SUBSPACE DECOMPOSITIONS FOR ASSOCIATION STRUCTURE LEARNING IN MULTIVARIATE CATEGORICAL RESPONSE REGRESSION

By Hongru Zhao[1,a], Aaron J. Molstad[1,b] and Adam J. Rothman[1,c]

[1]School of Statistics, University of Minnesota, [a]zhao1118@umn.edu; [b]amolstad@umn.edu; [c]arothman@umn.edu

Modeling the complex relationships between multiple categorical response variables as a function of predictors is a fundamental task in categorical data analysis. However, existing methods can be difficult to interpret and may lack computational efficiency. To address these challenges, we introduce a penalized likelihood method for multivariate categorical response regression that relies on a novel subspace decomposition to uncover interpretable association structures. Our approach models the relationships between categorical responses by identifying mutual, joint, and conditionally independent associations, which yields a linear problem within a tensor product space. We establish theoretical guarantees for our estimator, including error bounds in high-dimensional settings, and validate the method's effectiveness in enhancing both interpretability and prediction accuracy through comprehensive simulation studies.

1. Introduction. 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

$$\pi_{\boldsymbol{j}}(\boldsymbol{x}) := \mathbb{P}(Z_1 = j_1, \cdots, Z_q = j_q \big| \boldsymbol{X} = \boldsymbol{x}) \geq 0, \tag{1}$$

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) 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) using standard methods for the analysis of q-way contingency tables, a classical problem in categorical data analysis [1, 3, 11]. However, when one needs to model (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) using existing methods for multinomial regression (or in statistical learning terminology, multiclass classification). Notice that (1) could be equivalently defined in terms of a "univariate" categorical response variable, $\boldsymbol{Z}^\star$, with $|\boldsymbol{J}|$ 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 $\mathbb{P}(\boldsymbol{Z}^\star = f(\boldsymbol{j}) \mid \boldsymbol{X} = \boldsymbol{x}) = \pi_{\boldsymbol{j}}(\boldsymbol{x}) = \mathbb{P}(Z_1 = j_1, \cdots, Z_q =$

MSC2020 subject classifications: Primary 62H20, 62J12; secondary 62F10, 62H12.
Keywords and phrases: multinomial logistic regression, high-dimensional statistics.

$j_q \big| \boldsymbol{X} = \boldsymbol{x})$ using multinomial logistic regression [1, 23]; linear or quadratic discriminant analysis [6, 10]; 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 amongst 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, the cardinality of $\boldsymbol{J}$, $|\boldsymbol{J}|$, 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 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) that allows practitioners to discover parsimonious and interpretable dependence structures amongst responses.

To motivate our approach, consider a multinomial logistic regression model for (1) with $\boldsymbol{x} \in \mathbb{R}$ (i.e., $p = 1$),

$$(2)\quad \pi_{\boldsymbol{j}}(\boldsymbol{x}) = \mathbb{P}\left(Z_1 = j_1, \ldots, Z_q = j_q | \boldsymbol{X} = \boldsymbol{x}\right) = \frac{\exp(\boldsymbol{x} \cdot \boldsymbol{\zeta}_{\boldsymbol{j}})}{\sum_{\boldsymbol{j} \in \boldsymbol{J}} \exp(\boldsymbol{x} \cdot \boldsymbol{\zeta}_{\boldsymbol{j}})}, \quad \boldsymbol{j} \in \boldsymbol{J}, \quad \sum_{\boldsymbol{j} \in \boldsymbol{J}} \boldsymbol{\zeta}_{\boldsymbol{j}} = 0,$$

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) 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 [13] 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 $q \leq 3$, [13] 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 which are “intermediate" to mutual independence and arbitrary depedence. In this work, we generalize the approach of [13], allowing practicioners to discover much more complex, interpretable dependence structures.

As we just described, to learn the association structure for (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: external and internal. 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 regularizing $\boldsymbol{v}$ towards the subspace $\ker(\boldsymbol{A})$ can be achieved by penalizing the term $\|\boldsymbol{A}\boldsymbol{v}\|_2$. In this sense, [13] achieve structure learning via an external subspace representation. In contrast, for the internal representation, we can set $\mathfrak{L} = \operatorname{span}\left(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s\}\right)$ and $\boldsymbol{v} = v_1\boldsymbol{e}_1 + v_2\boldsymbol{e}_2 + \cdots + v_k\boldsymbol{e}_k$, where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k$ form an orthonormal basis for $\mathbb{R}^k$. Then regularizing $\boldsymbol{v}$ towards the subspace $\operatorname{span}\left(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s\}\right)$ 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

$$\boldsymbol{\zeta}^* = v_{11} g_1 g_1^\top + v_{12} g_1 g_2^\top + v_{21} g_2 g_1^\top + v_{22} g_2 g_2^\top,$$

where $g_i = \frac{1}{\sqrt{2}}\left(1, (-1)^{i+1}\right)^\top, i = 1, 2$. Here, $g_1 g_1^\top$ represents the overall effect, $g_1 g_2^\top$ denotes the main effect of category 1, $g_2 g_1^\top$ is for the main effect of category 2, and $g_2 g_2^\top$ captures the interaction effect between categories 1 and 2. Because $\mathfrak{C}^0 = \operatorname{span}\{g_1 g_1^\top, g_1 g_2^\top, g_2 g_1^\top\}$, 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 amongst 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 [3, 11].

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. We apply group lasso penalty [26] and overlapping group lasso [8, 28] over reparameterization parameters. We apply the accelerated proximal gradient descent algorithm to solve the convex optimization problem. We prove an error bound that illustrates our estimator's performance in high-dimensional settings. A key theoretical advancement in our research is the 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.

We conclude this section by introducing notation to be used for the remainder of the article. First, 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 $\|\boldsymbol{X}\| = \sqrt{\operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{X})}$. 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 $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{Y})$ 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. Let $\mathbf{1}_m$ denote a vector of ones of length m. 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 $\boldsymbol{v}$.

2. Association structure learning via subspace decomposition.

2.1. *Overview.* 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 cardinally 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

$$\operatorname{vec}_{\boldsymbol{J}}(\boldsymbol{y}^{\boldsymbol{J}}) := (y_{1,1,\cdots,1}, y_{2,1,\cdots,1}, \cdots, y_{J_1,1,\cdots,1}, y_{1,2,\cdots,1}, \cdots, y_{J_1,J_2,\cdots,J_q})^\top \in \mathbb{R}^{|\boldsymbol{J}|}. \tag{3}$$

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

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 $\boldsymbol{y}_i = \operatorname{vec}_{\boldsymbol{J}}(\boldsymbol{y}_i^{\boldsymbol{J}}) \in \mathbb{N}^{|\boldsymbol{J}|}$. We assume that $\boldsymbol{y}_1^{\boldsymbol{J}}, \cdots, \boldsymbol{y}_n^{\boldsymbol{J}}$ are independent. Similar to [13], we use a multinomial logistic regression model for the q response variables. Specifically, we assume that $\boldsymbol{y}_i$ is a realization of a multinomial random vector with index $n_i \geq 1$ and category probabilities

$$\operatorname{vec}_{\boldsymbol{J}}\{\boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x}_i)\} = \frac{e^{\boldsymbol{\theta}\boldsymbol{x}_i}}{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta}\boldsymbol{x}_i} \right\rangle} \in \mathbb{R}^{|\boldsymbol{J}|}, \tag{4}$$

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

$$\operatorname{vec}_{\boldsymbol{J}}\{\boldsymbol{\mu}^{\boldsymbol{J}}(\boldsymbol{x}_i)\} = e^{\boldsymbol{\theta}\boldsymbol{x}_i} \in \mathbb{R}^{|\boldsymbol{J}|}, \tag{5}$$

and category probabilities $\operatorname{vec}_{\boldsymbol{J}}\{\boldsymbol{\pi}^{\boldsymbol{J}}(\boldsymbol{x}_i)\} = e^{\boldsymbol{\theta}\boldsymbol{x}_i}\{\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta}\boldsymbol{x}_i} \right\rangle\}^{-1}$, where $\boldsymbol{\mu}^{\boldsymbol{J}}(\boldsymbol{x}) = \{\mu_{\boldsymbol{j}}(\boldsymbol{x})\}_{\boldsymbol{j} \in \boldsymbol{J}}$ 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

$$\ell_{\text{Mult}}(\boldsymbol{\theta}\boldsymbol{x}_i, \boldsymbol{y}_i) = -\left\langle \boldsymbol{y}_i, \boldsymbol{\theta}\boldsymbol{x}_i \right\rangle + n_i \cdot \log\left(\left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta}\boldsymbol{x}_i} \right\rangle \right), \qquad n_i = \left\langle \mathbf{1}_{|\boldsymbol{J}|}, \boldsymbol{y}_i \right\rangle, \tag{6}$$

and

$$\ell_{\text{Pois}}(\boldsymbol{\theta}\boldsymbol{x}_i, \boldsymbol{y}_i) = -\left\langle \boldsymbol{y}_i, \boldsymbol{\theta}\boldsymbol{x}_i \right\rangle + \left\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{\theta}\boldsymbol{x}_i} \right\rangle. \tag{7}$$

The function in (6) is sometimes referred to as cross-entropy loss.

When considering the multinomial categorical response model, we impose the constraint $\mathbf{1}_{|\boldsymbol{J}|}^{\top}\boldsymbol{\theta} = \mathbf{0}$ to address the identifiability issue. 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 $\boldsymbol{P}_{\mathcal{V}} = (I - |\boldsymbol{J}|^{-1}\mathbf{1}_{|\boldsymbol{J}|}\mathbf{1}_{|\boldsymbol{J}|}^{\top})$, where I denotes the identity matrix of order $|\boldsymbol{J}|$. Notice that $\ell_{\text{Mult}}(\boldsymbol{\theta}\boldsymbol{x}, \boldsymbol{y}) = \ell_{\text{Mult}}(\boldsymbol{\theta}'\boldsymbol{x}, \boldsymbol{y})$ for any $(\boldsymbol{x}, \boldsymbol{y})$ if and only if $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta} = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}'$.

2.2. *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 variable are indepedendent. We demonstrate how to construct such bases $\boldsymbol{H}$ in the following example.

EXAMPLE 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 component and $J_2 = 3$ categories for the second. We can write $\boldsymbol{\theta} = (a_{11}, \ldots, a_{J_1 J_2})^{\top} \in \mathbb{R}^{J_1 J_2}$ as

$$\left(\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{\theta}) \right)_{j_1, j_2} = a_{j_1, j_2}, \tag{8}$$

where $a_{j_1,j_2} \in \mathbb{R}$ for any $(j_1, j_2) \in \boldsymbol{J}$. Accordingly, we can rewrite (8) as

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

where $\boldsymbol{e}_{j_i}^{J_i}$ is the j_i-th standard basis vector for $\mathbb{R}^{J_i}$ (i.e., the j_i-th column of I_{J_i}) for $i \in [2]$, and $\boldsymbol{E}_{j_1 j_2}^{\boldsymbol{J}}$ 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 S1 of [27].

Due to the bilinearity of 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 $\{\otimes_t(\boldsymbol{e}_{j_2}^{J_2}, \boldsymbol{e}_{j_1}^{J_1}); 1 \le j_1 \le J_1, 1 \le j_2 \le J_2\}$.

Applying the isomorphism T onto each subspace, we obtain that $T(\mathcal{R}(\boldsymbol{V}_2) \otimes \mathcal{R}(\boldsymbol{V}_1)) = \mathcal{R}(\boldsymbol{V}_2 \otimes \boldsymbol{V}_1)$, where $\boldsymbol{V}_i \in \{\frac{1}{\sqrt{J_i}}\mathbf{1}_{J_i}, \boldsymbol{U}_{J_i}\}$ for $i \in [2]$. Hence, for $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|}$, we can write

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

for vectors $\boldsymbol{\alpha}_{\boldsymbol{V}_2, \boldsymbol{V}_1}$ of appropriate size. More simply, we may write

$$\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}, \text{ where } \boldsymbol{H}_{\{0\}} = \frac{1}{\sqrt{|\boldsymbol{J}|}} \mathbf{1}_{|\boldsymbol{J}|}, \quad \boldsymbol{H}_{\{1\}} = \frac{1}{\sqrt{J_2}} \mathbf{1}_{J_2} \otimes \boldsymbol{U}_{J_1}, \quad \boldsymbol{H}_{\{2\}} = \boldsymbol{U}_{J_2} \otimes \frac{1}{\sqrt{J_1}} \mathbf{1}_{J_1},$$

$$\boldsymbol{H}_{\{1,2\}} = \boldsymbol{U}_{J_2} \otimes \boldsymbol{U}_{J_1}, \quad \mathcal{K} = \{\{0\}, \{1\}, \{2\}, \{1,2\}\},$$

where each $\boldsymbol{\beta}_{\boldsymbol{k}}$ is simply the corresponding $\boldsymbol{\alpha}_{\boldsymbol{V}_2, \boldsymbol{V}_1}$. As we will formalize in Lemma 3, Lemma 2, and Theorem 3, $\operatorname{span}(\boldsymbol{H}_{\{0\}})$ is the subspace for overall effect, $\operatorname{span}(\boldsymbol{H}_{\{1\}}), \operatorname{span}(\boldsymbol{H}_{\{2\}})$ are the subspaces for marginal effect on Z_1 and Z_2, respecticely, and $\operatorname{span}(\boldsymbol{H}_{\{1,2\}})$ is the subspace for joint effect on (Z_1, Z_2). Because of this, sparsity in 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 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*

$$T_{\boldsymbol{J}}\Big(\otimes_t(\boldsymbol{v}_q, \cdots, \boldsymbol{v}_1)\Big) = \boldsymbol{v}_q \otimes \cdots \otimes \boldsymbol{v}_1; \tag{9}$$

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

$$T_{\boldsymbol{J}}\Big(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)\Big) = \mathcal{R}\Big(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1\Big). \tag{10}$$

In Lemma 1, the isomorphism $T_{\boldsymbol{J}}$ is not only between vector spaces $\mathbb{R}^{J_q} \otimes \cdots \otimes \mathbb{R}^{J_1}$ and $\mathbb{R}^{|\boldsymbol{J}|}$, i.e.,

$$T_{\boldsymbol{J}}(\boldsymbol{v}+\boldsymbol{w}) = T_{\boldsymbol{J}}(\boldsymbol{v}) + T_{\boldsymbol{J}}(\boldsymbol{w}), \quad T_{\boldsymbol{J}}(a\boldsymbol{v}) = a \cdot T_{\boldsymbol{J}}(\boldsymbol{v}), \quad \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{J_q} \otimes \cdots \otimes \mathbb{R}^{J_1}, \quad a \in \mathbb{R},$$

but also preserves the bilinear function, i.e., the statement of (9) holds. Here, the Kronecker product serves as the bilinear function.

To summarize Example 1 and Lemma 1, we define $\boldsymbol{U}_m$ as any matrix such that $[\frac{1}{\sqrt{m}}\mathbf{1}_m, \boldsymbol{U}_m]$ 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 $\mathcal{K}_s = \{\boldsymbol{k} = \{k_1,\cdots,k_s\} \subset [q]; 1 \le k_1 < k_2 < \cdots < k_s \le q\}$ 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 effects 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 s-th order effects 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

$$\boldsymbol{H}_0 = \frac{\mathbf{1}_{|\boldsymbol{J}|}}{\sqrt{|\boldsymbol{J}|}} = \frac{\mathbf{1}_{J_q}}{\sqrt{J_q}} \otimes \frac{\mathbf{1}_{J_{q-1}}}{\sqrt{J_{q-1}}} \otimes \cdots \otimes \frac{\mathbf{1}_{J_2}}{\sqrt{J_2}} \otimes \frac{\mathbf{1}_{J_1}}{\sqrt{J_1}}.$$

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

$$\boldsymbol{H}_{\boldsymbol{k}} = \boldsymbol{V}_q \otimes \boldsymbol{V}_{q-1} \otimes \cdots \otimes \boldsymbol{V}_2 \otimes \boldsymbol{V}_1, \qquad \boldsymbol{V}_i = \begin{cases} \boldsymbol{U}_{J_i} & i \in \boldsymbol{k} \\ \frac{\mathbf{1}_{J_i}}{\sqrt{J_i}} & i \in [q]\backslash\boldsymbol{k} \end{cases}. \tag{11}$$

Following from Example 1, we see that $\mathrm{span}(\boldsymbol{H}_{\boldsymbol{k}})$ is the subspace corresponding to the $\boldsymbol{k} = \{k_1, k_2, \cdots, k_s\}$-joint effects. Importantly, it can be verified that the columns of $\boldsymbol{H}_{\boldsymbol{k}}$ are orthonormal.

LEMMA 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 S1 of [27]. 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 has two key implications. First, it ensures invariance of our proposed estimator (Subsection 4.2), and second, it allows us to identify the effect encoded by the span on of each $\boldsymbol{H}_{\boldsymbol{k}}$, as we discuss in the following subsection.

2.3. *Reparameterization via subspace decomposition.*

When fitting (4), it is common to restrict the hypothesis space of models to include joint effects of at most order d for some $0 \le d \le q$. For example, if $q = 3$, we only want to consider models with all possible main effects and two-way interaction effects. In this case, we would take $d = 2$. As such, define the space of possible effects is $\mathcal{K} = \cup_{s=0}^{d}\mathcal{K}_s$. We call $\mathcal{K}$ the association index space. Though $\mathcal{K}$ is a function of d, the maximial order of effect considered, we omit notion indicating this dependence for improved display.

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

THEOREM 3 (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 $\boldsymbol{\beta_k}$ be a matrix in $\mathbb{R}^{|\boldsymbol{k}|_J \times p}$. For any $\mathcal{I} = \{i_1, \cdots, i_s\} \subset [q]$ such that $1 \le i_1 < \cdots < i_s \le q$, define $\boldsymbol{J}^{\mathcal{I}} = [J_{i_1}] \times \cdots \times [J_{i_s}]$.*

1. *(Mutual and joint independence) Let $\mathcal{S}_{\text{joint}} = \{\boldsymbol{k} \in \mathcal{K} : \exists\, i \in [m] \text{ such that } \boldsymbol{k} \subset \mathcal{I}_i\}$. If $\boldsymbol{\beta_k} = \boldsymbol{0}$ for all $\boldsymbol{k} \notin \mathcal{S}_{\text{joint}}$, i.e., the parameter $\boldsymbol{\theta}$ under either Poisson (5) categorical response model or multinomial (4) categorical response model is given by $\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H_k \beta_k} = \sum_{\boldsymbol{k} \in \mathcal{S}_{\text{joint}}} \boldsymbol{H_k \beta_k}$, then*

$$\pi_{\boldsymbol{j}}(\boldsymbol{x}) = \prod_{l=1}^{m} \pi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}), \tag{12}$$

   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 pmf of the responses corresponding to $\mathcal{I}_l$.
2. *(Conditional independence) Let $\mathcal{S}_{\text{joint}|\mathcal{I}_{\text{m}}} = \{\boldsymbol{k} \in \mathcal{K} : \exists\, i \in [m-1] \text{ such that } \boldsymbol{k} \subset \mathcal{I}_i \cup \mathcal{I}_m\}$. If $\boldsymbol{\beta_k} = \boldsymbol{0}$ for all $\boldsymbol{k} \notin \mathcal{S}_{\text{joint}|\mathcal{I}_{\text{m}}}$, i.e., the parameter $\boldsymbol{\theta}$ under either Poisson (5) categorical response model or multinomial (4) categorical response model is given by $\boldsymbol{\theta} = \sum_{\boldsymbol{k} \in \mathcal{K}} \boldsymbol{H_k \beta_k} = \sum_{\boldsymbol{k} \in \mathcal{S}_{\text{joint}|\mathcal{I}_{\text{m}}}} \boldsymbol{H_k \beta_k}$, then*

$$\pi_{\boldsymbol{j}_{\mathcal{I}_1}, \boldsymbol{j}_{\mathcal{I}_2}, \cdots, \boldsymbol{j}_{\mathcal{I}_{m-1}} | \boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \prod_{l=1}^{m-1} \pi_{\boldsymbol{j}_{\mathcal{I}_l},+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}), \tag{13}$$

   where

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

To illustrate the practical implications and applications of Theorem 3, we present the following example. This example is specifically designed to clarify the theorem's underlying principles and to showcase its utility within a hierarchical model, see Section 3.3.

EXAMPLE 2. Suppose $q = 4$. The following types of dependence structures—akin to those in Chapter 6 of [11]—are encoded in the sparsity of the $\boldsymbol{\beta_k}$. 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 $\boldsymbol{x} \in \mathcal{X}$

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

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

The neat interpretations in Example 2 rely partly on a hierarchical structure of the effects. That is, high-order effects are included only if all the corresponding low-order effects are included. Formally, if effect $\boldsymbol{k}$ 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 effect $\{1,2,3\}$ is included in the model, then for the hierarchy to be enforced, the effects $\{0\}, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}$, 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

$$\left\{ \boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}|\times p} : \mathcal{R}(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}\in\mathcal{N}} \mathcal{R}(\boldsymbol{H}_{\boldsymbol{k}}) \subset \mathbb{R}^{|\boldsymbol{J}|},\ \mathcal{N} \subset \mathcal{K} \text{ s.t. } \boldsymbol{k} \in \mathcal{N} \implies \mathcal{P}(\boldsymbol{k}) \subset \mathcal{N} \right\}. \tag{14}$$

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.

3. Penalized likelihood-based association learning.

3.1. *Penalized maximum likelihood estimation.* Define the negative log-likelihoods as L_n^{Mult}, and its reparametarized versions $\mathcal{L}_n^{\text{Mult}}$ 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
$$\mathcal{L}_n(\boldsymbol{\beta}) = \begin{cases} \mathcal{L}_n^{\text{Mult}}(\boldsymbol{\beta}) : \ell = \ell_{\text{Mult}} \\ \mathcal{L}_n^{\text{Pois}}(\boldsymbol{\beta})\ : \ell = \ell_{\text{Pois}} \end{cases}.$$
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
$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \left\{ \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega(\boldsymbol{\beta}) \right\}, \tag{15}$$
for convex penalties $\Omega : \mathbb{R}^{\sum_{s=0}^{d} L_s \times p} \to [0, \infty)$ to be discussed in the next subsection.

3.2. *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, we parameterize $\boldsymbol{\theta}$ as
$$\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} =: \boldsymbol{H}\boldsymbol{\beta}, \tag{16}$$

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 $\boldsymbol{\beta}_{\boldsymbol{k}} = [\boldsymbol{\beta}_{\boldsymbol{k},1}, \boldsymbol{\beta}_{\boldsymbol{k},2}, \cdots, \boldsymbol{\beta}_{\boldsymbol{k},t}]$ and vector $\boldsymbol{x}_i = (\boldsymbol{x}_{i(1)}^{\top}, \dots, \boldsymbol{x}_{i(t)}^{\top})^{\top} \in \mathbb{R}^p$ so that

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

As discussed in the previous section, if $\boldsymbol{\beta}_{\boldsymbol{k}} = 0$, then the corresponding effect 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 effects such that all predictors contribute to the effect, or none contribute to the effect. 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 [26] with a positive sequence $\{w_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{K}}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively, as

$$\Omega_{\text{global}}(\boldsymbol{\beta}) = \sum_{\boldsymbol{k}\in\mathcal{K}} w_{\boldsymbol{k}} \left\|\boldsymbol{\beta}_{\boldsymbol{k}}\right\|, \tag{17}$$

$$\Phi_{\text{global}}(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}=\boldsymbol{H}\boldsymbol{\beta}} \Omega_{\text{global}}(\boldsymbol{\beta}) = \Omega_{\mathcal{G}}(\boldsymbol{H}^{\top}\boldsymbol{\theta}). \tag{18}$$

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 (18) 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 effect, or no predictors contribute to an effect. 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 [13]. Our local association learning is much more general: we can discover which predictors modify certain high-order effects, and which predictors (or groups of predictors) only affect lower-order effects.

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

$$\Omega_{\text{local}}(\boldsymbol{\beta}) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}_{\text{local}}} w_{\boldsymbol{k},j} \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\|, \tag{19}$$

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 effect 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.

3.3. *Association learning with hierarchical constraints.* As mentioned in Section 2.3, it is often desirable to enforce a hierarchical structure for the effects. 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 effect $\boldsymbol{k}$ 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

$$\Omega_{\mathcal{G}}^{H}(\boldsymbol{\beta}) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j} \sqrt{\sum_{\boldsymbol{k}':\boldsymbol{k}\subset\boldsymbol{k}'} \|\boldsymbol{\beta}_{\boldsymbol{k}',j}\|^2} \tag{20}$$

and

$$\Phi_{\mathcal{G}}^{H}(\boldsymbol{\theta}) = \Omega_{\mathcal{G}}^{H}(\boldsymbol{H}^{\top}\boldsymbol{\theta}). \tag{21}$$

The term $\sqrt{\sum_{\boldsymbol{k}':\boldsymbol{k}\subset\boldsymbol{k}'} \|\boldsymbol{\beta}_{\boldsymbol{k}',j}\|^2}$ is a group lasso penalty on the entire set of coefficients corresponding to effects that include $\boldsymbol{k}$ in their powerset. For example, if $q = 3$ and $\boldsymbol{k} = \{1\}$, then $\sqrt{\sum_{\boldsymbol{k}':\boldsymbol{k}\subset\boldsymbol{k}'} \|\boldsymbol{\beta}_{\boldsymbol{k}',j}\|^2} = \sqrt{\|\boldsymbol{\beta}_{\{1\},j}\|^2 + \left\|\boldsymbol{\beta}_{\{1,2\},j}\right\|^2 + \|\boldsymbol{\beta}_{\{1,3\},j}\|^2 + \|\boldsymbol{\beta}_{\{1,2,3\},j}\|^2}$. 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 effects $\widehat{\boldsymbol{\beta}}_{\{1,2\},j} = \widehat{\boldsymbol{\beta}}_{\{1,3\},j} = \widehat{\boldsymbol{\beta}}_{\{1,2,3\},j} = 0$ as well. See [25] for a comprehensive review of how hierarchical structures can be enforced with the overlapping group lasso and related penalties.

4. Relation to existing work.

4.1. *Alternative parametric links.* Multivariate categorical response regression is a classical problem in categorical data analysis (e.g., see Chapter 6 of [11]). 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 contigency table).

One popular 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 $Z_1, \cdots, Z_q$ [4, 11, 12]. For a given $\boldsymbol{\pi}$, the transformation $\boldsymbol{\pi} \to \boldsymbol{\eta}$ can be expressed as a matrix equation:

$$\boldsymbol{\eta} = C \log(M\boldsymbol{\pi}), \tag{22}$$

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 [9]. According to the definition of [2], a numerical value assigned to $\boldsymbol{\eta}$ is considered *strongly compatible* if there exists a valid probability distribution $\boldsymbol{\pi}$ that corresponds to it. [17] showed that excluding the cases when $q = 2$ with $J_1 = J_2$, no explicit solution is available. [5] pointed out the difficulty in solving (22) 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 $Z_1, \cdots, Z_q$, [19] 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. Desirata (i) is addressed by Theorem 3, 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).

4.2. *Generalizing log-linear models for contigency 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 effects 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

$$\log(\mu_{j_1 j_2}) = \boldsymbol{\Lambda}_{j_1,j_2} := \mu + \mu_{j_1}^{Z_1} + \mu_{j_2}^{Z_2} + \mu_{j_1 j_2}^{Z_1 Z_2} \tag{23}$$

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 effects of variables Z_1 and Z_2, respectively, and $\mu_{j_1 j_2}^{Z_1 Z_2}$ denotes the interaction effect between Z_1 and Z_2. Under a multinomial sampling scheme, the model can be written as

$$\pi_{j_1 j_2} = \frac{\exp(\boldsymbol{\Lambda}_{j_1,j_2})}{\sum_{j_1,j_2}\exp(\boldsymbol{\Lambda}_{j_1,j_2})}. \tag{24}$$

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

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 effects and interaction effects for each variable is set to zero. For example, for the main effects, 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 effects:

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

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 maxmimum 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 $\boldsymbol{U}'_m = [\boldsymbol{e}_2^m, \cdots, \boldsymbol{e}_m^m]$. Similar to $\boldsymbol{H}_{\boldsymbol{k}}$ defined in (11), for any $\boldsymbol{k} = \{k_1, \cdots, k_s\} \in \mathcal{K}_s, s \geq 1$, define

$$\boldsymbol{H}'_{\boldsymbol{k}} = \boldsymbol{V}_q \otimes \boldsymbol{V}_{q-1} \otimes \cdots \otimes \boldsymbol{V}_2 \otimes \boldsymbol{V}_1, \qquad \boldsymbol{V}_i = \begin{cases} \boldsymbol{U}'_{J_i} & i \in \boldsymbol{k} \\ \boldsymbol{1}_{J_i} & i \in [q]\backslash\boldsymbol{k} \end{cases}. \tag{25}$$

We can thus rewrite (23) in matrix form as

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

where $\{\boldsymbol{H}'_{\boldsymbol{k}}\}_{\boldsymbol{k}}$ are defined in (25) with $q=2$. Similarly, recall that $\boldsymbol{\pi}^{\boldsymbol{J}} = \{\pi_{j_1,j_2}\}_{(j_1,j_2)\in\boldsymbol{J}}$ for the multinomial log-linear model so that

$$\mathrm{vec}_{\boldsymbol{J}}\{\boldsymbol{\pi}^{\boldsymbol{J}}\} = \frac{\exp(\boldsymbol{\theta})}{\langle \exp(\boldsymbol{\theta}), \mathbf{1}_{J_1J_2}\rangle}, \quad \boldsymbol{\theta} = \boldsymbol{H}'_{\{0\}}\boldsymbol{\beta}'_{\{0\}} + \boldsymbol{H}'_{\{1\}}\boldsymbol{\beta}'_{\{1\}} + \boldsymbol{H}'_{\{2\}}\boldsymbol{\beta}'_{\{2\}} + \boldsymbol{H}'_{\{1,2\}}\boldsymbol{\beta}'_{\{1,2\}}.$$

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, $\mu^{Z_1}_{j_1}$, $\mu^{Z_2}_{j_2}$, and $\mu^{Z_1Z_2}_{j_1j_2}$, 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 under the choice of identifability constraints. To be more specific, if $\max_{i\in[q]} J_i > 2$, $\boldsymbol{U}''_m = [\boldsymbol{e}^m_1, \cdots, \boldsymbol{e}^m_{m-1}]$, and we define $\boldsymbol{H}''$ accordingly, then

$$\boldsymbol{H}'_{\boldsymbol{k}}\boldsymbol{\beta}'_{\boldsymbol{k}} \not\equiv \boldsymbol{H}''_{\boldsymbol{k}}\boldsymbol{\beta}''_{\boldsymbol{k}}, \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}'_{\boldsymbol{k}}\| \not\equiv \sum_{\boldsymbol{k}} \|\boldsymbol{\beta}''_{\boldsymbol{k}}\|, \text{where } \boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}'_{\boldsymbol{k}}\boldsymbol{\beta}'_{\boldsymbol{k}} = \sum_{\boldsymbol{k}\in\mathcal{K}} \boldsymbol{H}''_{\boldsymbol{k}}\boldsymbol{\beta}''_{\boldsymbol{k}}.$$

Choosing $\boldsymbol{H}''$ instead of $\boldsymbol{H}'$ changes how the $\boldsymbol{k}$-joint effect 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 parameter 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 choice 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 (11), then

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

4.3. *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 14], or classifier chains [21].

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 $Z_1, \ldots, Z_q \mid \boldsymbol{X}$, as can the coefficients from our fitted model. Moreover, the order of effects in the latent class method cannot, generally speaking, be easily identified unless the effect 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 [7], namely, classifier chains [21]. A classifier chain estimates $Z_1, \ldots, Z_q \mid \boldsymbol{X}$ 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 effects in a straightforward way, which is the primary motivation for our work.

5. Computation.

In this section, we propose a proximal gradient descent algorithm—described in Chapter 4 of [18]—to calculate the group lasso estimator, and the overlapping group lasso estimator.

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

$$\mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega(\boldsymbol{\beta}) \leq \mathcal{L}_n(\boldsymbol{\beta}^k) + \left\langle \nabla\mathcal{L}_n(\boldsymbol{\beta}^k), \boldsymbol{\beta} - \boldsymbol{\beta}^k \right\rangle + \frac{L}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^k\right\|^2 + \lambda\Omega(\boldsymbol{\beta}) \tag{26}$$

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

$$\boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2}\left\|\boldsymbol{\beta} - \left\{\boldsymbol{\beta}^k - \frac{1}{L}\nabla\mathcal{L}_n(\boldsymbol{\beta}^k)\right\}\right\|^2 + \frac{\lambda}{L}\Omega(\boldsymbol{\beta})\right], \tag{27}$$

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 $\Omega(\boldsymbol{\beta})$ is group lasso penalty, then the proximal problem (27) has closed form solution

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

In Lemma S6 of [27], we show that for all $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$, $\left\|\nabla\mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta}) - \nabla\mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta}')\right\| \leq \frac{\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X})}{2n}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$ with $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$, which implies that with $L \geq \frac{\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X})}{2n}$, (26) will hold. However, in the case of a Poisson categorical response model, the inequality (26) cannot hold globally for any L. Therefore, we use a proximal gradient descent algorithm with the step size determined adaptively by a backtracking line search.

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 S2 of [27]. To summarize, we discuss Algorithm 1, Algorithm 2, and Algorithm 3. These algorithms are motivated by Section 4.3 in [18] and Algorithm 2 in [22]. We present them in Section S2 of our supplementary materials [27]. These algorithms are designed for the accelerated proximal gradient descent with backtracking search for the step size.

6. Statistical Properties.

In this section, we examine the statistical properties of the group lasso estimator, as defined in (15), considering variations in n, p, and $\boldsymbol{J}$. Let $\boldsymbol{\theta}^* = \boldsymbol{H}_{\mathrm{full}}\boldsymbol{\beta}^*$ represent the data generation parameter, where $\boldsymbol{H}_{\mathrm{full}} = \{\boldsymbol{H}_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\cup_{s=0}^{q}\mathcal{K}_s} \in \mathbb{R}^{|\boldsymbol{J}|\times|\boldsymbol{J}|}$. 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,

$$\begin{aligned}\mathcal{F}_{\boldsymbol{\theta}^*} &= \left\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{\sum_{i=0}^{q} L_i \times p}; \ell(\boldsymbol{\theta}^*\boldsymbol{x}, \boldsymbol{y}) = \ell(\boldsymbol{H}_{\mathrm{full}}\widetilde{\boldsymbol{\beta}}\boldsymbol{x}, \boldsymbol{y}), \forall(\boldsymbol{x}, \boldsymbol{y})\right\} \\ &= \left\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{\sum_{i=0}^{q} L_i \times p}; \boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}^* = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\mathrm{full}}\widetilde{\boldsymbol{\beta}}\right\},\end{aligned} \tag{28}$$

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 $\boldsymbol{\beta}^\dagger = \boldsymbol{H}_{\mathrm{full}}^\top\boldsymbol{\theta}^\dagger$. By Lemma S8 from [27], we know that

$$\boldsymbol{\beta}^\dagger \in \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \mathcal{L}_n(\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathcal{F}_{\boldsymbol{\theta}^*}} \Omega_{\mathcal{G}}(\boldsymbol{\beta}). \tag{29}$$

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

ASSUMPTION 1 (Predictor scaling). *The predictors are scaled so that for any $1 \in [n], j \in [t]$, 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. *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 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 $\Lambda := \max_{i\in[n]} \|e^{\boldsymbol{\theta}^\dagger \boldsymbol{x}_i}\|_\infty \le C_1$.*

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 [24, 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 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., $\Delta\boldsymbol{\theta} \in \mathcal{M}_H := \{\boldsymbol{\theta} : \boldsymbol{\theta} = \sum_{g\in\mathcal{G}} \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_g\}$,*

$$(30)\quad \mathcal{E}_n(\Delta\boldsymbol{\theta}) \ge \frac{\kappa}{2}\|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\|^2 - A\cdot C^2\left(\frac{\log|\mathcal{G}|}{n} + \frac{m}{n}\right)\cdot \inf_{\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta} = \boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}'} \Phi_{\mathcal{G}}^2(\Delta\boldsymbol{\theta}'), \|\Delta\boldsymbol{\theta}\| \le R,$$

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 $\kappa = \kappa_{\boldsymbol{J}}^{Pois}$, whereas under (ii), the multinomial categorical response model, $\boldsymbol{P}_{\mathcal{V}} = I - |\boldsymbol{J}|^{-1}\mathbf{1}_{|\boldsymbol{J}|}\mathbf{1}_{|\boldsymbol{J}|}^\top$, and denote $\kappa = \kappa_{\boldsymbol{J}}^{Mult}$.

Restricted strong convexity condition is a well-understood condition in penalized regression. Effectively, this condition requires that in a neighborhood of the true parameter, the negative log-likelihood has sufficient curvature.

REMARK 1. *In Lemma S4 of [27], 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 $\mathcal{S} = \{g \in \mathcal{G}; \boldsymbol{\beta}_g^\dagger \ne 0\}$ and define $\Psi(\mathcal{S})^2 = \sum_{(\boldsymbol{k},j)\in\mathcal{S}} w_{\boldsymbol{k},j}^2$. 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 [24, 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 $\boldsymbol{\beta}^\dagger \in \mathbb{R}^{p\times|\boldsymbol{J}|}$. Thus, to simplify notation, let $\widehat{\boldsymbol{\beta}}_0$ denote the version of $\widehat{\boldsymbol{\beta}}$ where all effects of order higher than d have been set to zero (i.e., $\widehat{\boldsymbol{\beta}}_0 = [\widehat{\boldsymbol{\beta}}, 0] \in \mathbb{R}^{p\times|\boldsymbol{J}|}$. 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 $d^* = \{\|\boldsymbol{k}\|_0 ; (\boldsymbol{k}, j) \in \mathcal{S}\}$.

THEOREM 4. *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 1-4 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 $1 - e^{-B'(\xi-1)^2(m+\log|\mathcal{G}|)}$.

(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 $1 - e^{-B'(\xi-1)^2(m+\log|\mathcal{G}|)}$.

In Lemma 5 of [27], we show that, under certain regularity assumptions, $\kappa_{\boldsymbol{J}}^{Mult} \asymp \frac{1}{|\boldsymbol{J}|}\kappa_{\boldsymbol{J}}^{Pois}$ and $\kappa_{\boldsymbol{J}}^{Pois} = O(1)$. However, we cannot conclude that Poisson sampling scheme is better than multinomial sampling scheme, because the nature of the data generating models are different.

The result of Theorem 4 indicates that under the Poisson or multinomial sampling scheme, assuming $\kappa_{\boldsymbol{J}}^{Pois} = O(1)$ or $\kappa_{\boldsymbol{J}}^{Mult} = (\frac{1}{|\boldsymbol{J}|})$, 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 (15) 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 effect 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 effects to be zero.

The following corollary is a special case of Theorem 4 for multinomial categorical response model, letting $\mathcal{G} = \mathcal{G}_{\text{global}}$ or $\mathcal{G}_{\text{local}}$. Here, we replace the quantities from Theorem 4 with more explicit versions.

COROLLARY 1. *Under the conditions of Theorem 4, assuming the multinomial categorical response model, if tuning parameters are chosen in accordance with Theorem 4(ii) and $w_g = 1$ for all $g \in \mathcal{G}$, then*

1. *For global association learning, with probability as specified in Theorem 4,*

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

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 4,*

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

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 4, particularly regarding its validity under the assumption of random predictors. In section 9 of [24], the restricted strong convexity condition has been derived under a GLM setting (See Theorem 9.36 in [24]). 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 S7 of [27] and incorporate both the multinomial and Poisson categorical response settings into the following lemma.

LEMMA 5. *Under Assumptions 1–3 and equation (S31) from [27], and assuming that* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are independent and identically distributed with zero mean, we have the following result. For both multinomial and Poisson categorical response models with the reparameterized group lasso penalty* $\Phi_{\mathcal{G}}$*, the restricted strong convexity condition* (30) *in Assumption 4 holds with probability at least* $1 - c_1 e^{-c_2 n}$*. Furthermore,* $\kappa_{\boldsymbol{J}}^{Mult} \asymp \frac{1}{|\boldsymbol{J}|}\kappa_{\boldsymbol{J}}^{Pois}$ *and* $\kappa_{\boldsymbol{J}}^{Pois} = O(1)$.

The above lemma justifies condition (30) and the typical behavior of the curvature κ in Assumption 4, showing that both will hold with high probability under mild assumptions.

7. Numerical studies.

7.1. *Data generating models and competitors.* We present a series of simulations designed to evaluate the performance of the proposed methods and classical 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 S3 of [27].

Parameter setup and simulation. For $N_{\text{rep}} = 100$ independent replications, we simulate data based on the multivariate multinomial logistic regression framework, specifying $d = 4$, $q = 4$ categories, and dimensions $J_1 = J_2 = J_3 = 2$, $J_4 = 3$. With $n \in \{100, 300, 500, 1000, 2000\}$ training samples, each observation $\boldsymbol{x}_i$ is drawn from a multivariate normal distribution $N_p(\mathbf{0}, \Sigma_X)$, where the covariance entries $\Sigma_{jk} = 0.5^{|j-k|}$ are defined for all pairs $(j,k) \in [p]^2$. Given a coefficient matrix $\boldsymbol{\beta}^* \in \mathbb{R}^{\sum_{s=0}^{d} L_s \times p}$, the probability vector is given by

$$\mathrm{vec}_{\boldsymbol{J}}\left(\boldsymbol{\pi}^{\boldsymbol{J}}_{j_1,j_2,j_3}(\boldsymbol{x})\right) = \frac{\exp(\boldsymbol{H}\boldsymbol{\beta}^*\boldsymbol{x})}{\langle \mathbf{1}_{|\boldsymbol{J}|}, \exp(\boldsymbol{H}\boldsymbol{\beta}^*\boldsymbol{x})\rangle},$$

from which we generate the response vectors $\boldsymbol{y}_i \in \mathbb{R}^p$ based on

$$\boldsymbol{y}_i | \boldsymbol{x}_i \sim \text{Multinomial}\left(n_i, \mathrm{vec}_{\boldsymbol{J}}\left(\boldsymbol{\pi}^{\boldsymbol{J}}_{j_1,j_2,j_3,j_4}(\boldsymbol{x}_i)\right)\right), \quad n_i = 1. \tag{31}$$

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. These correspond to the three interpretable models—mutual independence, joint independence, and conditional independence, respectively—as presented in Example 2.

Candidate estimators. 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.

Tuning criteria. We employ a train-validation split in order to select tuning parameters in our simulation study. Specifically, we select the candidate tuning parameters that minimize cross-entropy loss on the validation set.

7.2. *Results.* The estimators' performances, evaluated based on Hellinger distance and (joint) misclassification rate on a test set, is displayed in Figures 1 and 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 effects. 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 should do not suggest that estimators assuming the multinomial data generating model are uniformly preferable to those assuming a Poisson data generating model. In this case, the estimators minimizing a multinomial negative log-likelihood assume a correctly specified model, and thus, as the sample size increases, tend to outperform their Poisson counterparts.

7.3. *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

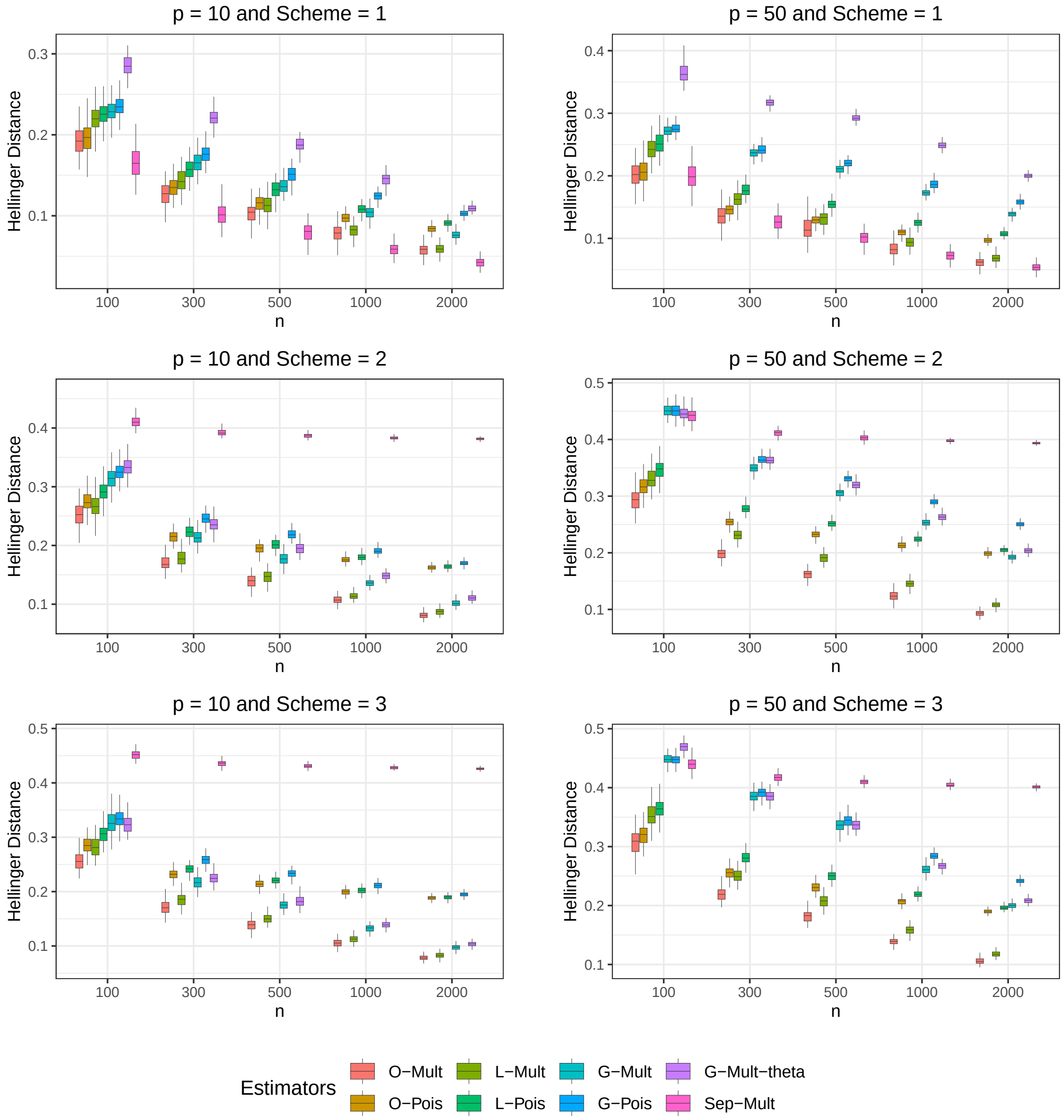

FIG 1. *Hellinger distances for the competing estimators with* $p \in \{10, 50\}$ *and Scheme* $\in \{1, 2, 3\}$ *as* n *varies.*

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 5. 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.

8. Discussion.

This article introduces an alternative approach to multivariate categorical response regression by introducing a subspace decomposition. Our proposed decomposition

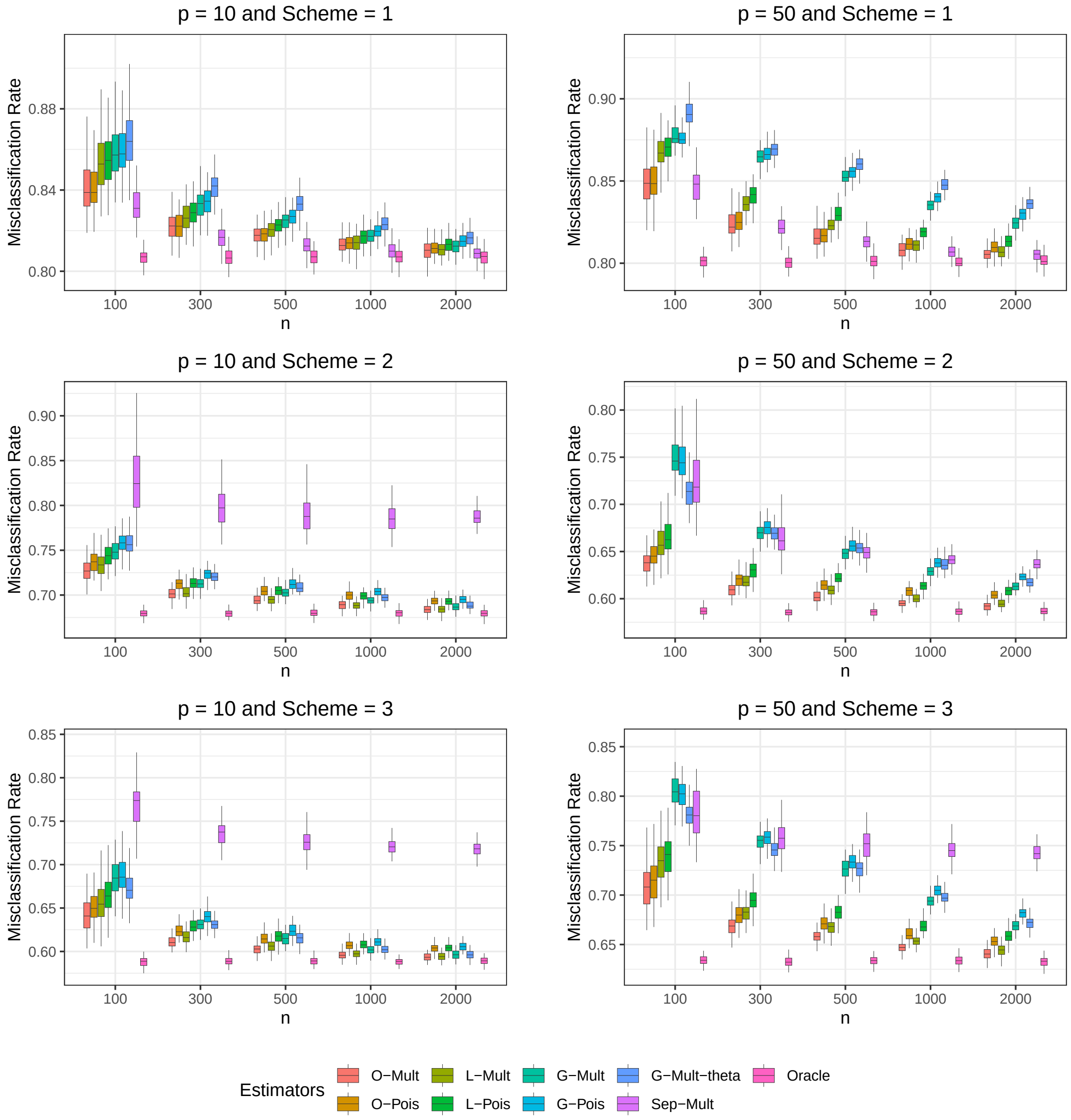

FIG 2. *Misclassification rates for the competing estimators with* $p \in \{10, 50\}$ *and Scheme* $\in \{1, 2, 3\}$ *as* n *varies.*

allows practitioners to use standard regularization techniques to select the order of effects, and bypasses the issue of dependence on choice of identifability constraints. There are two key directions for future research.

8.1. *More computationally efficient approaches to hierarchically-structured effect selection.* The use of the overlapping group lasso penalty Ω^H to select effects adhering to a hierarchy is especially appealing in practice, but more computationally intensive than the estimator excluding hierarchical constraints. In the future, it is important to consider alternative approaches to regularization that may be less computationally intensive, but encourage the desired hierarchy. One such approach may be to utilize the latent overlapping group lasso penalty [16], which allows the optimization problem to be separable across the (latent) parameters being penalized. This can allow for more efficient computational algorithms and schemes to be developed. The estimator based on latent overlapping group lasso penalty is distinct from that based on the overlapping group lasso penalty in the sense that their solution paths are fundamentally distinct, but both can be used to enforce hierarchical constraints. Consequently, the theoretical properties of the estimator based on the latent overlapping group lasso penalty are not immediate from the results we derived in Section 6, so this direction is nontrivial.

Another approach is to use a separable (non-overlapping) approximation to the overlapping group lasso penalty. Specifically, [20] recently proposed a separable relaxation of the overlapping group lasso penalty, and showed that in terms of squared estimator error, the estimator using their relaxation is statistically equivalent to that using the overlapping group lasso penalty. Notably, because the relaxation is separable, the corresponding estimator can be computed much more efficiently—roughly at the same cost as estimators using nonoverlapping group lasso penalization schemes.

8.2. *Other representations in predictors.* Recall that in our model (7) and (6), $\boldsymbol{\theta x} = \boldsymbol{H \beta x}$ is linear in $\boldsymbol{x}$. The subspace decomposition model can be extended to accommodate scenarios where the relationship with $\boldsymbol{x}$ is not necessarily linear, i.e.,

$$\theta(\boldsymbol{x}) = \boldsymbol{H}\beta(\boldsymbol{x}),$$

where $\theta : \mathbb{R}^p \to \mathbb{R}^{|\boldsymbol{J}|}$ and $\beta : \mathbb{R}^p \to \mathbb{R}^{\sum_{s=0}^{d} L_s}$. Here, β can be associated with both parametric models, such as polynomial regression, and non-parametric models, including splines, kernel-based models, additive models, and deep learning architectures.

8.3. *Application to the analysis of large contingency tables.* Finally, a direction not explored in this article is the use of our estimator for fitting traditional log-linear models for contingency tables. The traditional log-linear model is a special case of our model with the predictor consisting of the intercept only. Effect selection in standard log-linear models has been studied in the past [e.g, see 15], but in the asymptotic regime with $n \to \infty$ and all other model dimensions fixed. Thus, it is of particular interest to study whether our finite sample error bounds can be applied, or even refined, in this context.

Acknowledgments. A. J. Molstad's contributions were supported in part by a grant from the National Science Foundation (DMS 2415067). H. Zhao would like to thank Jinwen Fu, Robert Oliver Vandenberg, Yinzhao Wang, Yuyuan Wang, and Xiangyu Zhang for useful discussions.

REFERENCES

[1] AGRESTI, A. (2002). *Categorical Data Analysis*, 2 ed. John Wiley & Sons, New York.

[2] BERGSMA, W. P. and RUDAS, T. (2002). Marginal models for categorical data. *The Annals of Statistics* **30** 140–159.

[3] CHRISTENSEN, R. (1997). *Log-linear models and logistic regression*. Springer Science & Business Media.

[4] GLONEK, G. F. (1996). A class of regression models for multivariate categorical responses. *Biometrika* **83** 15–28.

[5] GLONEK, G. F. and MCCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 533–546.

[6] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

[7] HERRERA, F., CHARTE, F., RIVERA, A. J., DEL JESUS, M. J., HERRERA, F., CHARTE, F., RIVERA, A. J. and DEL JESUS, M. J. (2016). *Multilabel classification*. Springer.

[8] JENATTON, R., MAIRAL, J., OBOZINSKI, G. and BACH, F. (2011). Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research* **12** 2297–2334.

[9] LANG, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *The Annals of Statistics* **24** 726–752.

[10] MAI, Q., YANG, Y. and ZOU, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica* **29** 97–111.

[11] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2 ed. Chapman and Hall, London.

[12] Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine* **18** 2237–2255.

[13] Molstad, A. J. and Rothman, A. J. (2023). A likelihood-based approach for multivariate categorical response regression in high dimensions. *Journal of the American Statistical Association* **118** 1402–1414.

[14] Molstad, A. J. and Zhang, X. (2022). Conditional probability tensor decompositions for multivariate categorical response regression. *arXiv preprint arXiv:2206.10676*.

[15] Nardi, Y. and Rinaldo, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974.

[16] Obozinski, G., Jacob, L. and Vert, J.-P. (2011). Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*.

[17] Palmgren, J. (1989). Regression models for bivariate binary responses Technical Report No. 101, Department of Biostatistics, University of Washington, Seattle.

[18] Parikh, N., Boyd, S. et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization* **1** 127–239.

[19] Qaqish, B. F. and Ivanova, A. (2006). Multivariate logistic models. *Biometrika* **93** 1011–1017.

[20] Qi, M. and Li, T. (2024). The Non-Overlapping Statistical Approximation to Overlapping Group Lasso. *Journal of Machine Learning Research* **25** 1–70.

[21] Read, J., Pfahringer, B., Holmes, G. and Frank, E. (2021). Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research* **70** 683–718.

[22] Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization* **2**.

[23] Vincent, M. and Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis* **71** 771–786.

[24] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.

[25] Yan, X. and Bien, J. (2017). Hierarchical Sparse Modeling: A Choice of Two Group Lasso Formulations. *Statistical Science* **32** 531–560.

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

[27] Zhao, H., Molstad, A. J. and Rothman, A. J. (2024). Supplementary Materials to "Subspace decompositions for association structure learning in multivariate categorical response regression".

[28] Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* **37** 3468–3497.

SUPPLEMENTARY MATERIALS TO "SUBSPACE DECOMPOSITIONS FOR ASSOCIATION STRUCTURE LEARNING IN MULTIVARIATE CATEGORICAL RESPONSE REGRESSION"

By Hongru Zhao[†], Aaron J. Molstad[†,⋆], and Adam J. Rothman[†]
School of Statistics, University of Minnesota, Minneapolis, MN[†]
Department of Statistics, University of Florida, Gainesville, FL[⋆]

S1. 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 $\mathbb{R}$ with a basis consisting of elements of the form $\otimes_t(v_{j_1}^1, v_{j_2}^2, \cdots, v_{j_q}^q)$ for $j_l = 1, 2, \ldots, J_l$ and $l = 1, 2, \ldots, q$.

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 $i \neq j$.

Here, 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 $\otimes_t(v_{j_1}^1, v_{j_2}^2, \cdots, v_{j_q}^q)$ the same as $\boldsymbol{E}_{j_1,\cdots,j_q}^{\boldsymbol{J}}$, where $\boldsymbol{E}_{j_1,\cdots,j_q}^{\boldsymbol{J}}$ 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 $\boldsymbol{E}_{\boldsymbol{j}}^{\boldsymbol{J}} \sim \otimes_t(v_{j_1}^1, v_{j_2}^2, \cdots, v_{j_q}^q)$ 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 inner product space.

S2. Details about computation.

S2.1. *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 $n_\lambda = 100$, $\text{ratio}_\lambda \in (0,1)$, $d \in \{1, 2, \cdots, q\}$, $\gamma \in (0,1)$. 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 (S27) and (S28) in Lemma S6.

Let λ_1 be a value such that $\widehat{\boldsymbol{\theta}}_{\lambda_1} = 0$. Let $\lambda_{i+1} = \text{ratio}_\lambda \lambda_i, 1 \le i \le n_\lambda - 1$. 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 S7.

S2.2. *Algorithm Formulation.* The accelerated backtracking proximal gradient descent algorithm is presented in Algorithm 1.

Algorithm 1: Accelerated Backtracking Proximal Gradient Descent Algorithm

```
1  for $i \in \{2, 3, \ldots, n_\lambda\}$ do
2      Set $k = 0$, $\widehat{\boldsymbol{\beta}}^0_\eta = \widehat{\boldsymbol{\beta}}^{-1}_\eta = \widehat{\boldsymbol{\beta}}_{\lambda_{i-1}}$ ;
3      $\boldsymbol{z}^{k+1} = \widehat{\boldsymbol{\beta}}^k_\eta + \frac{k}{k+3}\left(\widehat{\boldsymbol{\beta}}^k_\eta - \widehat{\boldsymbol{\beta}}^{k-1}_\eta\right)$ ;
4      $\widehat{\boldsymbol{\beta}}^{k+1}_\eta \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{z}^{k+1} - \eta\nabla\mathcal{L}_n(\boldsymbol{z}^{k+1})\right)\right\|^2 + \lambda_i\eta \cdot \Omega(\boldsymbol{\beta})$ ;
5      while
         $\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{k+1}_\eta) > \mathcal{L}_n(\boldsymbol{z}^{k+1}) + \left\langle \nabla\mathcal{L}_n(\boldsymbol{z}^{k+1}), \widehat{\boldsymbol{\beta}}^{k+1}_\eta - \boldsymbol{z}^{k+1} \right\rangle + \frac{1}{2\eta}\left\|\widehat{\boldsymbol{\beta}}^{k+1}_\eta - \boldsymbol{z}^{k+1}\right\|^2$
         do
6          Shrink $\eta \leftarrow \gamma\eta$ ;
7      if $\widehat{\boldsymbol{\beta}}^{k+1}_\eta$ has not converged then
8          Set $k \leftarrow k + 1$ and return to step 3 ;
9      $\widehat{\boldsymbol{\beta}}_{\lambda_i} \leftarrow \widehat{\boldsymbol{\beta}}^{k+1}_\eta$,    $\widehat{\boldsymbol{\theta}}_{\lambda_i} \leftarrow \boldsymbol{H}\widehat{\boldsymbol{\beta}}_{\lambda_i}$ ;
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 (17), we replace line 4 in Algorithm 1 with the following line.

Algorithm 2: Accelerated Backtracking Proximal Gradient Descent Algorithm (Group Lasso)

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

In addressing the proximal problem for the overlapping group lasso penalty, as defined in (20), we adopt the methodology outlined in Jenatton et al. (2011), utilizing block coordinate descent. We modify line 4 of Algorithm 1 accordingly.

Algorithm 3: Accelerated Backtracking Proximal Gradient Descent Algorithm (Overlapping Group Lasso)

```
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          $\boldsymbol{\xi}^{d(g)} \leftarrow \boldsymbol{\Pi}_{\lambda w_g}(\boldsymbol{\zeta}^{j+1})$ ;
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, $\boldsymbol{\Pi}_\rho(\cdot)$ 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 S6. Similarly, the value of λ_1 in these algorithms is suggested by Lemma S7.

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.

S3. Detailed setup for Simulation. **Parameter setup and 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, 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, 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, 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

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

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

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

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

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

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

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

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

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta}) + \lambda \Omega_{\mathcal{G}_0}(\boldsymbol{\beta}).$$

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

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_n^{\mathrm{Pois}}(\boldsymbol{\beta}) + \lambda \Omega_{\mathcal{G}_0}(\boldsymbol{\beta}).$$

   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

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

   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
$$\widehat{\boldsymbol{\eta}}^{(m)} = \arg\min_{\boldsymbol{\eta}^{(m)}} \sum_{i=1}^{n} \frac{1}{n}\Big\{ - \Big\langle \boldsymbol{y}_i^{J_m}, \boldsymbol{\eta}^{(m)}\boldsymbol{x}_i \Big\rangle + n_i \log\Big(\Big\langle \mathbf{1}_{J_m}, e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}_i} \Big\rangle \Big) \Big\} + \lambda^{(m)} \sum_{j=1}^{J_m} \Big\| \boldsymbol{\eta}_{j,:}^{(m)} \Big\|,$$
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 $\widehat{\boldsymbol{\pi}}_{j_1,\cdots,j_q}^{\boldsymbol{J}}(\boldsymbol{x}) = \prod_{m=1}^{q} [e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}}]_{j_m} \Big(\sum_{\boldsymbol{j} \in \boldsymbol{J}} \prod_{m=1}^{q} [e^{\boldsymbol{\eta}^{(m)}\boldsymbol{x}}]_{j_m} \Big)^{-1}$, 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
$$\operatorname{vec}_{\boldsymbol{J}}\Big(\boldsymbol{\pi}_{j_1,\cdots,j_q}^{\boldsymbol{J}}(\boldsymbol{x}) \Big) = \frac{e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}}}{\Big\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}} \Big\rangle}.$$

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 $n_{\text{val}} = 1000$ 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
$$\arg\min_{\lambda \in \Lambda} \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} \ell_{\text{Mult}}(\widehat{\boldsymbol{\theta}}_{\lambda}\boldsymbol{x}_i, \boldsymbol{y}_i).$$

Model evaluation: The Hellinger distance and misclassification rate, considered in Section 7, 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 $H(\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{J}}(\boldsymbol{x}), \widehat{\boldsymbol{\pi}}_{\boldsymbol{j}}^{\boldsymbol{J}}(\boldsymbol{x})) = \frac{1}{\sqrt{2}} \Big\| \sqrt{\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{J}}(\boldsymbol{x})} - \sqrt{\widehat{\boldsymbol{\pi}}_{\boldsymbol{j}}^{\boldsymbol{J}}(\boldsymbol{x})} \Big\|$. Generate samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N_{\text{test}}} \sim N_p(\mathbf{0}, \Sigma_X)$, 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 (31). The misclassification rate is given by
$$\frac{\sum_{i=1}^{N_{\text{test}}} I\Big(\arg\max(\boldsymbol{y}_i) \neq \arg\max\big(\operatorname{vec}_{\boldsymbol{J}}(\widehat{\boldsymbol{\pi}}^{\boldsymbol{J}}(\boldsymbol{x}_i)) \big) \Big)}{N_{\text{test}}},$$
where $\arg\min \boldsymbol{x} = \arg\min_j x_j, \boldsymbol{x} = (x_1, \cdots, x_{|\boldsymbol{J}|})$.

S4. Proofs of the Isomorphism and Subspace Decomposition Lemmas. In this section, we provide the proofs for Lemma 1 and Lemma 2. We will use the following simple result from linear algebra.

LEMMA S1. *For subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$, if (i) $\boldsymbol{U} \subset \boldsymbol{V}$ and (ii) $\dim(\boldsymbol{U}) = \dim(\boldsymbol{V}) < \infty$, then $\boldsymbol{U} = \boldsymbol{V}$.*

S4.1. *Proof of Lemma 1.*

PROOF OF LEMMA 1. Let $\boldsymbol{v}_i = (v_i(1), \cdots, v_i(J_i))^\top \in \mathbb{R}^{J_i}$ for any $1 \le i \le q$. 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 (9). Recall that $\mathcal{R}(\boldsymbol{U})$ denotes the column space of the matrix $\boldsymbol{U}$. For any vectors $\boldsymbol{v}_i \in \mathcal{R}(\boldsymbol{V}_i), 1 \le i \le q$, there exists $\boldsymbol{\alpha}_i$ such that $\boldsymbol{v}_i = \boldsymbol{V}_i \boldsymbol{\alpha}_i$. Combining identity (9), 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 S1, 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 $\dim T_{\boldsymbol{J}}(\mathcal{R}(\boldsymbol{V}_q) \otimes \cdots \otimes \mathcal{R}(\boldsymbol{V}_1)) = \dim \mathcal{R}(\boldsymbol{V}_q \otimes \cdots \otimes \boldsymbol{V}_1)$. □

S4.2. *Proof of Lemma 2.*

PROOF OF LEMMA 2. Let $\mathcal{K}' = \cup_{s=0}^q \mathcal{K}_s$. Let $\boldsymbol{P}_{\boldsymbol{k}} = \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{H}_{\boldsymbol{k}}^\top$ 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 $\boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{H_k}$ 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

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

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

$$\text{(S1)} \qquad \boldsymbol{H}_{\boldsymbol{k}}^\top \boldsymbol{H_{k'}} = (\boldsymbol{A}^\top \boldsymbol{A}') \otimes (\frac{1}{\sqrt{J_l}} \boldsymbol{U}_{J_l}^\top \mathbf{1}_{J_l}) \otimes (\boldsymbol{B}^\top \boldsymbol{B}') = (\boldsymbol{A}^\top \boldsymbol{A}') \otimes \mathbf{0} \otimes (\boldsymbol{B}^\top \boldsymbol{B}') = \mathbf{0},$$

which completes the proof of statement 2.

Furthermore, we also know that the columns of $\boldsymbol{H_k}$ 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

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

Combined with (S1), we further obtain that

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

In summary, the columns of $\boldsymbol{H_k}$ for any $\boldsymbol{k} \in \mathcal{K}'$ form an orthonormal basis of $\mathbb{R}^{|\boldsymbol{J}|}$, so by Lemma S1 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) $\boldsymbol{y} = I\boldsymbol{y} = \sum_{\boldsymbol{k}\in\mathcal{K}'} \boldsymbol{P_k}\boldsymbol{y}$,
(ii) $\boldsymbol{P_k}\boldsymbol{y} \in \mathcal{R}(\boldsymbol{H_k})$ for any $\boldsymbol{k} \in \mathcal{K}'$,
(iii) $\mathcal{R}(\boldsymbol{H_k}) \cap \Big(\sum_{\boldsymbol{k}'\in\mathcal{K}', \boldsymbol{k}'\neq\boldsymbol{k}} \mathcal{R}(\boldsymbol{H_{k'}})\Big) = \{\mathbf{0}\}$ for any $\boldsymbol{k} \in \mathcal{K}'$, and
(iv) $\mathcal{R}(\boldsymbol{H_k}) \perp \mathcal{R}(\boldsymbol{H_{k'}})$ 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 $\{\mathcal{R}(\boldsymbol{H_k})\}_{\boldsymbol{k}\in\mathcal{K}'}$ 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. □

S5. Proof of Theorem 3 . In this section, we first prove the inverse vectorization identities (S2) and (S3) for tensor reshaping, which plays a crucial role in the proof of Theorem 3.

LEMMA S2. *(Inverse Vectorization Identity for Nonentangled Tensor) Let $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ be a disjoint partition of $[q]$. For any $\mathcal{I} = \{i_1, \cdots, i_s\} \subset [q]$ such that $1 \leq i_1 < \cdots < i_s \leq q$, define $\boldsymbol{j}_{\mathcal{I}} = (j_{i_1}, \cdots j_{i_s})$ 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

$$\text{(S2)} \qquad \left(\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{v})\right)_{\boldsymbol{j}} = \prod_{l=1}^{m} \left(\mathrm{vec}_{\boldsymbol{J}^{\mathcal{I}^l}}^{-1}(\boldsymbol{v}_{\mathcal{I}_l})\right)_{\boldsymbol{j}_{\mathcal{I}_l}},$$

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

PROOF OF LEMMA S2. 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

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

□

LEMMA S3. *(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* $\mathrm{vec}_{\boldsymbol{J_k}}^{-1}$ *such that*

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

then the following inverse vectorization identity holds,

(S3)

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

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 S3. Recall the definition of $\boldsymbol{H}$ in (16) and $\boldsymbol{H_k}$ in (11). Combined with the linearity of $\mathrm{vec}_{\boldsymbol{J}}^{-1}$, we know that

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

To show the inverse vectorization identity (S3), it suffices to show that

$$\text{(S4)} \quad \left(\mathrm{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H_k\beta_k x})\right)_{j_1,\cdots,j_q} = \frac{1}{\sqrt{|\boldsymbol{J}|}}\left(\mathrm{vec}_{\boldsymbol{J_k}}^{-1}\left(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta_k x}\right)\right)_{j_{i_1},\cdots,j_{i_s}}, (j_1, \cdots, j_q) \in \boldsymbol{J},$$

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 (11), we know that

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

Thus, we obtain that

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

where the last equation holds due to identity (S2) in Lemma S2.

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 (S4), thereby completing the proof of Lemma S3 as well. □

Utilizing Lemma S3, we proceed to prove Theorem 3.

PROOF OF THEOREM 3. Consider the representation

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

Applying the inverse vectorization identity (S3) in Lemma S3, we know that there exists some $B(\boldsymbol{x})$ and $B'(\boldsymbol{x})$ not depending on $\boldsymbol{j}$, such that

$$\begin{aligned}\pi_{\boldsymbol{j}}(\boldsymbol{x}) &= \exp\{B(\boldsymbol{x})\}\left[\exp\{\operatorname{vec}_{\boldsymbol{J}}^{-1}(\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x})\}\right]_{j_1,\cdots,j_q}\\ &= \exp\{B'(\boldsymbol{x})\}\underbrace{\exp\left[\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{1\le s\le q}\sum_{\boldsymbol{k}\in\mathcal{K}_s}\{\operatorname{vec}_{\boldsymbol{J}_{\boldsymbol{k}}}^{-1}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\}_{j_{i_1},\cdots,j_{i_s}}\right]}_{=:\xi_{\boldsymbol{j}}(\boldsymbol{x})}.\end{aligned} \tag{S5}$$

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

For any $\mathcal{I} = \{i_1,\cdots,i_s\}\subset[q]$ such that $1\le i_1<\cdots<i_s\le q$, define $\boldsymbol{J}^{\mathcal{I}} = [J_{i_1}]\times\cdots\times[J_{i_s}]$, 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 (12), it suffices to show that

$$\log\xi_{\boldsymbol{j}}(\boldsymbol{x}) = \sum_{l=1}^{m}\log\xi_{\boldsymbol{j}_{\mathcal{I}_l},+}(\boldsymbol{x}) + C(\boldsymbol{x}),\ \text{for all } \boldsymbol{j}\in\boldsymbol{J}, \tag{S6}$$

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

$$\boldsymbol{\theta} = \sum_{\boldsymbol{k}\in\mathcal{S}_{\text{joint}}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}} = \boldsymbol{H}_{\{0\}}\boldsymbol{\beta}_{\{0\}} + \sum_{l=1}^{m}\sum_{\boldsymbol{k}\subset\mathcal{I}_l,\|\boldsymbol{k}\|_0\ge1}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}.$$

Similar to (S5), we obtain that

$$\xi_{\boldsymbol{j}}(\boldsymbol{x}) = \prod_{l=1}^{m}\exp\Bigg(\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\boldsymbol{k}\subset\mathcal{I}_l,\|\boldsymbol{k}\|_0\ge1}\left(\operatorname{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\right)_{j_{i_1},\cdots,j_{i_s}}\Bigg).$$

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

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

where

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

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

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

which implies (S6), by setting $C(\boldsymbol{x}) = -\sum_{l=1}^{m} C_l(\boldsymbol{x})$. Thus, we complete the proof of (12).
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 (S5), we obtain that

$$\begin{aligned}\xi_{\boldsymbol{j}}(\boldsymbol{x}) = &\prod_{l=1}^{m-1}\exp\Big(\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\substack{|\boldsymbol{k}\cap\mathcal{I}_l|>0\\ \boldsymbol{k}\subset\mathcal{I}_l\cup\mathcal{I}_m}}\Big(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\Big)_{j_{i_1},\cdots,j_{i_s}}\Big)\\ &\cdot\exp\Big(\frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\boldsymbol{k}\subset\mathcal{I}_m,\|\boldsymbol{k}\|_0\geq 1}\Big(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\Big)_{j_{i_1},\cdots,j_{i_s}}\Big).\end{aligned}\tag{S7}$$

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

$$\log \xi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},+}(\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{J}|}}\sum_{\substack{|\boldsymbol{k}\cap\mathcal{I}_l|>0\\ \boldsymbol{k}\subset\mathcal{I}_l\cup\mathcal{I}_m}}\Big(\mathrm{vec}_{\boldsymbol{J}_{\boldsymbol{k}}^{-1}}(\mathcal{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\boldsymbol{x})\Big)_{j_{i_1},\cdots,j_{i_s}} + C_{l,\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}),\tag{S8}$$

where

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

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

$$\log \pi_{\boldsymbol{j}_{\mathcal{I}_1},\cdots,\boldsymbol{j}_{\mathcal{I}_{m-1}}|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) = \log \pi_{\boldsymbol{j}}(\boldsymbol{x}) - \log \pi_{\boldsymbol{j}_{\mathcal{I}_m}+}(\boldsymbol{x}) = \log \xi_{\boldsymbol{j}}(\boldsymbol{x}) + C_{\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}),\tag{S9}$$

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

$$\begin{aligned}\log \pi_{\boldsymbol{j}_{\mathcal{I}_l}+|\boldsymbol{j}_{\mathcal{I}_m}}(\boldsymbol{x}) &= \log \pi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m}+}(\boldsymbol{x}) - \log \pi_{\boldsymbol{j}_{\mathcal{I}_m}+}(\boldsymbol{x}) \\ &= \log \xi_{\boldsymbol{j}_{\mathcal{I}_l\cup\mathcal{I}_m},+}(\boldsymbol{x}) + C'_{\mathcal{I}_l}(\boldsymbol{x}),\end{aligned} \tag{S10}$$

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 (S7), (S8), (S9) and (S10), 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 (13). □

S6. Proof of Theorem 4. To complete the proof of Theorem 4, we will show the following error bound for the group lasso without assuming $w_g \equiv 1$.

THEOREM S1. *Let B, B_1, B_2 and B' be some positive absolute constants. Let $\lambda_\delta = BC\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\big), \delta \ge 0$ for multinomial categorical response model, and let $\lambda_\delta = BC\sqrt{\Lambda}\big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\big), 0 \le \delta \le B_2$ for Poisson categorical response model.*

Assume $d^* \le d$. Under Assumptions 1-4, if $\frac{m}{n} + \frac{\log|\mathcal{G}|}{n} \le B_1 \cdot \min\big(1, \frac{\kappa}{AC^2\Psi^2(\mathcal{S})}\big)$ and $\lambda_\delta \le \frac{R\cdot\kappa}{6\Psi(\mathcal{S})}$, then

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

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

Consequently, Theorem 4 can be viewed as a specific case of Theorem S1.

PROOF OF THEOREM 4. Applying Theorem S1 and assuming $w_g \equiv 1$, we complete the proof of Theorem 4. □

S7. 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 S1. Consider the multivariate GLM model: conditionally on $\boldsymbol{x}_i$, each response $\boldsymbol{y}_i$ is i.i.d. according to a conditional distribution of the following form:

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

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 $M > 0$,

$$\gamma_M = \inf_{\substack{\|\boldsymbol{\theta x}\|\le M \\ \boldsymbol{\theta}\in\mathcal{M}_H}} \inf_{\substack{\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}\neq 0 \\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \frac{\big\langle \nabla^2\psi(\boldsymbol{\theta x})\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}\big\rangle}{\|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}\|^2} > 0. \tag{S11}$$

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

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

THEOREM S2. *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*

$$\mathbb{E}\,\|\Delta\boldsymbol{\theta x}_i\|^2 \ge \alpha \text{ and } \mathbb{E}\,\|\Delta\boldsymbol{\theta x}_i\|^4 \le \beta, \tag{S12}$$

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 $\Phi(\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\theta}) \le \Phi(\boldsymbol{\theta})$, for any $\boldsymbol{\theta} \in \mathcal{M}_H$. Under Assumption S1, for any $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$, we have

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) \ge \frac{\kappa}{2}\|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\|^2 - c_0 \cdot \mu_n^2(\Phi, \mathcal{M}_H)\cdot \Phi^2(\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}), \|\Delta\boldsymbol{\theta}\| \le R, R > 0, \tag{S13}$$

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

$$\begin{aligned} \mu_n(\Phi, \mathcal{M}_H) =& \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i 1\le i\le n} \sup_{\substack{\Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\le 1 \\ \Delta\boldsymbol{\theta}\in\mathcal{M}_H}} \left(\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\varepsilon}_i, \boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i\rangle\right) \\ =& \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i 1\le i\le n} \Phi^*\Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{P}_{\mathcal{M}_H}\boldsymbol{P}_{\mathcal{V}}\big(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^\top\big)\Big), \end{aligned} \tag{S14}$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i|\boldsymbol{J}|})^\top \in \{-1, 1\}^{|\boldsymbol{J}|}, 1 \le i \le n$, ε_{ij} is a sequence of independent doubly indexed Rademacher variables, Φ^* denotes the duel 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 independent of the dimension, sample size, and regularizer Φ. Furthermore, for prespecified large value M, we know that $\kappa, c_0 \propto \gamma_M$, where γ_M is defined in (S11).

REMARK S1. The definition (S14) 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 $1 \le i \le n$ are multivariate responses. $\mu_n(\Phi, \mathcal{M}_H)$ is simply the Rademacher complexity of the class of linear operators $\boldsymbol{x} \mapsto \boldsymbol{\theta x}$ as $\boldsymbol{\theta} \in \mathcal{M}_H$ ranges over the unit ball of norm Φ.

LEMMA S4. *Under Assumptions 1-4, for both multinomial and Poisson categorical response models, there exists absolute constant A such that*

$$\mu_n(\Phi_{\mathcal{G}}, \mathcal{M}_H) = \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i, 1\le i\le n}\Omega^*_{\mathcal{G}}\Big(\frac{1}{n}\boldsymbol{H}^\top\sum_{i=1}^n \boldsymbol{P}_{\mathcal{V}}\big(\boldsymbol{\varepsilon}_i\boldsymbol{x}_i^\top\big)\Big) \le A\cdot C\Big(\sqrt{\frac{\log|\mathcal{G}|}{n}} + \sqrt{\frac{m}{n}}\Big), \tag{S15}$$

where $\Omega_{\mathcal{G}}^*$ denotes the duel 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.*

S7.1. *Proof of Theorem S2* . In this subsection, we provide the proofs for Theorem S2, Lemma S4 and Lemma 5.

PROOF OF THEOREM S2. 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

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

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 S1, to demonstrate that (S13) holds for any $\Delta\boldsymbol{\theta} \in \mathcal{M}_H$, it suffices to show that

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) = \mathcal{E}_n(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})) \geq \frac{\kappa}{2} \|\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\|^2 - c_0 \cdot \mu_n^2 \cdot \Phi^2(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})), \|\Delta\boldsymbol{\theta}\| \leq 1. \tag{S16}$$

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

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

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

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

where T is the second truncation parameter to be chosen. Let $\gamma := \gamma_{T+2K}$. By Assumption S1, we know that $\gamma > 0$.

Thus, we obtain that

$$\mathcal{E}_n(\Delta\boldsymbol{\theta}) \geq \frac{\gamma}{n}\sum_{i=1}^n \varphi_\tau(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\Big\|\boldsymbol{\theta}^\dagger\boldsymbol{x}_i\Big\| \leq T]. \tag{S17}$$

Applying (S17) and inequality $xy \leq \frac{x^2+y^2}{2}$, to show (S13), 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

$$\frac{1}{n}\sum_{i=1}^n \varphi_\tau(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\Big\|\boldsymbol{\theta}^\dagger\boldsymbol{x}_i\Big\| \leq T] \geq c_3\delta^2 - c_4\mu_n\Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta}))\delta. \tag{S18}$$

If the lower bound in (S18) holds, then inequality (S13) 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 (S18) holds for $\|\Delta\boldsymbol{\theta}\| = \delta = 1$. Define a new truncation function

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

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

$$\text{(S19)} \qquad \frac{1}{n}\sum_{i=1}^{n} \widetilde{\varphi_\tau}(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\right\| \le T] \ge c_3 - c_4 \cdot \mu_n \cdot \Phi(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})).$$

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

(S20)

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

Suppose that we can prove that

$$\text{(S21)} \qquad \mathbb{E}\left[\widetilde{\varphi_\tau}(\boldsymbol{P}_{\mathcal{V}}(\Delta\boldsymbol{\theta})\boldsymbol{x}_i) I[\left\|\boldsymbol{\theta}^\dagger \boldsymbol{x}_i\right\| \le T]\right] \ge \frac{3}{4}\alpha,$$

and

$$\text{(S22)} \qquad \mathbb{P}\left[Z_n(r) > \frac{\alpha}{2} + c_4 r \mu_n\right] \le \exp\left(-c_2 \frac{n r^2 \mu_n^2}{\sigma^2} - c_2 n\right).$$

Then with $c_3 = \alpha/4$, we obtain that for any $\|\Delta\boldsymbol{\theta}\| = 1$ and $\Phi(\Delta\boldsymbol{\theta}) \le r$, the bound (S19) hold. Similar to the argument in the proof of Theorem 9.34 in Wainwright (2019), we obtain that (S13) holds with probability at least $1 - c_1 e^{-c_2 n}$.

Now, we turn to the proof of inequalities (S21) and (S22). The proof of (S21) closely follows the argument for the expectation bound (9.99a) presented in Wainwright (2019); therefore, we omit its detailed proof here. However, for (S22), 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

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

With this modification of the proof of (9.99b) in Wainwright (2019), we complete the proof of (S22). In conclusion, we complete the proof of (S13), and obtain the expression (S14).

Hence, the proof is concluded. □

S7.2. *Proof of Lemma S4.*

PROOF OF LEMMA S4. 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 duel norms of $\Omega_{\mathcal{G}}$ and $\Phi_{\mathcal{G}}$, respectively. Because

(S23)

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

we know that

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

By the definition of dual norm $\Phi_{\mathcal{G}}^*$ and Lemma S12, we obtain

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

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,

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

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

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

Thus,

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

Notice that for Rademacher variable ε,

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

Thus,

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

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 $u > 0$,

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

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

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

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

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

□

S7.3. *Proof of Lemma 5.*

PROOF OF LEMMA 5. By applying Lemma S4, Lemma S10 and Theorem S2, we complete the proof of Lemma 5. □

S8. Auxiliary Lemmas.

S8.1. *Subspace Lipschitz Constant.* We first establish the subspace Lipschitz constant $\Psi(\mathcal{S})$.

LEMMA S5.

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

PROOF OF LEMMA S5. Assume $\boldsymbol{\theta x} = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} \boldsymbol{H}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k},j} \boldsymbol{x}^j$ for all vectors $\boldsymbol{x}$. Because columns of $\boldsymbol{H}$ are orthonormal, we know that

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

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

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

By the Cauchy–Schwarz inequality, we know that

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

and the equation is achievable. Thus,

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

□

S8.2. *Gradient and Hessian Expressions.* For completeness, we rederive the gradient and Hessian for negative log-likelihood of Poisson and multinomial categorical response models.

S8.2.0.1. 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

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

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

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

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

S8.2.0.2. 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 $\mathrm{vec}_{\boldsymbol{J}}\left(p^{\boldsymbol{J}}(\boldsymbol{x})\right) = \frac{e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}}{\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}}\rangle}$. Note that

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

Thus, we have

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

$$\begin{aligned}
&\frac{\partial^2\mathcal{L}_n^{\mathrm{Mult}}(\boldsymbol{\beta})}{\partial\,\mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top}\\
=&\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top \otimes \boldsymbol{H}^\top\Big\{\mathrm{diag}[\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i))] - \mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i))\{\mathrm{vec}_{\boldsymbol{J}}(p^{\boldsymbol{J}}(\boldsymbol{x}_i))\}^\top\Big\}\boldsymbol{H}.
\end{aligned}\tag{S24}$$

S8.3. *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

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

LEMMA S6. *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*

$$\left\|\nabla\mathcal{L}_n(\boldsymbol{\beta}) - \nabla\mathcal{L}_n(\boldsymbol{\beta}')\right\| \le \left\|\boldsymbol{\beta} - \boldsymbol{\beta}'\right\| \cdot \max_{\substack{0\le t\le 1;\\ \boldsymbol{\beta}_t = t\boldsymbol{\beta} + (1-t)\boldsymbol{\beta}'}} \left\|\frac{\partial^2\mathcal{L}_n(\boldsymbol{\beta}_t)}{\partial\,\mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}}.\tag{S25}$$

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

$$\left\|\frac{\partial^2\mathcal{L}_n(\boldsymbol{\beta})}{\partial\,\mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}} \le \frac{1}{2n}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}) \textit{ and}\tag{S26}$$

$$\left\|\frac{\partial^2\mathcal{L}_n(\mathbf{0})}{\partial\,\mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}} = \frac{1}{n\cdot|\boldsymbol{J}|}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}).\tag{S27}$$

For Poisson categorical response model, we have

$$\left\|\frac{\partial^2\mathcal{L}_n(\mathbf{0})}{\partial\,\mathrm{vec}(\boldsymbol{\beta})\partial\{\mathrm{vec}(\boldsymbol{\beta})\}^\top}\right\|_{\mathrm{op}} = \frac{1}{n}\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}).\tag{S28}$$

PROOF OF LEMMA S6. 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 (S25).
Let $p(\boldsymbol{x}) = \frac{1}{\langle \mathbf{1}_{|\boldsymbol{J}|}, e^{\boldsymbol{H\beta x}}\rangle} e^{\boldsymbol{H\beta x}}$. By (S24), we obtain

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

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 (S26).
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}\Big\|p^\top(\mathbf{0})\boldsymbol{H}\Delta\boldsymbol{\beta}\boldsymbol{X}\Big\|^2\\
= &\frac{1}{n\cdot|\boldsymbol{J}|}\max_{\|\Delta\boldsymbol{\beta}\|=1}\Big\|\big(\boldsymbol{X}^\top\otimes I\big)\operatorname{vec}(\Delta\boldsymbol{\beta})\Big\|^2 - \Big\|\boldsymbol{e}_1^\top\Delta\boldsymbol{\beta}\boldsymbol{X}\Big\|^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,

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

□

The next lemma establishes the critical value of the regularization parameter, λ_1 defined in Appendix S2, for the group lasso penalty $\Omega_{\mathcal{G}}(\boldsymbol{\beta})$.

LEMMA S7. *If $\Omega_{\mathcal{G}}(\boldsymbol{\beta})$ denotes the group lasso penalty defined in (17), 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 (20), 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 S7. Assume

$$\mathbf{0}\in\frac{\partial}{\partial\boldsymbol{\beta}}\mathcal{L}_n(\mathbf{0})+\lambda\partial\Omega_{\mathcal{G}}(\mathbf{0}),\tag{S29}$$

where ∂ denote the subgradient.

We know that (S29) 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,

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

which implies that

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

□

S8.4. *Monotone Property of Reparameterized Grouped 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 S8. *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 S8. 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
$$\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}\right\| \leq \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{\beta}_{\boldsymbol{k},j}\right\| = \Omega_{\mathcal{G}}(\boldsymbol{\beta}),$$
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
$$\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger)) = \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}^\dagger\right\| \leq \sum_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}\left\|\boldsymbol{\beta}_{\boldsymbol{k},j}^\dagger\right\| = \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger) \leq \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta}^\dagger)),$$
we obtain that if $w_{\boldsymbol{k},j} > 0$, then $\left\|\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}^\dagger\right\| = \left\|\boldsymbol{\beta}_{\boldsymbol{k},j}^\dagger\right\|$, and thus $\boldsymbol{H}_{\boldsymbol{k}}^\top\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k},j}^\dagger = \boldsymbol{\beta}_{\boldsymbol{k},j}^\dagger$ 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
$$\mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) = L_n(\boldsymbol{H}\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) = L_n(\boldsymbol{P}_{\mathcal{V}}\boldsymbol{H}\widehat{\boldsymbol{\beta}}) = L_n(\boldsymbol{H}\widehat{\boldsymbol{\beta}}) = \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) \text{ and}$$
$$\begin{aligned}&\mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) = L_n(\boldsymbol{H}\widehat{\boldsymbol{\beta}}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}}))\\ \leq& L_n(\boldsymbol{H}\widehat{\boldsymbol{\beta}}) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) = \mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}})\\ =&\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda\Omega_{\mathcal{G}}(\widehat{\boldsymbol{\beta}}) \leq \mathcal{L}_n(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\widehat{\boldsymbol{\beta}})).\end{aligned}$$
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 (28), we have $\boldsymbol{\beta}^\dagger \in \mathcal{F}_{\boldsymbol{\theta}^*}$,

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

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 S8, 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.

□

S8.5. *Proof of Lemma S10.* To show that the γ_M in (S11) for multinomial and Poisson categorical response model both satisfy Assumptions S1, we only need to show Lemma S10. To show Lemma S10 for multinomial categorical response model, we introduce the following lemma.

LEMMA S9. *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 $u_{\max} = \max_{1\le i\le m} u_i$ and $u_{\min} = \min_{1\le i\le m} u_i$. Then*

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

PROOF OF LEMMA S9. First of all, we know that for any $\boldsymbol{u}$,

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

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

$$1 = \sum_{i=1}^m p_i \le m\cdot p_{\max} = m\cdot e^{u_{\max}-u_{\min}}\cdot p_{\min}, \tag{S30}$$

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

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

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

Combined with inequality (S30), we obtain that

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

□

LEMMA S10. *For multinomial categorical response model with $n_i = 1$, and $\psi(\boldsymbol{u}) = \log(\sum_{l=1}^{|\boldsymbol{J}|} e^{u_l})$, we have*

$$\gamma_M \geq \frac{1}{|\boldsymbol{J}| \cdot \exp\left(4M\right)}. \tag{S31}$$

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

$$\gamma_M \geq e^{-M}. \tag{S32}$$

PROOF OF LEMMA S10. Let $\boldsymbol{u} = (u_1, \cdots, u_{|\boldsymbol{J}|})^\top = \boldsymbol{\theta x}$. We know that

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

where $u_{\max} = \max_{1 \leq l \leq |\boldsymbol{J}|} u_l$ and $u_{\min} = \min_{1 \leq l \leq |\boldsymbol{J}|} u_l$. For multinomial categorical response model, by Lemma S9 we obtain (S31). For Poisson categorical response model, we know that $\lambda_{\min}\left(\nabla^2 \psi(\boldsymbol{u})\right) = e^{u_{min}}$. This completes the proof of (S32). □

S8.6. *Norm and Dual Norm for Group Lasso.* Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_t] \in \mathcal{M}_H \subset \mathbb{R}^{|\boldsymbol{J}| \times p}$, $\boldsymbol{\theta}_j \in \mathbb{R}^{|\boldsymbol{J}| \times p_j}, 1 \leq j \leq t$. Recall that $\mathcal{M}_H := \{\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{J}| \times p}; \boldsymbol{\theta x} = \sum_{(\boldsymbol{k},j) \in \mathcal{G}} \boldsymbol{H_k \beta_{k,j} x}^j, \forall \boldsymbol{x} \in \mathbb{R}^p\}$. Due to mapping $\boldsymbol{\beta} \mapsto \boldsymbol{H\beta}$ is one to one, by (18), we know that

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

LEMMA S11. *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 S11. 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

$$\Phi_{\mathcal{G}}^*(\boldsymbol{\alpha}) = \max\left\{ \langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle ; \Phi_{\mathcal{G}}(\boldsymbol{\theta}) \leq 1, \forall \boldsymbol{\theta} \in \mathcal{M}_H \right\}.$$

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

LEMMA S12. *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 S12. 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:

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

□

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 $\boldsymbol{\theta}\in\mathcal{M}_H$

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

which leads to the following Lemma.

LEMMA S13. *For any $\boldsymbol{\theta}\in\mathcal{M}_H$,*

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

S8.7. *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 S14. *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 S14. 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

$$\Omega_{\mathcal{G}}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta})) = \Phi_{\mathcal{G}}^*(\boldsymbol{H}\boldsymbol{H}^\top \nabla L_n(\boldsymbol{H}\boldsymbol{\beta})) = \max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}_{\boldsymbol{k}}^\top \frac{\partial}{\partial \boldsymbol{u}} \ell(\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{x}_i, \boldsymbol{y}_i)\{\boldsymbol{x}_{i(j)}\}^\top \right\|$$

$$= \max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}) \right\|.$$

Applying Hölder's inequality twice, we know that

$$|\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}), \Delta\boldsymbol{\beta}\rangle| \le \sum_{g\in\mathcal{G}} \left\| \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_g} \right\| \|\Delta\boldsymbol{\beta}_g\| \le \Omega_{\mathcal{G}}^*(\nabla \mathcal{L}_n(\boldsymbol{\beta})) \cdot \Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}).$$

□

S9. Tail Bound Estimates. Before we show Theorem S1, we first prove the sub-exponential tail bound of $\Omega_{\mathcal{G}}^*(\mathcal{L}_n(\boldsymbol{\beta}^\dagger))$ for the Poisson categorical response model, and the sub-Gaussian tail bound for the multinomial categorical response model.

LEMMA S15. *For Poisson categorical response model, there exists absolute constants A_1, A_2 such that for any $u \ge 0$,*

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \right\| \ge u \Big) \le \exp\Big\{ -\frac{A_1 \cdot n}{C^2\Lambda} \big(\min\{u, A_2 C\sqrt{\Lambda}\}\big)^2 + \log(5)m + \log|\mathcal{G}| \Big\}.$$

For multinomial categorical response model, for any $u \ge 0$, we have

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \right\| \ge u \Big) \le \exp\Big(-\frac{nu^2}{4C^2} + \log(5)m + \log|\mathcal{G}| \Big).$$

COROLLARY S1. *For Poisson categorical response model, there exists absolute constants B, B_1, B_2, B' such that if $\frac{m}{n} + \frac{\log|\mathcal{G}|}{n} \le B_1$ and $\lambda = BC\sqrt{\Lambda}\Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\Big)$, then*

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1} \left\| \frac{\partial}{\partial \boldsymbol{\beta}_{\boldsymbol{k},j}} \mathcal{L}_n(\boldsymbol{\beta}^\dagger) \right\| \ge \frac{\lambda}{2} \Big) \le \exp\big(-B'\delta^2 n\big), 0 \le \delta \le B_2.$$

For multinomial categorical response model, there exists absolute constants B, B_1, B' such that if $\frac{m}{n} + \frac{\log|\mathcal{G}|}{n} \le B_1$ and $\lambda = BC\Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\Big)$, then

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1}\Big\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^\dagger)\Big\| \ge \frac{\lambda}{2}\Big) \le \exp\big(-B'\delta^2 n\big), \delta \ge 0.$$

PROOF OF LEMMA S15. **Case 1: Poisson categorical response model** First of all, define the Orlicz space for mean zero random variable

$$L_{\psi_1}^0 := \{X; \mathbb{E}X = 0, \exists K > 0, \mathbb{E}\exp(|X/K|) \le 2\}.$$

Consider the following two norm on Orlicz space $L_{\psi_1}^0$

$$\|X\|_{\psi_1} := \inf\{K > 0; \mathbb{E}\exp(|X/K|) \le 2\},$$
$$\tau_{\varphi_1}(X) := \inf\{K > 0; \forall t \in [-1,1], \mathbb{E}\exp(tX) \le \exp(K^2t^2/2)\}.$$

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

$$\frac{d}{d\xi}\frac{e^\xi - \xi - 1}{\xi^2} = \frac{e^\xi(\xi-2)+\xi+2}{\xi^3} = \frac{\int_0^\xi du_1\int_0^{u_1} e^{u_2}u_2du_2}{\xi^3} > 0, \forall\xi \in \mathbb{R}\backslash\{0\},$$

and

$$\begin{aligned}&\inf\{K > 0; \forall t \in [-1,1], \exp(\lambda(e^t - t - 1)) \le \exp(K^2t^2/2)\}\\ =&\inf\{K > 0; \forall t \in [-1,1], 2\lambda\frac{e^t - t - 1}{t^2} \le K^2\} = \sqrt{2(e-2)\lambda},\end{aligned}$$

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 $L_{\psi_1}^0$. 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 $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N \in \mathbb{R}^{|\boldsymbol{k}|_J\cdot p_j}$ such that $N \le 5^{|\boldsymbol{k}|_J\cdot p_j}$, $\|\boldsymbol{v}_l\| \le 1, 1 \le l \le N$, for any $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{k}|_J\cdot p_j}$ 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,

$$\begin{aligned}&\mathbb{P}\Big(w_{\boldsymbol{k},j}^{-1}\Big\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^\dagger)\Big\| \ge u\Big)\\ \le&\sum_{1\le l\le N}\mathbb{P}\Big(\sum_{i=1}^n\Big\langle\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}, \boldsymbol{y}_i - e^{\boldsymbol{H}\boldsymbol{\beta}^\dagger\boldsymbol{x}_i}\Big\rangle \ge \frac{w_{\boldsymbol{k},j}nu}{2}\Big).\end{aligned}$$

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 $1\le i\le n$. 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 $\lambda_{i,\boldsymbol{j}}, 1\le i\le n, \boldsymbol{j}\in\boldsymbol{J}$.

We aim to use Bernstein's inequality for sub-exponential random variables to estimate the following tail probability
(S33)

$$\mathbb{P}\Big(\sum_{i=1}^{n}\Big\langle \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)},\boldsymbol{y}_i-e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\Big\rangle\ge\frac{w_{\boldsymbol{k},j}nu}{2}\Big)=\mathbb{P}\Big(\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\boldsymbol{J}}v_{i,\boldsymbol{j}}(y_{i,\boldsymbol{j}}-\lambda_{i,\boldsymbol{j}})\ge\frac{w_{\boldsymbol{k},j}nu}{2}\Big).$$

Notice that $\sum_{\boldsymbol{j}\in\boldsymbol{J}}v_{i,\boldsymbol{j}}^2=\left\|\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|^2\le\left\|\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|^2\le\left\|\boldsymbol{x}_{i(j)}\right\|^2$, for any $1\le l\le N$, $1\le j\le t$, and $1\le i\le n$.

It is straightforward to show that

$$\max_{\boldsymbol{j}\in\boldsymbol{J}}|v_{i,\boldsymbol{j}}|=\left\|\boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|_\infty\le\left\|\boldsymbol{v}_l\boldsymbol{x}_{i(j)}\right\|\le\left\|\boldsymbol{x}_{i(j)}\right\|$$

In conclusion, we know that

$$\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\boldsymbol{J}}\|v_{i,\boldsymbol{j}}(y_{i,\boldsymbol{j}}-\lambda_{i,\boldsymbol{j}})\|_{\psi_1}^2\le C_0^2\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\boldsymbol{J}}v_{i,\boldsymbol{j}}^2\lambda_{i,\boldsymbol{j}}\le C_0^2\Lambda\sum_{i=1}^{n}\left\|\boldsymbol{x}_{i(j)}\right\|^2\le C_0^2C^2w_{\boldsymbol{k},j}^2\Lambda n,$$

and

$$\max_{1\le i\le N}\max_{\boldsymbol{j}\in\boldsymbol{J}}\|v_{i,\boldsymbol{j}}(y_{i,\boldsymbol{j}}-\lambda_{i,\boldsymbol{j}})\|_{\psi_1}\le C_0\sqrt{\Lambda}\max_{1\le i\le n}\left\|\boldsymbol{x}_{i(j)}\right\|\le C_0Cw_{\boldsymbol{k},j}\sqrt{\Lambda},$$

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

$$\begin{aligned}&\mathbb{P}\Big(\sum_{i=1}^{n}\Big\langle \boldsymbol{H}_{\boldsymbol{k}}\boldsymbol{v}_l\boldsymbol{x}_{i(j)},\boldsymbol{y}_i-e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\Big\rangle\ge\frac{(w_{\boldsymbol{k},j}\cdot n)u}{2}\Big)\\ \le&\exp\Big\{-c\cdot\min\Big(\frac{nu^2}{4C_0^2C^2\Lambda},\frac{nu}{2C_0C\sqrt{\Lambda}}\Big)\Big\}\\ \le&\exp\Big\{-\frac{c\cdot n}{4C_0^2C^2\Lambda}\big(\min\{u,2C_0C\sqrt{\Lambda}\}\big)^2\Big\}.\end{aligned}$$

By the union bound, we obtain

$$\begin{aligned}&\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}}w_{\boldsymbol{k},j}^{-1}\left\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\right\|\ge u\Big)\\ \le&\exp\Big(-\frac{c\cdot n}{4C_0^2C^2\Lambda}\big(\min\{u,2C_0C\sqrt{\Lambda}\}\big)^2+\log(5)m+\log|\mathcal{G}|\Big).\end{aligned}$$

Let absolute constants $A_1=\frac{c}{4C_0^2}$ and $A_2=2C_0$, we complete the proof of Lemma S15.

Case 2: Multinomial categorical response model

Similar to previous discussion, we obtain

$$\left\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\right\|\le\frac{2}{n}\max_{1\le l\le N}\sum_{i=1}^{n}\Big\langle\boldsymbol{v}_l,\boldsymbol{H}_{\boldsymbol{k}}^{\top}\big(\boldsymbol{y}_i-p_{\boldsymbol{J}}(\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i)\big)\{\boldsymbol{x}_{i(j)}\}^{\top}\Big\rangle,$$

where $p_{\boldsymbol{J}}(\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i)=\frac{1}{\langle 1,e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}\rangle}e^{\boldsymbol{H}\boldsymbol{\beta}^{\dagger}\boldsymbol{x}_i}$.

By Lemma S16, 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

$$\mathbb{P}\Big(\max_{(\boldsymbol{k},j)\in\mathcal{G}} w_{\boldsymbol{k},j}^{-1}\Big\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^\dagger)\Big\| \ge u\Big) \le \exp\Big(-\frac{nu^2}{4C^2}+\log(5)m+\log|\mathcal{G}|\Big).$$

□

LEMMA S16. *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 $\boldsymbol{P}_{\boldsymbol{1}} = I - \frac{1}{k}\boldsymbol{1}_k\boldsymbol{1}_k^\top$. Then for any $\boldsymbol{t}\in\mathbb{R}^k$, we have*

$$\mathbb{E}\exp\big(\sum_{i=1}^{k} t_i(Y_i - np_i)\big) \le \exp\big(\frac{n\left\|\boldsymbol{P}_{\boldsymbol{1}}\boldsymbol{t}\right\|^2}{4}\big).$$

PROOF OF LEMMA S16. Based on the moment generation function of multinomial distribution random vector, we obtain

(S34)

$$\log\mathbb{E}\exp\big(\sum_{i=1}^{k} t_i(Y_i - np_i)\big) = n\Big(\log\big(\sum_{i=1}^{k} p_i e^{t_i}\big) - \sum_{i=1}^{k} p_i t_i\Big) = n\big(f(1) - f(0) - f'(0)\big),$$

where $f(x) = \log(\sum_{i=1}^k p_i e^{xt_i})$.

Applying Taylor series with Lagrange remainder, we obtain

$$f(1) - f(0) - f'(0) \le \frac{1}{2}\max_{0\le x\le 1} f''(x). \tag{S35}$$

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 $1 \le j \le k$. It is easy to check that $\widetilde{p}_j \le 1$ and $\sum_{j=1}^k \widetilde{p}_j = 1$.

We can rewrite $f''(x)$ as below,

$$f''(x) = \big(\sum_{i=1}^{k} t_i\widetilde{p}_i\big)^2 - \big(\sum_{i=1}^{k} t_i^2\widetilde{p}_i\big) = \boldsymbol{t}^\top[\mathrm{diag}(\widetilde{\boldsymbol{p}}) - \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^\top]\boldsymbol{t},$$

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 $[\mathrm{diag}(\widetilde{\boldsymbol{p}}) - \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^\top]$, with the corresponding eigenvector $\boldsymbol{1}_k$. By Geršgorin circle theorem (see Theorem 1.1 in Varga (2010)), we know that the maximum eigenvalue is upper bounded by

$$\max_{1\le i\le n}\widetilde{p}_i(1-\widetilde{p}_i) + \sum_{j:j\ne i}\widetilde{p}_j\widetilde{p}_i = 2\max_{1\le i\le n}\widetilde{p}_i(1-\widetilde{p}_i) \le \frac{1}{2}.$$

Hence, we know that

$$\boldsymbol{t}^\top[\mathrm{diag}(\widetilde{\boldsymbol{p}}) - \widetilde{\boldsymbol{p}}\widetilde{\boldsymbol{p}}^\top]\boldsymbol{t} \le \frac{1}{2}\left\|\boldsymbol{P}_{\boldsymbol{1}}\boldsymbol{t}\right\|^2.$$

Combined with (S34) and (S35), we complete the proof of Lemma S16. □

S10. Proof of Theorem S1.

S10.1. *Proof Overview.* The proof of Theorem S1 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 S17. Next, we provide a general bound for $\Delta\boldsymbol{\beta}$ in Lemma S18. Finally, as a consequence of the above lemmas, we obtain the proofs of Theorem S1 and its special case, Theorem 4.

S10.2. *A Star-Shaped Set.* Define orthogonal projection operator

$$\Big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta})\Big)_{\boldsymbol{k},j} = \boldsymbol{\beta}_{\boldsymbol{k},j}, (\boldsymbol{k}, j) \in \mathcal{S} \text{ and } \Big(\boldsymbol{Q}_{\mathcal{S}}(\boldsymbol{\beta})\Big)_{\boldsymbol{k},j} = \mathbf{0}, (\boldsymbol{k}, j) \in \mathcal{G}\backslash\mathcal{S},$$

where we denote the support of $\boldsymbol{\beta}^{\dagger}$ as $\mathcal{S} = \{g \in \mathcal{G}; \boldsymbol{\beta}_g^{\dagger} \neq 0\}$

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 S8.

We know that if $\Delta\boldsymbol{\beta} \in \mathbb{C}(\mathcal{S}, \phi)$, then

$$\Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}) \leq (\phi + 1)\sum_{g\in\mathcal{S}} w_g \left\|\Delta\boldsymbol{\beta}_g\right\|. \tag{S36}$$

Similar to Proposition 9.13 in Wainwright (2019), we obtain the following lemma.

LEMMA S17. *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,*

$$\Omega_{\mathcal{G}}^*(\nabla\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})) = \max_{(\boldsymbol{k},j)\in\mathcal{G}} \quad w_{\boldsymbol{k},j}^{-1}\left\|\frac{\partial}{\partial\boldsymbol{\beta}_{\boldsymbol{k},j}}\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\right\|.$$

PROOF OF LEMMA S17. 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 S8, we know that $\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta}) = \Delta\boldsymbol{\beta}$. Set $u = \frac{\phi-1}{\phi+1}$. Combined with Lemma S14, we know that

$$\begin{aligned}
0 &\geq \mathcal{L}_n(\boldsymbol{\beta}^{\dagger} + \Delta\boldsymbol{\beta}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}^{\dagger} + \Delta\boldsymbol{\beta}) - (\mathcal{L}_n(\boldsymbol{\beta}^{\dagger}) + \lambda\Omega_{\mathcal{G}}(\boldsymbol{\beta}^{\dagger}))\\
&\geq \Big\langle \Delta\boldsymbol{\beta}, \nabla\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})\Big\rangle + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^{\dagger} + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^{\dagger}))\\
&\geq -\Omega_{\mathcal{G}}^*(\nabla\mathcal{L}_n(\boldsymbol{\beta}^{\dagger})) \cdot \Big(\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})) + \Omega_{\mathcal{G}}\big((I - \boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big)\Big) + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^{\dagger} + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^{\dagger})).
\end{aligned}$$

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

$$0 \ge \lambda\Big((1-u)\Omega_{\mathcal{G}}\big((I-\boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) - (1+u)\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big)\Big),$$

which implies

$$\Omega_{\mathcal{G}}\big((I-\boldsymbol{Q}_{\mathcal{S}})(\Delta\boldsymbol{\beta})\big) \le \frac{1+u}{1-u}\Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big) = \phi \cdot \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})\big).$$

Combined with Lemma S8, we obtain that $\Delta\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\dagger \in \mathbb{C}(\mathcal{S}, \phi)$. □

S10.3. *Non-Asymptotic bound.*

LEMMA S18. *Under Assumptions 1-4 and over event* $\mathbb{G}(\lambda) = \{\Omega^*_{\mathcal{G}}(\nabla\mathcal{L}_n(\boldsymbol{\beta}^\dagger)) \le \frac{\lambda}{2}\}$*, if* $A \cdot C^2\left(\frac{\log|\mathcal{G}|}{n} + \frac{m}{n}\right) \le \frac{\kappa}{64\Psi^2(\mathcal{S})}$ *and* $\lambda \le \frac{R\cdot\kappa}{6\Psi(\mathcal{S})}$*, then*

$$\left\|\boldsymbol{\beta}^\dagger - \widehat{\boldsymbol{\beta}}\right\| \le \frac{6\lambda\Psi(\mathcal{S})}{\kappa}.$$

PROOF OF LEMMA S18. 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 S8 and Assumption 4, we know that

$$\mathcal{F}(\Delta\boldsymbol{\theta}) \ge \left\langle \nabla L_n(\boldsymbol{\theta}^\dagger), \boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta} \right\rangle + \frac{\kappa}{2}\|\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \Delta\boldsymbol{\beta}) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)).$$

where $\tau_n^2 = A \cdot C^2\left(\frac{\log|\mathcal{G}|}{n} + \frac{m}{n}\right)$.

By Lemma S8, 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 $\boldsymbol{\beta}$,

$$\Omega_{\mathcal{G}}(\boldsymbol{\beta}) \ge \Omega_{\mathcal{G}}\big(\boldsymbol{Q}_{\mathcal{V}}(\boldsymbol{\beta})\big).$$

Thus, we obtain that

$$\begin{aligned}
&\mathcal{F}(\Delta\boldsymbol{\theta}) \\
\ge&\left\langle \nabla L_n(\boldsymbol{\theta}^\dagger), \boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta} \right\rangle + \frac{\kappa}{2}\|\boldsymbol{P}_{\mathcal{V}}\Delta\boldsymbol{\theta}\|^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)) \\
\ge& - \Omega^*_{\mathcal{G}}(\nabla\mathcal{L}_n(\boldsymbol{\beta}^\dagger)) \cdot \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \frac{\kappa}{2}\|\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) \\
&+ \lambda(\Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger + \boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) - \Omega_{\mathcal{G}}(\boldsymbol{\beta}^\dagger)) \\
\ge& - \frac{\lambda}{2}\Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \frac{\kappa}{2}\|\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})\|^2 - \tau_n^2\Omega_{\mathcal{G}}^2(\boldsymbol{Q}_{\mathcal{V}}(\Delta\boldsymbol{\beta})) + \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}\|\Delta\boldsymbol{\beta}\|^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) \\
\ge&\frac{\kappa}{2}\|\Delta\boldsymbol{\beta}\|^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) \\
\ge&\frac{\kappa}{2}\|\Delta\boldsymbol{\beta}\|^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 (S36) and Lemma S5, we have

$$\Omega_{\mathcal{G}}(\Delta\boldsymbol{\beta}) \leq 4 \cdot \Omega_{\mathcal{G}}(\boldsymbol{Q}_{\mathcal{S}}(\Delta\boldsymbol{\beta})) \leq 4 \cdot \Psi(\mathcal{S}) \cdot \|\Delta\boldsymbol{\beta}\| \,.$$

This implies that

$$\mathcal{F}(\Delta\boldsymbol{\theta}) \geq \left(\frac{\kappa}{2} - 16\tau_n^2\Psi^2(\mathcal{S})\right) \cdot \|\Delta\boldsymbol{\beta}\|^2 - \frac{3\lambda}{2}\Psi(\mathcal{S}) \cdot \|\Delta\boldsymbol{\beta}\| \,.$$

Applying the assumed bound $16\tau_n^2\Psi^2(\mathcal{S}) \leq \frac{\kappa}{4}$, we obtain that

$$\mathcal{F}(\Delta\boldsymbol{\theta}) \geq \left(\frac{\kappa}{4}\right)\left(\|\Delta\boldsymbol{\beta}\| - \frac{6\lambda}{\kappa}\Psi(\mathcal{S})\right)\|\Delta\boldsymbol{\beta}\| > 0,$$

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 out 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

$$\|\Delta\boldsymbol{\theta}\|^2 = \mathrm{tr}\left(\{\Delta\boldsymbol{\theta}\}^\top\Delta\boldsymbol{\theta}\right) = \mathrm{tr}\left(\{\Delta\boldsymbol{\beta}\}^\top\boldsymbol{H}^\top\boldsymbol{H}\Delta\boldsymbol{\beta}\right) = \mathrm{tr}\left(\{\Delta\boldsymbol{\beta}\}^\top\Delta\boldsymbol{\beta}\right) = \|\Delta\boldsymbol{\beta}\|^2 \,.$$

All the above argument is valid as along as $\frac{6\lambda}{\kappa}\Psi(\mathcal{S}) \leq R$ by Assumption 4. □

S10.4. *Proof of Theorem S1.*

PROOF OF THEOREM S1. The proof of Theorem S1 is a straightforward application of Lemma S18 and Corollary S1, under Assumptions 1-4.

□

REFERENCES

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

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

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*, pages 3–17. Springer.

Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*.

Roman, S., Axler, S., and Gehring, F. (2005). *Advanced linear algebra*, volume 3. Springer.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

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.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Zajkowski, K. (2020). On norms in some class of exponential type orlicz spaces of random variables. *Positivity*, 24(5):1231–1240.