices to show

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_B - \boldsymbol{\beta}^0_B) \xrightarrow{d} N(0, \Sigma), \tag{5.26}$$

as $n \to \infty$.

Let $t \in \mathbb{R}^B$ and $u \in \mathbb{R}^{A^0\cup B}$ such that $u_B = t$ and $u_{B^c} = 0$. Note that

$$\widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B} = (\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})\boldsymbol{\beta}^0_{A^0\cup B} + (\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger\mathbf{X}^T_{A^0\cup B}\boldsymbol{\varepsilon}, \tag{5.27}$$

where $\mathbf{X}_{A^0\cup B}$ denotes the sub-matrix with columns of $A^0 \cup B$.

Due to $u \in \{\boldsymbol{\xi} \in \mathbb{R}^{A^0\cup B}; \boldsymbol{\xi}_i = 0, i \notin B\} \subset \mathcal{R}(\mathbf{X}^T_{A^0\cup B})$, we know that there exists vector v such that $u = \mathbf{X}^T_{A^0\cup B}v$ and

$$\begin{aligned}
&(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger u\\
=&\mathbf{X}^T_{A^0\cup B}\Big(\mathbf{X}_{A^0\cup B}(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger\mathbf{X}^T_{A^0\cup B}\Big)v = \mathbf{X}^T_{A^0\cup B}v = u.
\end{aligned}$$

By direct calculation of the characteristic function of $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_B - \boldsymbol{\beta}^0_B)$, we know that

$$\begin{aligned}
&\mathbb{E}e^{it^T\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_B - \boldsymbol{\beta}^0_B)} = \mathbb{E}e^{iu^T\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B} - \boldsymbol{\beta}^0_{A^0\cup B})} \\
=&\mathbb{E}_{\mathbf{X}}\mathbb{E}_{\boldsymbol{\varepsilon}}e^{iu^T\sqrt{n}(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger\mathbf{X}^T_{A^0\cup B}\boldsymbol{\varepsilon}} \\
=&\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2u^Tn(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B}(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger u\} \\
=&\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2u^T(1/n\cdot\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger u\} \\
=&\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2t^T(1/n\cdot\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger_{B,B}t\}.
\end{aligned}$$

Under the assumption of Moore–Penrose inverse $\sigma^2\left(\frac{1}{n}\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B}\right)^\dagger_{B,B}$ converges in distribution to some positive semi-definite matrix Σ, and by applying the continuous mapping theorem, we obtain that

$$\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2t^T(1/n\cdot\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger_{B,B}t\} \to e^{-1/2t^T\Sigma t}, \tag{5.28}$$

as $n\to\infty$, for any $t\in\mathbb{R}^B$.

By Lévy's continuity theorem, we complete the proof of weak convergence (5.26). Therefore, the proof of Theorem 3.3 is completed. $\square$

5.5 Top-κ index set Selection

Algorithm 4 Threshold-Based Top-κ index set Selection

1: **Inputs:**

- Current parameter estimates $\{\widetilde{\Gamma}_{k,j}^{[t]}\}_{(k,j)\in\mathcal{S}}$. Target sparsity level κ.
- Initial thresholds a, b such that

$$\left|\left\{(k,j) : |\widetilde{\Gamma}_{k,j}^{[t]}| \leq a\right\}\right| < \kappa \quad \text{and} \quad \left|\left\{(k,j) : |\widetilde{\Gamma}_{k,j}^{[t]}| \leq b\right\}\right| > \kappa.$$

2: **Initialize:** Set $c \leftarrow \frac{a+b}{2}$.

3: **Site-Level Operations:** Each site counts how many local parameters satisfy $|\widetilde{\Gamma}_{k,j}^{[t]}| \leq c$ and sends *only* this count to a central location.

4: **Aggregation and Update:**

- Sum local counts with the center count to obtain the total: $\left|\{(k,j) : |\widetilde{\Gamma}_{k,j}^{[t]}| \leq c\}\right|$.
- If the total is exactly κ, **break** this iteration.
- If the total is $< \kappa$, set $a \leftarrow c$; otherwise set $b \leftarrow c$. Then update $c \leftarrow \frac{a+b}{2}$.

5: **Check Condition:** If $(b-a)$ is still not sufficiently small, **repeat** steps 3–5. Otherwise, **proceed**.

6: **Finalize Threshold c^*:** Let $c^* \leftarrow c$ if the loop ended early due to an exact match, or set $c^* \leftarrow a$ if we finished the binary search without an exact match.

7: **Construct the ℓ_0 Projected Set:**

- Each site identifies which local parameters exceed $|\widetilde{\Gamma}_{k,j}^{[t]}| > c^*$.
- Let C be the union of those parameter indices across all sites,

$$C = \left\{(k,j) : |\widetilde{\Gamma}_{k,j}^{[t]}| \geq \text{threshold } c^*\right\}.$$

- If $|C| < \kappa$, select additional $\kappa - |C|$ parameters with $|\widetilde{\Gamma}_{k,j}^{[t]}| \in (a,b)$ until $|C| = \kappa$.

8: **Output:** The projected index set C.

References

Agresti, A. (2002). *Categorical Data Analysis* (2 ed.). New York: John Wiley & Sons.

Austin, E., W. Pan, and X. Shen (2020). A new semiparametric approach to finite mixture of regressions using penalized regression via fusion. *Statistica Sinica 30*(2), 783–807.

Barrows Jr, R. C. and P. D. Clayton (1996). Privacy, confidentiality, and electronic medical records. *Journal of the American medical informatics association 3*(2), 139–148.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of statistics 46*(3), 1352–1382.

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

Bergsma, W. P. and T. Rudas (2002). Marginal models for categorical data. *The Annals of Statistics 30*(1), 140–159.

Bhowal, S. and S. Mukherjee (2024). Rates of convergence of the magnetization in the tensor curie–weiss potts model. *Journal of Statistical Physics 192*(1), 2.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Bretto, A. (2013). Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer 1*.

Chen, Y., D. Ge, M. Wang, Z. Wang, Y. Ye, and H. Yin (2017). Strong np-hardness for sparse optimization with concave penalty functions. In *International Conference on Machine Learning*, pp. 740–747. PMLR.

Chen, Y., Y. Ye, and M. Wang (2019). Approximation hardness for a class of sparse optimization problems. *Journal of Machine Learning Research 20*(38), 1–27.

Cheng, A., D. Kessler, R. Mackinnon, T. P. Chang, V. M. Nadkarni, E. A. Hunt, J. Duval-Arnould, Y. Lin, M. Pusic, and M. Auerbach (2017). Conducting multicenter research in healthcare simulation: Lessons learned from the inspire network. *Advances in Simulation 2*(1), 1–14.

Cheng, J., E. Levina, P. Wang, and J. Zhu (2014). A sparse ising model with covariates. *Biometrics 70*(4), 943–953.

Christensen, R. (1997). *Log-linear models and logistic regression*. Springer Science & Business Media.

Daubechies, I., M. Defrise, and C. De Mol (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 57*(11), 1413–1457.

Deng, K., X. Zhang, and A. J. Molstad (2024). Multi-response linear discriminant analysis in high dimensions. *Journal of Machine Learning Research 25*(413), 1–66.

Duan, R., Y. Ning, and Y. Chen (2022). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika 109*(1), 67–83.

Ekeberg, M., C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics 87*(1), 012707.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics 42*(3), 819.

Glonek, G. F. (1996). A class of regression models for multivariate categorical responses. *Biometrika 83*(1), 15–28.

Glonek, G. F. and P. McCullagh (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(3), 533–546.

Guo, Z., X. Li, L. Han, and T. Cai (2025). Robust inference for federated meta-learning. *Journal of the American Statistical Association 0*(0), 1–16.

Han, L., J. Hou, K. Cho, R. Duan, and T. Cai (2025). Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association 0*(ja), 1–25.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Herrera, F., F. Charte, A. J. Rivera, M. J. Del Jesus, F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus (2016). *Multilabel classification*. Springer.

Hirakida, T., H. Kouno, J. Takahashi, and M. Yahiro (2016). Interplay between sign problem and z 3 symmetry in three-dimensional potts models. *Physical Review D 94*(1), 014011.

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

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Keetelaar, S., N. Sekulovski, D. Borsboom, and M. Marsman (2024). Comparing maximum likelihood and maximum pseudolikelihood estimators for the ising model. *advances. in/psychology 2*, e25745.

Khaled, A., K. Mishchenko, and P. Richtárik (2020). Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR.

Kim, S., W. Pan, and X. Shen (2013). Network-based penalized regression with application to genomic data. *Biometrics 69*(3), 582–593.

Klamt, S., U.-U. Haus, and F. Theis (2009). Hypergraphs and cellular networks. *PLoS computational biology 5*(5), e1000385.

Konecnỳ, J., H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492 8*.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *The Annals of Statistics 24*(2), 726–752.

Lange, K. (2016). *MM Optimization Algorithms*, Volume 147. SIAM.

Lee, G. M. and K. B. Lee (2005). Vector variational inequalities for nondifferentiable convex vector optimization problems. *Journal of Global Optimization 32*, 597–612.

Levada, A. L. M., N. D. A. Mascarenhas, and A. Tannús (2009). Pseudo-likelihood equations for Potts model on higher-order neighborhood systems: A quantitative approach for parameter estimation in image analysis. *Brazilian Journal of Probability and Statistics 23*(2), 120 – 140.

Li, C., X. Shen, and W. Pan (2023). Inference for a large directed acyclic graph with unspecified interventions. *Journal of Machine Learning Research 24*(73), 1–48.

Loh, P.-L. and M. J. Wainwright (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics 45*(6), 2455–2482.

Mai, Q., Y. Yang, and H. Zou (2019). Multiclass sparse discriminant analysis. *Statistica Sinica 29*(1), 97–111.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pp. 3–17. Springer.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman and Hall.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR.

Molenberghs, G. and E. Lesaffre (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine 18*(17-18), 2237–2255.

Molstad, A. J. and A. J. Rothman (2023). A likelihood-based approach for multivariate categorical response regression in high dimensions. *Journal of the American Statistical Association 118*(542), 1402–1414.

Molstad, A. J. and X. Zhang (2022). Conditional probability tensor decompositions for multivariate categorical response regression. *arXiv preprint arXiv:2206.10676*.

Mukherjee, S., J. Son, and B. B. Bhattacharya (2022). Estimation in tensor ising models. *Information and Inference: A Journal of the IMA 11*(4), 1457–1500.

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

Obozinski, G., L. Jacob, and J.-P. Vert (2011). Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*.

Okabayashi, S., L. Johnson, and C. J. Geyer (2011). Extending pseudo-likelihood for potts models. *Statistica Sinica*, 331–347.

Palmgren, J. (1989). Regression models for bivariate binary responses. Technical Report 101, Department of Biostatistics, University of Washington, Seattle.

Pan, W., X. Shen, and B. Liu (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of machine learning research 14*(7), 1865–1889.

Parikh, N., S. Boyd, et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization 1*(3), 127–239.

Qaqish, B. F. and A. Ivanova (2006). Multivariate logistic models. *Biometrika 93*(4), 1011–1017.

Razaee, Z. and A. Amini (2020). The potts-ising model for discrete multivariate data. *Advances in Neural Information Processing Systems 33*, 13727–13737.

Read, J., B. Pfahringer, G. Holmes, and E. Frank (2021). Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research 70*, 683–718.

Robiglio, T., L. Di Gaetano, A. Altieri, G. Petri, and F. Battiston (2024). Higher-order ising model on hypergraphs. *arXiv preprint arXiv:2411.19618*.

Roman, S., S. Axler, and F. Gehring (2005). *Advanced linear algebra*, Volume 3. Springer.

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shen, X., W. Pan, and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*(497), 223–232.

Shen, X., W. Pan, Y. Zhu, and H. Zhou (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics 65*(5), 807–832.

Shi, C., R. Song, Z. Chen, and R. Li (2019). Linear hypothesis testing for high dimensional generalized linear models. *Annals of statistics 47*(5), 2671–2703.

Shimagaki, K. and M. Weigt (2019). Selection of sequence motifs and generative hopfield-potts models for protein families. *Physical Review E 100*(3), 032128.

Sidransky, E., M. A. Nalls, J. O. Aasly, J. Aharon-Peretz, G. Annesi, E. R. Barbosa, A. Bar-Shira, D. Berg, J. Bras, A. Brice, et al. (2009). Multicenter analysis of glucocerebrosidase mutations in parkinson's disease. *New England Journal of Medicine 361*(17), 1651–1661.

Tanaka, S., R. Tamura, and N. Kawashima (2011). Phase transition of generalized ferromagnetic potts model-effect of invisible states. In *Journal of Physics: Conference Series*, Volume 297, pp. 012022. IOP Publishing.

Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization 2*(3).

Varga, R. S. (2010). *Geršgorin and his circles*, Volume 36. Springer Science & Business Media.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Vincent, M. and N. R. Hansen (2014). Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis 71*, 771–786.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge university press.

Wang, J., M. Kolar, N. Srebro, and T. Zhang (2017). Efficient distributed learning with sparsity. In *International conference on machine learning*, pp. 3636–3645. PMLR.

Wang, S., T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan (2019). Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications 37*(6), 1205–1221.

Wipf, A., T. Heinzl, T. Kaestner, C. Wozar, et al. (2007). Generalized potts-models and their relevance for gauge theories. *SIGMA. Symmetry, Integrability and Geometry: Methods and Applications 3*, 006.

Wu, C., S. Kwon, X. Shen, and W. Pan (2016). A new algorithm and theory for penalized regression-based clustering. *Journal of Machine Learning Research 17*(188), 1–25.

Wu, C., G. Xu, X. Shen, and W. Pan (2020). A regularization-based adaptive test for high-dimensional glms. *Journal of Machine Learning Research 21*(128), 1–67.

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

Yu, H., Y. Zhong, L. Hong, C. Xu, W. Ren, X. Gong, and H. Xiang (2024). Spin-dependent graph neural network potential for magnetic materials. *Physical Review B 109*(14), 144426.

Yu, S., G. Wang, and L. Wang (2024). Distributed heterogeneity learning for generalized partially linear models with spatially varying coefficients. *Journal of the American Statistical Association*, 1–15.

Yu, T., S. Ye, and R. Wang (2024). High-dimensional variable selection accounting for heterogeneity in regression coefficients across multiple data sources. *Canadian Journal of Statistics 52*(3), 900–923.

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

Zajkowski, K. (2020). On norms in some class of exponential type orlicz spaces of random variables. *Positivity 24*(5), 1231–1240.

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

Zhu, Y., X. Shen, and W. Pan (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association 115*(529), 217–230.