entropy

MDPI

Article

Sparse Multicategory Generalized Distance Weighted Discrimination in Ultra-High Dimensions

Tong Su [1], Yafei Wang [2], Yi Liu [2], William G. Branton [3], Eugene Asahchop [3], Christopher Power [3], Bei Jiang [2], Linglong Kong [2,*] and Niansheng Tang [1,*]

1 Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming 650091, China; sutong_366@sina.com
2 Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada; yafei2@ualberta.ca (Y.W.); yliu16@ualberta.ca (Y.L.); bei1@ualberta.ca (B.J.)
3 Department of Medicine (Neurology), University of Alberta, Edmonton, AB T6G 2G1, Canada; wbranton@ualberta.ca (W.G.B.); asahchop@ualberta.ca (E.A.); cp9@ualberta.ca (C.P.)
* Correspondence: lkong@ualberta.ca (L.K.); nstang@ynu.edu.cn (N.T.)

Received: 30 September 2020; Accepted: 2 November 2020; Published: 5 November 2020

Abstract: Distance weighted discrimination (DWD) is an appealing classification method that is capable of overcoming data piling problems in high-dimensional settings. Especially when various sparsity structures are assumed in these settings, variable selection in multicategory classification poses great challenges. In this paper, we propose a multicategory generalized DWD (MgDWD) method that maintains intrinsic variable group structures during selection using a sparse group lasso penalty. Theoretically, we derive minimizer uniqueness for the penalized MgDWD loss function and consistency properties for the proposed classifier. We further develop an efficient algorithm based on the proximal operator to solve the optimization problem. The performance of MgDWD is evaluated using finite sample simulations and miRNA data from an HIV study.

Keywords: high dimension; multicategory classification; DWD; sparse group lasso; L_2-consistency; proximal algorithm

1. Introduction

Classification problems appear in diverse practical applications, such as spam e-mail classification, disease diagnosis and drug discovery, among many others (e.g., [1–3]). In these classification problems, the goal is to predict class labels based on a given set of variables. Recent research has focused extensively on linear classification: see [4,5] for comprehensive introductions. Among many linear classification methods, support vector machines (SVMs) (see [6,7]) and distance-weighted discrimination (DWD) (see [8–10]) are two commonly used large-margin based classification methods.

Owing to the recent advent of new technologies for data acquisition and storage, classification with high dimensional features, i.e., a large number of variables, has become a ubiquitous problem in both theoretical and applied scientific studies. Typically, only a small number of instances are available, a setting we refer to as high-dimensional, low-sample size (HDLSS), as in [11]. In the HDLSS setting, a so-called "data-piling" phenomenon is observed in [8] for SVMs, occurring when projections of many training instances onto a vector normal to the separating hyperplane are nearly identical, suggesting severe overfitting. DWD was originally proposed to overcome data-piling in the HDLSS setting. In binary classification problems, linear SVMs seek a hyperplane maximizing the smallest margin for all data points, while DWD seeks a hyperplane minimizing the sum of inverse margins over all data points. Reference [8] suggests replacing the inverse margins by the q-th power of the inverse margins in a generalized DWD method; see [12] for a detailed description. Formally, for a training

data set $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^N$ of N observations, where $\boldsymbol{X}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, binary generalized linear DWD seeks a proper separating hyperplane $\{\boldsymbol{X} : a + \boldsymbol{X}^\top \boldsymbol{b} = 0\}$ through the optimization problem

$$\begin{aligned} \arg\max_{a,\boldsymbol{b}} & \sum_{i=1}^N \frac{1}{d_i^q} \\ \text{s.t. } & d_i = y_i \left(a + \mathbf{X}_i^T \boldsymbol{b}\right) + \eta_i \geq 0, \forall i, \\ & \eta_i \geq 0, \forall i, \sum_i \eta_i \leq c, \\ & \|\boldsymbol{b}\|_2^2 = 1, \end{aligned} \tag{1}$$

where a and $\boldsymbol{b}$ are the intercept and slope parameters, respectively. The slack variable η_i is introduced to ensure that the corresponding margin d_i is non-negative and the constant $c > 0$ is a tuning parameter to control the overlap between classes. Problem (1) can also be written in a loss-plus-penalty form (e.g., [12]) as

$$(\hat{a}, \hat{\boldsymbol{b}}) = \underset{a,\boldsymbol{b}}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^N \phi_q \left\{ y_i \left(a + \boldsymbol{X}_i^\top \boldsymbol{b}\right) \right\} + \lambda \|\boldsymbol{b}\|_2^2 \right], \tag{2}$$

where

$$\phi_q(u) = \begin{cases} 1 - u, & \text{if } u \leq Q \\ \varphi_q(u), & \text{if } u > Q, \end{cases} \tag{3}$$

with $Q = \frac{q}{q+1}$, $q > 0$ and $\varphi_q(u) = (1 - Q)(Qu^{-1})^q$. When $q = 1$, (1) becomes the standard DWD problem in [8] while problem (2) appears in [9,13].

The binary classification problem (1) is well studied. However, in many applications such as image classification [1], cancer diagnosis [2] and speech recognition [3], to name a few, problems with more than two categories are commonplace. To solve these multicategory problems with the DWD classifier, approaches based on either formulation (1) or (2) are common. One common strategy is to extend problem (1) to multiple classes by solving a series of binary problems in a one-versus-one (OVO) or one-versus-rest (OVR) method (e.g., [14]). Instead of reducing the multicategory problem to a binary one, another strategy based on problem (1) considers all classes at once. As shown in [14], this approach generally works better than the OVO and OVR methods. Based on an extension of problem (2), [15] proposes multicategory DWD, written in a loss-plus-penalty form as

$$\begin{aligned} \min_{a_k, b_k} & \frac{1}{N} \sum_{i=1}^N \phi_q \left(a_{y_i} + \boldsymbol{X}_i^\top \boldsymbol{b}_{y_i}\right) + \lambda \sum_{k=1}^K \|\boldsymbol{b}_k\|_2^2 \\ \text{s.t. } & \sum_{k=1}^K a_k = 0; \sum_{k=1}^K b_{jk} = 0, \quad \forall j = 1, \ldots, p, \end{aligned} \tag{4}$$

with y_i, $k \in \{1, \ldots, K\}$ and where a_k and $\boldsymbol{b}_k = (b_{1k}, \cdots, b_{pk})$ are the intercept and slope parameters for each category k, respectively. Although these methods can be applied to multicategory classification in the HDLSS setting, both problems (2) and (4) use the L_2 penalty and do not perform feature selection. As discussed in [16], for high dimensional classification, taking all features into consideration does not work well for two reasons. First, based on prior knowledge, only a small number of variables are relevant to the classification problem: a good classifier in high dimensions should have the ability to sparsely select important variables and discard redundant ones. Second, classifiers using all available variables in high-dimensional settings may have poor classification performance.

Much of the SVM literature has considered variable selection in high-dimensional classification problems to improve performance (e.g., [17–19]). Among the DWD literature, to the best of our knowledge, only [16] considered variables selection and classification simultaneously. Wang and Zou [16] considered an L_1 rather than an L_2 penalty in problem (2) to improve interpretability through sparsity in the binary classification. Moreover, [16] made selections based on the strengths of input variables within individual classes but ignored the strengths of input variable groupings, thereby selecting more factors than necessary for each class. To overcome this weakness in this paper, we developed a multicategory generalized DWD method that is capable of performing variable selection and classification simultaneously. Our approach incorporates sparsity and group structure information via the sparse group lasso penalty (see [20–24]).

Although DWD is well studied, it is less popular than the SVM for binary classification, arguably for computational and theoretical reasons. For an up-to-date list of works on DWD mostly focused on the $q = 1$ case, see [14,15]. Theoretical asymptotic properties of large-margin classifiers in high dimensional settings were studied in [25], and [26] derived an expression for asymptotic generalization error. In terms of computation, [8] solved the standard DWD problem in (1) as a second-order cone programming (SOCP) problem using a primal-dual interior-point method that is computationally expensive when N or p is large. To overcome computational bottlenecks, [12] proposed an approach based on a novel formulation of the primal DWD model in (1): this method, proposed in [12], does not scale to large data sets and requires further work. Lam et al. [27] designed a new algorithm for large DWD problems with $q \geq 2$ and $K = 2$ based on convergent multi-block ADMM-type methods (see [28]). Wang and Zou [16] solved the lasso-penalized binary DWD problem by combining majorization–minimization and coordinate descent methods: the lasso penalty does not directly permit a SOCP solution. In fact, solution identifiability in the generalized DWD problem with $q > 1$ requires more constraints and remains an open research problem (see [8]). To the best of our knowledge, no work focusing on computational aspects of lasso penalized multicategory generalized DWD (MgDWD) exists. The same holds for sparse group lasso-penalized MgDWD.

The theoretical and computational contributions of this paper are as follows. First, we establish the uniqueness of the minimizer in the population form of the MgDWD problem. Second, we prove a non-asymptotic L_2 estimation error bound for the sparse group lasso-regularized MgDWD loss function in the ultra-high dimensional setting under mild regularity conditions. Third, we develop a fast, efficient algorithm able to solve the sparse group lasso-penalized MgDWD problem using proximal methods.

The rest of this paper is organized as follows. In Section 2.1, we introduce the MgDWD problem with sparse group lasso penalty. In Sections 2.2 and 2.3, we establish theoretical properties of the population classifier and regularized empirical loss. We propose a computational algorithm in Section 2.4. Section 3 illustrates the finite sample performance of our method through simulation studies and a real data analysis. Proofs for major theorems are given in the Appendix A.

2. Methodology

2.1. Model Setup

We begin with some basic set-up and notation. Consider the multicategory classification problem for a random sample $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^N$ of N independent and identically distributed (i.i.d.) observations from some distribution $\mathbb{P}(y, \boldsymbol{X})$. Here, y is the categorical response taking values in $\mathscr{Y} = \{1, \ldots, K\}$, and $\boldsymbol{X} = (x_1, \ldots, x_p)^\top \in \mathscr{X} \subset \mathbb{R}^p$ is the covariate vector. We wish to obtain a proper separating hyperplane $\{\boldsymbol{X} \in \mathscr{X} | a_k + \boldsymbol{X}^\top \boldsymbol{b}_k = 0\}$ for each category $k \in \mathscr{Y}$, where a_k and $\boldsymbol{b}_k = (b_{1k}, \ldots, b_{pk})^\top$ are intercept and slope parameters, respectively.

In this paper, we consider MgDWD with sparse group lasso regularization. That is, we estimate a classification boundary by solving the constrained optimization problem

$$\min_{a_k,\, \boldsymbol{b}_k} \frac{1}{N}\sum_{i=1}^{N}\phi_q\left(a_{y_i}+\boldsymbol{X}_i^\top \boldsymbol{b}_{y_i}\right)+\lambda_1\sum_{k=1}^{K}\sum_{j=1}^{p}|b_{jk}|+\lambda_2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}b_{jk}^2}$$
$$s.t.\ \sum_{k=1}^{K}a_k=0;\ \sum_{k=1}^{K}b_{jk}=0,\quad \forall j=1,\ldots,p, \tag{5}$$

where ϕ_q is as defined in (3).

To approach this problem, we apply the concept of a "margin vector" to extend the definition of a (binary) margin to the multicategory case. Denote the margin vector of an observation $\boldsymbol{X}_i$ as $\boldsymbol{F}_i=(f_{i1},\ldots,f_{iK})^\top$, with $f_{ik}=a_k+\boldsymbol{X}_i^\top\boldsymbol{b}_k$ satisfying $\sum_{k=1}^{K}f_{ik}=0$. Let $\boldsymbol{E}_i=(e_{i1},\ldots,e_{iK})^\top$ be the class indicator vector with $e_{ik}=\mathbb{1}\{y_i=k\}$. The multicategory margin of the data point $(y_i,\boldsymbol{X_i})$ is then given as $f_{iy_i}=a_{y_i}+\boldsymbol{X}_i^\top\boldsymbol{b}_{y_i}=\boldsymbol{E}_i^\top\boldsymbol{F}_i$. Therefore, the MgDWD loss can be rewritten as

$$\phi_q(a_{y_i}+\boldsymbol{X}_i^\top\boldsymbol{b}_{y_i})=\phi_q(\boldsymbol{E}_i^\top\boldsymbol{F}_i)=\boldsymbol{E}_i^\top\phi_q(\boldsymbol{F}_i)=\sum_{k=1}^{K}\mathbb{1}\{y_i=k\}\phi_q(a_k+\boldsymbol{X}_i^\top\boldsymbol{b}_k). \tag{6}$$

Based on (6), Lemma 1 describes the Fisher consistency of the MgDWD loss.

Lemma 1. *Given $\boldsymbol{X}=\boldsymbol{u}$, the minimizer of the conditional expectation of (6) is $\tilde{\boldsymbol{F}}(\boldsymbol{u})=\left(\tilde{f}_1(\boldsymbol{u}),\ldots,\tilde{f}_K(\boldsymbol{u})\right)^\top$, satisfying*

$$\operatorname*{argmax}_{k\in\mathscr{Y}}\tilde{f}_k(\boldsymbol{u})=\operatorname*{argmax}_{k\in\mathscr{Y}}\Pr\{y=k|\boldsymbol{X}=\boldsymbol{u}\},$$

where

$$\tilde{f}_k(\boldsymbol{u})=\begin{cases}Q\sqrt[q]{\dfrac{\Pr\{y=k|\boldsymbol{X}=\boldsymbol{u}\}}{\Pr\{y=k_*|\boldsymbol{X}=\boldsymbol{u}\}}}, & k\neq k_*\\ -Q\displaystyle\sum_{l\neq k_*}\sqrt[q]{\dfrac{\Pr\{y=l|\boldsymbol{X}=\boldsymbol{u}\}}{\Pr\{y=k_*|\boldsymbol{X}=\boldsymbol{u}\}}}, & k=k_*.\end{cases}$$

and $k_*=\operatorname*{argmin}_{k\in\mathscr{Y}}\Pr\{y=k|\boldsymbol{X}=\boldsymbol{u}\}$.

Consequently, $\tilde{f}_k(\boldsymbol{u})$ can be treated as an effective proxy of $\Pr\{y=k|\boldsymbol{X}=\boldsymbol{u}\}$ and, for any new observation $\boldsymbol{X}_*$, a reasonable prediction of its label y_* is

$$\hat{y}_*=\operatorname*{argmax}_{k\in\mathscr{Y}}\{a_k+\boldsymbol{X}_*^\top\boldsymbol{b}_k\}.$$

Speaking to the sparse group lasso (SGL) regularization in (5), the L_1 penalty encourages an element-wise sparse estimator that selects important variables for each category, indicated by $\hat{b}_{jk}\neq 0$. Assuming that parameters in different categories share the same information, we use an L_2 penalty to encourage a group-wise sparsity structure that removes covariates that are irrelevant across all categories, that is, where $\hat{\boldsymbol{\beta}}_j=(b_{1j},\ldots,b_{Kj})^\top=\mathbf{0}$. Specifically, let $\boldsymbol{x}_j=(x_{1j},\cdots,x_{Nj})^T$ and $\mathbf{B}=(b_{jk})\in\mathbb{R}_{jk}^{p\times K}$, where the k-th column $\boldsymbol{b}_k$ is the slope vector for the category label k and the j-th row $\boldsymbol{\beta}_j^\top$ is the group coefficient for the variable $\boldsymbol{x}_j$. If $\boldsymbol{x}_j$ is noise in the classification problem or is not

relevant to category label k, then the entry b_{jk} of $\mathbf{B}$ should be shrunk to exactly zero. The SGL penalty of (5) can be written as a convex combination of the lasso and group lasso penalties in terms of $\boldsymbol{\beta}_j$ as

$$\lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p}|b_{jk}| + \lambda_2 \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K} b_{jk}^2} = \lambda \sum_{j=1}^{p}\left\{\tau\|\boldsymbol{\beta}_j\|_1 + (1-\tau)\|\boldsymbol{\beta}_j\|_2\right\}, \tag{7}$$

where $\lambda > 0$ is the scale of the penalty and $\tau \in [0,1]$ tunes the propensity between the element-wise and group-wise sparsity structure.

2.2. Population MgDWD

In this subsection, some basic results pertaining to unpenalized population MgDWD are given. These results are necessary for further theoretical analysis.

Denote the marginal probability mass of y as $\Pr(y = k) = \pi_k$ with $\pi_k > 0$ and $\sum_{k=1}^{K}\pi_k = 1$, and the conditional probability density functions of $\boldsymbol{X}$ given $y = k$ by $g(\boldsymbol{X} \mid y = k) = g_k(\boldsymbol{X})$. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ be the collection of coefficient vectors $\boldsymbol{\theta}_k = (a_k, \boldsymbol{b}_k^\top)^\top$ for all labels and $\boldsymbol{Z} = (1, \boldsymbol{X}^\top)^\top$. The population version of the MgDWD problem in (6) is

$$\mathcal{L}(\boldsymbol{\vartheta}) = \mathbb{E}\{\mathbb{I}(\mathscr{Y})^\top \phi_q(\boldsymbol{\Theta}^\top \boldsymbol{Z})\} = \sum_{k=1}^{K}\pi_k \int_{\mathscr{X}} \phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k) g_k(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{8}$$

where $\boldsymbol{\vartheta} = \mathrm{vec}\{\boldsymbol{\Theta}\}$ is the vectorization of the matrix $\boldsymbol{\Theta}$ and $\mathbb{I}(\mathscr{Y}) = (\mathbb{1}\{y = 1\}, \ldots, \mathbb{1}\{y = K\})^\top$ is a random vector. Denote the true parameter value $\boldsymbol{\vartheta}^*$ as a minimizer of the population MgDWD problem, namely,

$$\boldsymbol{\vartheta}^* \in \underset{\boldsymbol{\vartheta}\in\mathscr{C}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\vartheta}),$$

where $\mathscr{C} = \left\{\boldsymbol{\vartheta} \in \mathbb{R}^{K(p+1)} \;\middle|\; \mathbf{C}\boldsymbol{\vartheta} = \mathbf{0}_K\right\}$ is the set of sum-constrained $\boldsymbol{\vartheta}$ with $\mathbf{C} = \mathbf{1}_K^\top \otimes \mathbf{I}_{p+1}$, where $\otimes$ denotes the Kronecker product.

To facilitate our theoretical analysis, we first define the gradient vector and Hessian matrix of the population MgDWD loss function. We then introduce some regularity conditions necessary to derive theoretical properties of this problem. Let $\mathrm{diag}\{\boldsymbol{v}\}$ be a diagonal matrix constructed from the vector $\boldsymbol{v}$, and let $\circ$ and $\oplus$ be the Hadamard product and the direct matrix sum, respectively. Denote the gradient vector of the population MgDWD loss function (8) as

$$\mathcal{S}(\boldsymbol{\vartheta}) = \mathbb{E}\big(\{\mathbb{I}(\mathscr{Y}) \circ \phi_q'(\boldsymbol{\Theta}^\top\boldsymbol{Z})\} \otimes \boldsymbol{Z}\big) = \mathrm{vec}\big(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_K\big),$$

with

$$\boldsymbol{S}_k = \mathbb{E}\{\mathbb{1}\{y = k\}\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)\boldsymbol{Z}\} = \pi_k \int_{\mathscr{X}} \phi_q'(\boldsymbol{Z}^\top\theta_k)\boldsymbol{Z} g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X},$$

and its Hessian matrix as

$$\mathcal{H}(\boldsymbol{\vartheta}) = \mathbb{E}\Big\{\mathrm{diag}\{\mathbb{I}(\mathscr{Y},\ \mathscr{X}) \circ \varphi_q''(\boldsymbol{\Theta}^\top\boldsymbol{Z})\} \otimes (\boldsymbol{Z}\boldsymbol{Z}^\top)\Big\} = \bigoplus_{k=1}^{K}\boldsymbol{H}_k,$$

where φ_q'' denotes the second derivative of the function φ_q; $\mathbb{I}(\mathscr{Y},\ \mathscr{X}) = \mathbb{I}(\mathscr{Y}) \circ \mathbb{I}(\mathscr{X})$ is a random vector with $\mathbb{I}(\mathscr{X}) = (\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_1\}, \ldots, \mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k\})^\top$ and $\mathscr{X}_k = \left\{\boldsymbol{X} \in \mathscr{X} \middle| \boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\right\}$; and

$$\boldsymbol{H}_k = \mathbb{E}\{\mathbb{1}\{y = k,\ \boldsymbol{X} \in \mathscr{X}_k\}\varphi_q''(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)\boldsymbol{Z}\boldsymbol{Z}^\top\} = \pi_k \int_{\mathscr{X}_k} \varphi_q''(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)\boldsymbol{Z}\boldsymbol{Z}^\top g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X}.$$

The block structure of $\mathcal{H}(\boldsymbol{\vartheta})$ implies a parallel relationship between each category. The relationship between the $\boldsymbol{\theta}_k$ is reflected by the sum-to-zero constraint in the definition of $\mathscr{C}$.

We assume the following regularity conditions.

(C1) The densities of $\boldsymbol{X}$ given $y = k \in \mathscr{Y}$, i.e., the $g_k(\boldsymbol{X})$, are continuous and have finite second moments.

(C2) $0 < \Pr\{\boldsymbol{X} \in \mathscr{X}_k^* | y = k\} < 1$ for all $k \in \mathscr{Y}$, where $\mathscr{X}_k^* = \{\boldsymbol{X} \in \mathscr{X} \,|\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* > Q\}$.

(C3) $\mathrm{Var}\{\boldsymbol{X} \mid \boldsymbol{X} \in \mathscr{X}_k^*,\ y = k\} \succ \mathbf{O}_p$ for all $k \in \mathscr{Y}$.

Remark 1. *Condition (C1) ensures that $\mathcal{L}$, $\mathcal{S}$ and $\mathcal{H}$ are well defined and continuous in $\boldsymbol{\vartheta}$. For the theoretically optimal hyperplane $\{\boldsymbol{X} \in \mathscr{X} | \boldsymbol{Z}^\top \boldsymbol{\theta}_k^* = 0\}$, the case with $\boldsymbol{\theta}_k^* = \mathbf{0}_{p+1}$ leaves $\mathscr{X}$ useless for classification. On the other hand, when $a_k^* \neq 0$ and $\boldsymbol{b}_k^* = \mathbf{0}_p$, the hyperplane is the empty set and is similarly meaningless. Condition (C2) is proposed to avoid the case where $\boldsymbol{b}_k^* = \mathbf{0}_p$ so that $\boldsymbol{\vartheta}^*$ always contains information relevant to the classification problem. For bounded random variables, condition (C2) should be assumed with caution. Condition (C3) implies the positive definiteness of $\mathcal{H}(\boldsymbol{\vartheta}^*)$.*

By convexity and the second-order Lagrange condition, the following theorem shows that the local minimizer of the population MgDWD problem exists and is unique.

Theorem 1. *Under the regularity conditions (C1)-(C3), the true parameter $\boldsymbol{\vartheta}^* \in \mathscr{C}$ is the unique minimizer of $\mathcal{L}(\boldsymbol{\vartheta})$ with $\boldsymbol{b}_k^* \neq \mathbf{0}_p$, and*

$$\mathcal{L}(\boldsymbol{\vartheta}^*) = \sum_{k=1}^{K} A(k,q)\pi_k,$$

with $0 \leq u(k,q) \leq A(k,q) \leq v(k,q) \leq 1$, where

$$A(k,q) = 1 - \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k^*\}\Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^*}\Big)^q\Big\} \,\Big|\, y = k\Big\},$$

$$u(k,q) = \Pr\big\{\boldsymbol{X} \notin \mathscr{X}_k^* \big| y = k\big\} + Q^{2q} \Pr\big\{Q < \boldsymbol{Z}^\top \boldsymbol{\theta}_k^* \leq Q^{-1} \big| y = k\big\},$$

$$v(k,q) = \Pr\big\{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* \leq 1 \,\big|\, y = k\big\} + \inf_{\epsilon > 0} \Big(\frac{Q}{1+\epsilon}\Big)^q \Pr\big\{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* > 1 + \epsilon \,\big|\, y = k\big\}.$$

The bounds in Theorem 1 show how q affects the loss function $\mathcal{L}(\boldsymbol{\vartheta}^*)$. The upper bound $v(k,q)$ is a decreasing function of q with

$$\lim_{q \to 0} v(k,q) = 1 \text{ and } \lim_{q \to \infty} v(k,q) = \Pr\big\{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* \leq 1 \,\big|\, y = k\big\}.$$

In the lower bound $u(k,q)$, the first term $\Pr\big\{\boldsymbol{X} \notin \mathscr{X}_k^* \big| y = k\big\}$ is an increasing function of q and the last term $Q^{2q} \Pr\big\{Q < \boldsymbol{Z}^\top \boldsymbol{\theta}_k^* \leq Q^{-1} \big| y = k\big\}$ is a decreasing function of q, with

$$\lim_{q \to 0} u(k,q) = 1 \text{ and } \lim_{q \to \infty} u(k,q) = \Pr\big\{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* \leq 1 \,\big|\, y = k\big\}.$$

Consequently, for the given population $\mathbb{P}(y, \boldsymbol{X})$, a larger q encourages the population MgDWD estimator to focus more on the regions $\{\boldsymbol{X} \notin \mathscr{X}_k,\ y = k\}$ that correspond to misclassifications. As a result, the estimator's performance will be similar to the hinge loss as $q \to \infty$. Setting q too small will lead to an ineffective classifier due to the unreasonable penalty placed on the well classified region $\{\boldsymbol{X} \in \mathscr{X}_k,\ y = k\}$. This variation in the lower bound with respect to q provides a necessary condition for the existence of an optimal q.

Remark 2. *The explicit relationship between q and $\boldsymbol{\vartheta}^*$ is complicated. While it may be more desirable to prove that a greater value of q results in a smaller value of the loss function $\mathcal{L}(\boldsymbol{\vartheta})$, there is no explicit formula for the optimal value $\boldsymbol{\vartheta}^*$ in terms of q.*

2.3. Estimator Consistency

Under the unpenalized framework presented in the previous subsection, all covariates will contribute to the classification task for each category: this scenario may lead to a classifier that overfits to the training data set. In this subsection, we study the consistency of the estimator for (5) in ultra-high dimensional settings.

To achieve structural sparsity in the estimator, the regularization parameter λ in (7) must be large enough to dominate the gradient of the empirical MgDWD loss evaluated at the theoretical minimizer $\boldsymbol{\vartheta}^* = \text{vec}\{\boldsymbol{\Theta}^*\}$ with high probability. On the other hand, λ should also be as small as possible to reduce the bias incurred by the SGL regularization term

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \tau \|\boldsymbol{\beta}_j\|_1 + (1-\tau)\|\boldsymbol{\beta}_j\|_2.$$

Lemma 2 provides a suitable choice of λ under the following assumption.

(A1) The predictors $\boldsymbol{X} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ are independent sub-Gaussian random vectors satisfying $\mathbb{E}\boldsymbol{X} = \mathbf{0}_p$, and where $\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, there exists a constant $\kappa > 0$ such that for any $\boldsymbol{\gamma} \in \mathbb{R}^p$, $\mathbb{E}\exp(\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}) \leq \exp(\|\boldsymbol{\gamma}\|_2^2 \kappa^2/2)$. From here on, we define ς_1^2 as the largest eigenvalue of $\boldsymbol{\Sigma}$.

Lemma 2. *Denote $\boldsymbol{S}(\boldsymbol{\vartheta}^*) = (\mathbf{I}_K \otimes \mathbf{Z}^\top)\text{diag}(\text{vec}\{\mathbf{E}\})\text{vec}\{\phi_q'(\mathbf{Z}\boldsymbol{\Theta}^*)\}$, where $\mathbf{E} = \big(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_N\big)^\top$, $\mathbf{Z} = \big(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N\big)^\top$ with $\boldsymbol{Z}_i = (1, \boldsymbol{X}_i^\top)^\top$, and $\mathbf{I}_K$ is the identity matrix of size K. Under condition (A1),*

$$\tilde{P}\big\{\mathbf{P}\boldsymbol{S}(\boldsymbol{\vartheta}^*)\big\} \leq \tau\Lambda_1 + (1-\tau)\Lambda_2$$

with probability at least $1 - 2(Kp)^{1-c_1^2} - p^{1-c_2^2}$, where

$$\mathbf{P} = (\mathbf{I}_K - K^{-1}\mathbf{1}_K\mathbf{1}_K^\top) \otimes \mathbf{I}_{p+1},$$
$$\Lambda_1 = \max\{\varsigma_1\kappa, 1\}\Big(1 - \frac{1}{K}\Big)\sqrt{\frac{2\log(pK)}{N}},$$
$$\Lambda_2 = \max\{2\sqrt{2}\varsigma_1\kappa, 1\}\left\{c_2\sqrt{\Big(1 - \frac{1}{K}\Big)\frac{2\log(p)}{N}} + \sqrt{\frac{K-1}{N}}\right\},$$

for constants $c_1, c_2 > 1$.

It is difficult to obtain a closed form for the conjugate of the SGL penalty, say, $\bar{P}(\boldsymbol{v}) = \sup_{\boldsymbol{u} \in \mathscr{C}\setminus\{0\}} \frac{\langle \boldsymbol{u}, \boldsymbol{v}\rangle}{P(\boldsymbol{u})}$. Instead, we use a regularized upper bound $\tilde{P}(\boldsymbol{v}) \geq \bar{P}(\boldsymbol{v})$. Based on Lemma 2, we propose a theoretical tuning parameter value

$$\lambda = c_0\sqrt{\frac{\log(pK)}{N}}, \tag{9}$$

where c_0 is some given constant satisfying $\lambda > \tau\Lambda_1 + (1-\tau)\Lambda_2$.

Before we can derive an error bound for the estimator in (5), we impose two additional assumptions.

(A2) For the true parameter value $\boldsymbol{\vartheta}^*$, there is a (s_e, s_g)-sparse structure in the coefficients $\mathbf{B}^*$ with element-wise and group-wise support sets

$$\mathscr{E} = \big\{(j,k) \in \{1,\ldots,p\} \times \{1,\ldots,K\} | b_{jk}^* \neq 0\big\} \text{ and } \mathscr{G} = \big\{j \in \{1,\ldots,p\} | \boldsymbol{\beta}_j^* \neq \mathbf{0}_K\big\}$$

with cardinality $|\mathscr{E}| = s_e$ and $|\mathscr{G}| = s_g$, respectively.

(A3) There exist some positive constants ς_2 and ς_3 such that

$$\varsigma_2^2 = \max_{\gamma \in \mathscr{V}} \frac{\|\mathrm{diag}\{\mathrm{vec}(\mathbf{E}^\top)\}(\mathbf{Z} \otimes \mathbf{I}_K)\gamma\|_2^2}{N\|\gamma\|_2^2} \text{ and } \varsigma_3^2 = \min_{\gamma \in \mathscr{U}} \frac{\gamma^\top \mathcal{H}(\boldsymbol{\vartheta}^*)\gamma}{\gamma^\top \gamma}$$

with $\mathscr{V} = \{\boldsymbol{v} \in \mathbb{R}^{K(p+1)} | 0 < \|\boldsymbol{v}\|_0 \leq s_e + K\}$ and

$$\mathscr{U} = \left\{ \boldsymbol{\delta} \in \mathbb{R}^{K(p+1)} \;\middle|\; \frac{\tau}{1-\tau}\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + \sum_{j \in \mathscr{G}_+} \|\boldsymbol{\delta}_j\|_2 \geq C_0 \Big(\frac{\tau}{1-\tau}\|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1 + \sum_{j \notin \mathscr{G}} \|\boldsymbol{\delta}_j\|_2 \Big) \right\},$$

where $C_0 = \frac{(c_0-1)}{(c_0+1)}$, $\mathscr{E}^c$ is the complement of $\mathscr{E}$, $\mathscr{E}_+ = \mathscr{E} \cup \{l = 1 + (k-1)(p+1) | k = 1, \ldots, K\}$, and $\mathscr{G}_+ = \mathscr{G} \cup \{0\}$.

Under the choice of λ given in (9), we show the L_2-consistency of the estimator in (5).

Theorem 2. *Suppose that conditions (A1)-(A3) hold. Then with* $\lambda = c_0\sqrt{\frac{\log(pK)}{N}}$ *in* (5)*, we have that*

$$\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*\|_2 \leq \left\{ C_1\sqrt{s_e + K} + C_2\sqrt{s_g + 1} \right\} \sqrt{\frac{\log(pK)}{N}}$$

with probability at least $1 - 2(Kp)^{2(s_e+K)(1-c_3^2)}$*, where* $C_1 = 2\varsigma_3^{-2}\{c_0\tau + (\sqrt{2} + 2c_3)\varsigma_2\}$ *and* $C_2 = 2\varsigma_3^{-2}c_0(1-\tau)$.

Remark 3. *The sub-Gaussian distribution assumption (A1) is common in high-dimensional scenarios. This assumption characterizes the tail behavior of a collection of random variables including Gaussian, Bernoulli, and bounded variables as special cases. Assumption (A2) describes structural sparsity at two levels. The element-wise size* $s_e < p$ *is the size of the underlying generative model, and the group-wise size* $s_g < pK$ *is the size of the signal covariate set. Both* s_e *and* s_g *are allowed to depend on the sample size* N*. As a result, the dimension* p *is allowed to increase with the sample size* N*. Assumption (A3) guarantees that eigenvalues are positive in this sparse scenario.*

Remark 4. *In practice, the tuning parameters* λ *and* τ *in* (7) *are commonly chosen by M-fold cross validation. That is, we choose the pair* (τ, λ) *with the highest prediction accuracy among the sub-data sets* $\mathcal{D}_m$*, specifically,*

$$CV(\tau, \lambda) = \sum_{m=1}^{M} \sum_{i \in \mathcal{D}_m}, \mathbb{1}\{y_i = \hat{y}_i(\tau, \lambda)\}$$

where $\hat{y}_i(\tau, \lambda) = \underset{k \in \mathscr{Y}}{\mathrm{argmax}}\, \boldsymbol{Z}_i^\top \hat{\boldsymbol{\theta}}_k(\tau, \lambda)$.

2.4. Computational Algorithm

In this section, we propose an efficient algorithm to solve problem (5). Our approach uses the proximal algorithm (see [29]) for solving high dimensional regularization problems. In two main steps, this approach obtains a solution to the constrained optimization problem by applying the proximal operator to the solution to the unconstrained problem.

Since regularization is not needed for the intercept terms $\boldsymbol{\alpha} = (a_1, \ldots, a_K)^\top$, it can be separated from the coefficients in $\mathbf{B}$. The empirical MgDWD loss of (8) is given by

$$L(\boldsymbol{\vartheta}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{E}_i^\top \phi_q(\boldsymbol{F}_i) = \frac{1}{N}\mathrm{tr}\left\{\mathbf{E}\phi_q(\mathbf{F}^\top)\right\} = \frac{1}{N}\mathrm{vec}\{\mathbf{E}^\top\}^\top \mathrm{vec}\{\phi_q(\mathbf{F}^\top)\}$$

where $\mathbf{F} = (f_{ik})_{N\times K} = \mathbf{Z\Theta} = \mathbf{1}_N\boldsymbol{\alpha}^\top + \mathbf{XB}$. Various properties of the loss function $L(\boldsymbol{\vartheta})$ follow below.

Lemma 3. *The loss function $L(\boldsymbol{\vartheta})$ has Lipschitz continuous partial derivatives. In particular, for $\boldsymbol{S}(\boldsymbol{\alpha}) = \dfrac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \dfrac{1}{N}\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}^\top\mathbf{1}_N$ and any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^K$, we have that*

$$\big\|\boldsymbol{S}(\boldsymbol{u}) - \boldsymbol{S}(\boldsymbol{v})\big\|_2 \le \sqrt{\frac{n_{\max}}{N}}\frac{(q+1)^2}{q}\|\boldsymbol{u}-\boldsymbol{v}\|_2,$$

where $n_{\max}$ is the largest group sample size. For $\boldsymbol{S}(\mathbf{B}) = \dfrac{\partial L(\boldsymbol{\theta})}{\partial \mathbf{B}} = \dfrac{1}{N}\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}^\top\mathbf{X}$ and any $\mathbf{U},\mathbf{V}\in\mathbb{R}^{p\times K}$, we have that

$$\big\|\mathrm{vec}\{\boldsymbol{S}(\mathbf{U}) - \boldsymbol{S}(\mathbf{V})\}\big\|_2 \le \frac{\max_k\|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2}{N}\frac{(q+1)^2}{q}\|\mathrm{vec}\{\mathbf{U}-\mathbf{V}\}\|_2,$$

where $\boldsymbol{e}_k$ is the k-th column of $\mathbf{E}$ and indicates the observations belonging to the k-th group.

Hence, following the majorization–minimization scheme, we can majorize the empirical MgDWD loss $L(\boldsymbol{\vartheta})$ by a quadratic function, that is,

$$\begin{aligned} L(\boldsymbol{\vartheta}) \le & L(\boldsymbol{\vartheta}_*) + \boldsymbol{S}(\boldsymbol{\alpha}_*)^\top(\boldsymbol{\alpha}-\boldsymbol{\alpha}_*) + \frac{L_{\boldsymbol{\alpha}}}{2}\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_*\|_2^2 \\ & + \mathrm{vec}\{\boldsymbol{S}(\mathbf{B})\}^\top\mathrm{vec}\{\mathbf{B}-\mathbf{B}_*\} + \frac{L_{\mathbf{B}_*}}{2}\|\mathrm{vec}\{\mathbf{B}-\mathbf{B}_*\}\|_2^2, \end{aligned}$$

for some $\boldsymbol{\vartheta}_* = \mathrm{vec}\{(\boldsymbol{\alpha}_*, \mathbf{B}_*^\top)^\top\}$, where $L_{\boldsymbol{\alpha}}$ and $L_{\mathbf{B}}$ denote the Lipschitz constants in Lemma 3. Instead of minimizing $L(\boldsymbol{\vartheta})$ directly, we apply gradient descent to minimize its surrogate upper bound function. The gradient descent updates are given by

$$\boldsymbol{\alpha}_* = \boldsymbol{\alpha} - \frac{q(q+1)^{-2}}{\sqrt{n_{\max}N}}\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}^\top\mathbf{1}_N, \tag{10}$$

$$\mathbf{B}_* = \mathbf{B} - \frac{q(q+1)^{-2}}{\max_k\|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2}\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}^\top\mathbf{X}. \tag{11}$$

Next, we address the problem's constraints and regularization simultaneously by applying the proximal operator. For $\boldsymbol{\alpha}_*$, it is clear that

$$\boldsymbol{\alpha}_{\mathrm{new}} = \underset{\boldsymbol{\alpha}^\top\mathbf{1}_K=0}{\mathrm{argmin}}\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_*\|_2^2 = \mathbf{P}_K\boldsymbol{\alpha}_*, \tag{12}$$

where $\mathbf{P}_K = \mathbf{I}_K - k^{-1}\mathbf{1}_K\mathbf{1}_K$. For $\mathbf{B}_* = (\boldsymbol{\beta}_{1*},\ldots,\boldsymbol{\beta}_{p*})^\top$, the minimization problem can be expressed as

$$\begin{aligned} \mathbf{B}_{\mathrm{new}} &= \underset{\mathbf{B}\mathbf{1}_K=\mathbf{0}_p}{\mathrm{argmin}}\frac{1}{2}\|\mathrm{vec}\{\mathbf{B}-\mathbf{B}_*\}\|_2^2 + \frac{\lambda_1}{L_{\mathbf{B}}}\|\mathrm{vec}\{\mathbf{B}\}\|_1 + \frac{\lambda_2}{L_{\mathbf{B}}}\|\mathrm{vec}\{\mathbf{B}\}\|_{1,2} \\ &= \underset{\mathbf{B}\mathbf{1}_K=\mathbf{0}_p}{\mathrm{argmin}}\sum_{j=1}^{p}\frac{1}{2}\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j*}\|_2^2 + \frac{\lambda_1}{L_{\mathbf{B}}}\|\boldsymbol{\beta}_j\|_1 + \frac{\lambda_2}{L_{\mathbf{B}}}\|\boldsymbol{\beta}_j\|_2, \end{aligned} \tag{13}$$

which implies that we can implement minimization for p groups in parallel. The following theorem provides the solution to (13).

Theorem 3. *Let $\rho_1,\ \rho_2 \geq 0$ and $\boldsymbol{\beta}_* \in \mathbb{R}^K$. Then the constrained regularization problem*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \frac{1}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_*\|_2^2 + \rho_1\|\boldsymbol{\beta}\|_1 + \rho_2\|\boldsymbol{\beta}\|_2$$
$$s.t.\ \ \boldsymbol{\beta}^\top \mathbf{1}_K = 0$$

has a solution of the form

$$\boldsymbol{\beta}^* = \Big\{1 - \frac{\rho_2}{\|\mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1\boldsymbol{u})\|_2}\Big\}_+ \mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1\boldsymbol{u}) \tag{14}$$

for some $\boldsymbol{u} \in \partial\|\boldsymbol{\beta}\|_1$.

In the special case with $\rho_2 = 0$, the constrained regularization problem in Theorem 3 reduces to the constrained lasso problem with solution $\tilde{\boldsymbol{\beta}}^* = \mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1\boldsymbol{u})$. Combined with (14), the proximal operator $\mathcal{U}$, given by

$$\boldsymbol{\beta}^* = \mathcal{U}(\tilde{\boldsymbol{\beta}}^*, \rho_2) = \Big\{1 - \frac{\rho_2}{\|\tilde{\boldsymbol{\beta}}^*\|_2}\Big\}_+ \tilde{\boldsymbol{\beta}}^*, \tag{15}$$

can be introduced to realize the group sparsity of $\tilde{\boldsymbol{\beta}}^*$.

For the standard lasso problem, the subgradient $\boldsymbol{u}$ has a closed form given by $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{\beta}_* - \rho_1\boldsymbol{u} = \mathcal{S}(\boldsymbol{\beta}_*, \rho_1)$, with $\mathcal{S}(u, v) = \text{sign}(u)(|u| - v)_+$. However, under the constraint on $\tilde{\boldsymbol{\beta}}^*$, the naive solution $\mathbf{P}_K\mathcal{S}(\boldsymbol{\beta}_*, \rho_1)$ is misleading in that it satisfies the constraint but does not achieve shrinkage, let alone loss function minimization. The term $\mathbf{P}_K\boldsymbol{u}$ is suggestive of the intersection between the subdifferential set $\partial\|\boldsymbol{\beta}\|_1$ and the constraint set $\{\boldsymbol{\beta} \in \mathbb{R}^K | \boldsymbol{\beta}^\top \mathbf{1}_K = 0\}$; in this sense, $\tilde{\boldsymbol{\beta}}^*$ might not have a closed form. Here we consider using coordinate descent to solve the constrained lasso problem. For some fixed coordinate m, since $\boldsymbol{\beta}^\top \mathbf{1}_K = 0$, we have that $b_m = -\sum_{l\neq m} b_l$. Rewriting the objective function of the lasso-constrained problem in a coordinate-wise form, we obtain

$$\begin{aligned}\sum_{l=1}^K \frac{1}{2}(b_l - b_{l*})^2 + \rho_1|b_l| =& \Big(b_k - \frac{(b_{k*} - b_{m*})}{2} + \frac{1}{2}\sum_{l\neq k,m}^K b_l\Big)^2 + \rho_1\Big\{|b_k| + \Big|b_k + \sum_{l\neq k,m}^K b_l\Big|\Big\} \\ &+ \frac{1}{4}\Big(b_{k*} + b_{m*} + \sum_{l\neq k,m}^K b_l\Big)^2 + \sum_{l\neq k,m}^K \frac{1}{2}(b_l - b_{l*})^2 + \rho_1|b_l|.\end{aligned} \tag{16}$$

Next, Theorem 4 provides the solution to the optimization problem (16).

Theorem 4. *Suppose that $t,\ s \in \mathbb{R}$ and $\varrho \geq 0$. Then the regularization problem*

$$\min_{b\in\mathbb{R}} \frac{1}{2}(b-t)^2 + \varrho\{|b| + |b+s|\}$$

has solution

$$\begin{aligned} b^* &= \begin{cases} t, & |t| < C(s,t) \\ -C(s,t), & C(s,t) \leq |t| \leq C(s,t) + 2\varrho \\ \text{sign}(t)(|t| - 2\varrho), & |t| > C(s,t) + 2\varrho \end{cases} \\ &= t - \mathcal{S}\big(t, C(s,t)\big) + \mathcal{S}\big\{\mathcal{S}\big(t, C(s,t)\big), 2\varrho\big\}, \end{aligned}$$

where $C(s,t) = \dfrac{1 - \text{sign}(s)\text{sign}(t)}{2}|s|$.

By Theorem 4, given some $m \in \{1, \ldots, K\}$, the coordinate-wise minimizer for any $k \neq m$ can be expressed as the proximal operator

$$b_k = \mathcal{T}(t, s, \rho_1) = t - \mathcal{S}\big(t, C(s,t)\big) + \mathcal{S}\big\{\mathcal{S}\big(t, C(s,t)\big), \rho_1\big\}, \tag{17}$$

with $s = \sum_{l \neq k,m} b_l$ and $t = (b_{k*} - b_{m*} - s)/2$. If we fix m during iteration, then the shrinkage of b_m will be indirectly reflected in the other b_k. We propose that m change with k in the coordinate-wise minimization process to ensure that every coordinate can be equally shrunk. We summarize our proposed algorithm in Algorithm 1.

Algorithm 1 Proximal gradient descent algorithm for SGL-MgDWD.

Input: λ_1, λ_2.
Initialization: $\boldsymbol{\alpha}^{(0)} = \mathbf{0}_K$, $\mathbf{B}^{(0)} = \mathbf{O}_{p\times K}$, $l = 0$.

repeat
  Update $\boldsymbol{\alpha}$ according to (10) and (12):

$$\boldsymbol{\alpha}^{(l+1)} = \mathbf{P}_K\{\boldsymbol{\alpha}^{(l)} - L_{\boldsymbol{\alpha}}^{-1} S(\boldsymbol{\alpha}^{(l)})\}.$$

  Update $\tilde{\mathbf{B}}$ according to (11):

$$\tilde{\mathbf{B}} = \mathbf{B}^{(l)} - L_{\mathbf{B}}^{-1} S(\mathbf{B}^{(l)}).$$

  Set $\mathbf{B}^{(l+1)} \leftarrow \tilde{\mathbf{B}}$.
  repeat
    for $m = 1$ to K **do**
      for k in $\{1, \ldots, K\} \setminus m$ **do**
        Update $(\boldsymbol{t}, \boldsymbol{s})$:

$$\boldsymbol{t} = \tilde{\boldsymbol{b}}_k - \tilde{\boldsymbol{b}}_m, \quad \boldsymbol{s} = \sum_{r=1}^{K} \boldsymbol{b}_r^{(l+1)} - \boldsymbol{b}_k^{(l+1)} - \boldsymbol{b}_m^{(l+1)}.$$

        Update $\boldsymbol{b}_k^{(l+1)}$ according to (17) and $\boldsymbol{b}_m^{(l+1)}$ by constraint:

$$\boldsymbol{b}_k^{(l+1)} = \mathcal{T}(\boldsymbol{t}, \boldsymbol{s}, L_{\mathbf{B}}^{-1}\lambda_1), \quad \boldsymbol{b}_m^{(l+1)} = -\boldsymbol{s} - \boldsymbol{b}_k^{(l+1)}.$$

      end for
    end for
  until $\mathbf{B}^{(l+1)}$ convergence.
  Update $\mathbf{B}^{(l+1)}$ according to (15):

$$\mathbf{B}^{(l+1)} = \mathcal{U}(\mathbf{B}^{(l+1)}, L_{\mathbf{B}}^{-1}\lambda_2).$$

  Set $l \leftarrow l + 1$.
until some condition is met.
Output: $\boldsymbol{\alpha}^{(l)}$ and $\mathbf{B}^{(l)}$.

3. Numerical Analysis

In the following section, we use both simulated and real data sets to evaluate the finite sample properties of our proposed method. We compare the finite sample performance of SGL-MgDWD with L_1-regularized multinomial logistic regression (L_1-logistic).

3.1. Simulation Studies

The data is generated from the following model. Consider the K-category classification problem where $\pi_k = K^{-1}$ and $g_k(\boldsymbol{X})$ is the density function of a normal distribution with mean vector $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k}, \mathbf{0}_{p-2}^\top)^\top$ and covariance matrix $\mathbf{I}_p$, where $(\mu_{1k}, \mu_{2k}) = (2\cos(\pi r_k),\ 2\sin(\pi r_k))$ with $r_k = \frac{2(k-1)}{K}$, for $k = 1, \ldots, K$. In this model, only the first two variables contribute to the classification and their corresponding parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ form two groups of coefficients. The true model has the sparsity structure $(s_e, s_g) = (2K, 2)$ for a total of $K(p+1)$ coefficients. We set the sample size for each category to $n_k = 50, 100, 200$ and 400, and the number of classes to $K = 5$ and 11. We consider dimensionality $p = 100$ and 1000.

In what follows, we compare the proposed SGL-MgDWD method with the OVR method based on SGL-MgDWD with $K = 2$ (OVR-SGL-gDWD). For SGL-MgDWD, logistic regression and OVR, the tuning parameter λ is optimized over a discrete set by minimizing the prediction error using 5-fold cross validation. In each simulation, we conduct 100 runs and use a testing set of equal size to evaluate each method's performance using the following criteria:

- Testing set accuracy, measuring the rate of correct classification;
- Signal, as the average number of correctly-selected element-wise and group-wise signals, that is, with $\hat{b}_{jk} \neq 0$ and $\hat{\boldsymbol{\beta}}_j \neq \mathbf{0}$, respectively, denoted by the pair (s_e^+, s_g^+);
- Noise, as the average number of incorrectly-selected element-wise and group-wise components, that is, with $\hat{b}_{jk} = 0$ and $\hat{\boldsymbol{\beta}}_j = 0$, respectively, denoted by the pair (n_e^+, n_g^+).

Simulation results are summarized in Tables 1 and 2.

As shown in Tables 1 and 2, the proposed SGL-MgDWD method performs better than the L_1-logistic and OVR methods. Specifically, in each scenario, predictions from the SGL-MgDWD method had higher accuracy relative to the other two methods. Similarly, the SGL-MgDWD method correctly selected the signal components of the model with fewer incorrectly-selected noise components, again relative to the L_1-logistic and OVR methods. These simulation results also demonstrate that test accuracy increases with increasing sample size n_k and that test accuracy decreases with higher dimension p at fixed n_k. This is consistent with the derived theoretical properties. All computations were performed on a Tensorflow 2.3 CPU on Threadripper 2950X at 4.1 Ghz.

3.2. HIV Data Analysis

Symptomatic distal sensory polyneuropathy (sDSP) is a common debilitating condition among people with HIV. This condition leads to neuropathic pain and is associated with substantial comorbidities and increased health care costs. Plasma miRNA profiles show differences between HIV patients with and without sDSP, and several miRNA biomarkers are reported to be associated with the presence of sDSP in HIV patients (see [30]). The corresponding binary classification problem was analyzed in [30] using random forest classifiers. However, the HIV data set can be further classified into four classes. The HIV data set has 1715 miRNA measures for 40 patients and is partitioned into four groups ($K = 4$) with $n_k = 10$ patients each category: non-HIV, HIV with no brain damage (HIVNBD), HIV with brain damage but stable (HIVBDS) and HIV with brain damage and unstable (HIVBDU). In the following analysis, we apply our proposed method to this classification problem. The primary aim was to identify critical miRNA biomarkers for each of the four groups. Beyond achieving a finer classification, this analysis is helpful in assessing related pathogenic effects for each patient group.

Given the small sample size of $N = 40$, we chose the tuning parameter λ by maximizing leave-one-out cross validation accuracy. We fixed $(q, \tau) = (1, 0.1)$. Table 3 shows the signal for

coefficient estimates obtained from the SGL-MgDWD method using the selected λ. We conclude that there are 22 critical miRNA biomarkers important to the classification problem. In particular, the biomarkers miR-25-star, miR-3171, miR-3924 and miR-4307 are not relevant to the non-HIV group; miR-4641, miR-4655-3p and miR-660 are not relevant to the HIVNBD group; miR-217 and miR-4683 are not relevant to the HIVBDS group; and miR-217 and miR-4307 are not relevant to the HIVBDU group.

Table 1. Simulation results for the SGL-MgDWD, L_1-logistic, and OVR methods with $K = 5$. Time is measured relative to a baseline logistic regression model with $K = 5$, $p = 100$, and $N = 50$. Numbers in parentheses denote standard deviations.

n_k	p	Method	Test Accuracy	Signal (s_e^+, s_g^+)	Noise (n_e^+, n_g^+)	Time (SD)
50	100	SGL-MgDWD	0.980	(9.99, 2)	(0, 0)	1.150 (0.173)
		L_1-logistic	0.979	(9.00, 2)	(116.98, 26.17)	1.000 (0.153)
		OVR-SGL-gDWD	0.912	-	-	-
	1000	SGL-MgDWD	0.979	(10, 2)	(6.96, 1.94)	5.290 (0.166)
		L_1-logistic	0.966	(10, 2)	(2793.65, 722.38)	5.130 (0.063)
		OVR-SGL-gDWD	0.740	-	-	-
100	100	SGL-MgDWD	0.981	(10, 2)	(0.07, 0.03)	1.453 (0.155)
		L_1-logistic	0.980	(8.82, 2)	(35.18, 3.98)	1.258 (0.127)
		OVR-SGL-gDWD	0.828	-	-	-
	1000	SGL-MgDWD	0.980	(10, 2)	(1.01, 0.25)	4.863 (0.150)
		L_1-logistic	0.978	(9.93, 2)	(1380.38, 192.37)	4.703 (0.061)
		OVR-SGL-gDWD	0.546	-	-	-
200	100	SGL-MgDWD	0.980	(10, 2)	(7.67, 2.08)	1.776 (0.164)
		L_1-logistic	0.980	(9.39, 2)	(13.1, 0.72)	1.709 (0.175)
		OVR-SGL-gDWD	0.934	-	-	-
	1000	SGL-MgDWD	0.982	(10, 2)	(1.09, 0.29)	8.641 (0.186)
		L_1-logistic	0.981	(9.79, 2)	(199.02, 2.51)	2.505 (0.121)
		OVR-SGL-gDWD	0.950	-	-	-
400	100	SGL-MgDWD	0.981	(10, 2)	(0.02, 0)	2.792 (0.159)
		L_1-logistic	0.981	(10, 2)	(4.72, 3.95)	2.828 (0.115)
		OVR-SGL-gDWD	0.979	-	-	-
	1000	SGL-MgDWD	0.981	(10, 2)	(4.72, 3.95)	15.800 (0.221)
		L_1-logistic	0.981	(9.6, 2)	(16.17, 0.02)	17.915 (1.585)
		OVR-SGL-gDWD	0.964	-	-	-

Table 2. Simulation results for the SGL-MgDWD, L_1-logistic, and OVR methods with $K = 11$. Time is measured relative to a baseline logistic regression model with $K = 5$, $p = 100$, and $N = 50$. Numbers in parentheses denote standard deviations.

n_k	p	Method	Test Accuracy	Signal (s_e^+, s_g^+)	Noise (n_e^+, n_g^+)	Time (SD)
50	100	SGL-MgDWD	0.735	(21.41, 2)	(0.14, 0.02)	1.661 (0.143)
		L_1-logistic	0.735	(20.13, 2)	(337.77, 22.07)	1.610 (0.110)
		OVR-SGL-gDWD	0.647	-	-	-
	1000	SGL-MgDWD	0.733	(21.25, 2)	(0, 0)	7.105 (0.205)
		L_1-logistic	0.566	(20.67, 2)	(3805.97, 265.82)	6.933 (0.205)
		OVR-SGL-gDWD	0.382	-	-	-
100	100	SGL-MgDWD	0.737	(21.82, 2)	(0.06, 0.01)	2.518 (0.099)
		L_1-logistic	0.721	(20, 2)	(173.17, 5.81)	2.418 (0.103)
		OVR-SGL-gDWD	0.609	-	-	-
	1000	SGL-MgDWD	0.737	(21.88, 2)	(5.4, 0.77)	12.371 (0.109)
		L_1-logistic	0.697	(20.15, 2)	(1859.51, 9.04)	12.279 (0.114)
		OVR-SGL-gDWD	0.214	-	-	-
200	100	SGL-MgDWD	0.738	(22, 2)	(0, 0)	5.191 (0.079)
		L_1-logistic	0.730	(20, 2)	(50.7, 0.08)	4.246 (0.100)
		OVR-SGL-gDWD	0.609	-	-	-
	1000	SGL-MgDWD	0.738	(21.98, 2)	(0.23, 0.04)	21.950 (0.241)
		L_1-logistic	0.730	(20, 2)	(523.08, 1.07)	22.158 (0.163)
		OVR-SGL-gDWD	0.490	-	-	-
400	100	SGL-MgDWD	0.740	(22, 2)	(0, 0)	7.025 (0.172)
		L_1-logistic	0.738	(20, 2)	(3.71, 3.48)	7.997 (0.122)
		OVR-SGL-gDWD	0.709	-	-	-
	1000	SGL-MgDWD	0.738	(22, 2)	(0.68, 0.11)	38.301 (0.200)
		L_1-logistic	0.734	(20, 2)	(38.84, 35.37)	41.059 (2.064)
		OVR-SGL-gDWD	0.556	-	-	-

Table 3. Signal for the coefficient estimates obtained from the SGL-MgDWD method with $(q, \tau) = (1, 0.1)$ for the HIV data set. The symbols "+" and "-" denote positive and negative coefficient estimates, respectively, while "0" denotes a zero coefficient (i.e., an irrelevant variable).

	Non-HIV	HIVNBD	HIVBDS	HIVBDU
interception	+	+	-	+
miR-255b	-	+	-	+
miR-217	+	-	0	0
miR-25-star	0	+	+	-
miR-3136-5p	-	-	+	-
miR-3152-3p	+	-	-	+
miR-3159	-	-	-	+
miR-3171	0	+	-	-

Table 3. *Cont.*

	Non-HIV	HIVNBD	HIVBDS	HIVBDU
miR-33b	-	-	-	+
miR-34c-3p	-	-	+	+
miR-3545-5p	-	+	-	+
miR-3654	-	-	-	+
miR-3924	0	-	+	-
miR-4307	0	-	+	0
miR-4474-5p	-	+	+	+
miR-4526	+	-	-	-
miR-4641	+	0	-	-
miR-4655-3p	+	0	-	-
miR-4680-5p	-	-	+	-
miR-4683	-	-	0	+
miR-589	-	+	+	-
miR-619	+	-	-	+
miR-660	+	0	-	+

Author Contributions: Conceptualization, L.K. and N.T.; Methodology, T.S., L.K. and N.T.; Formal Analysis, Y.L.; Data Curation, W.G.B., E.A. and C.P.; Writing—Review & Editing, Y.W., B.J. and L.K.; Supervision, B.J., L.K. and N.T. All authors have read and agreed to the published version of the manuscript.

Funding: A Canadian Institutes of Health Research Team Grant and Canadian HIV-Ageing Multidisciplinary Programmatic Strategy (CHAMPS) in NeuroHIV (Christopher Power) supported these studies. Bei Jiang and Linglong Kong were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Christopher Power and Linglong Kong were supported by Canada Research Chairs in Neurological Infection and Immunity and Statistical Learning, respectively. Niansheng Tang was supported by grants from the National Natural Science Foundation of China (grant number: 11671349) and the Key Projects of the National Natural Science Foundation of China (grant number: 11731011).

Acknowledgments: The authors are thankful for the invitation of the two guest editors, Farouk Nathoo and Ejaz Ahmed. This work has also benefited from two anonymous reviewers' constructive comments and valuable feedback. The authors also thank the great help of Matthew Pietrosanu with editing.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Appendix A.1. Proof of Lemma 1

Proof. For simplicity, we write $p_j = P(y = j|\boldsymbol{X})$ and $f_k = f_k(\boldsymbol{X})$. Using the Lagrange multiplier method, we define

$$L(\boldsymbol{F}) = \mathbb{E}\Big\{ \sum_{k=1}^{K} \mathbb{1}\{y = k\}\phi_q\{\boldsymbol{F}(\boldsymbol{X})\}\Big|\boldsymbol{X} = \boldsymbol{u}\Big\} + \mu \mathbf{1}_K^\top \boldsymbol{F}(\boldsymbol{X}) = \sum_{k=1}^{K} p_k \phi_q(f_k) + \mu f_k.$$

Then for each k,

$$\frac{\partial L(\boldsymbol{F})}{\partial f_k} = \phi_q'(f_k)p_k + \mu = 0 \tag{A1}$$

with

$$\phi_q'(f_j) = \begin{cases} -1, & f_k \le Q \\ -(Qf_k^{-1})^q, & f_k > Q. \end{cases}$$

Without loss of generality, assume that $p_1 > p_2 \ge p_3 \ge \cdots \ge p_{K-1} > p_K$. Note that $-1 \le \phi_q' < 0$, and so $p_j \ge -\phi_q'(f_k)p_k = \mu > 0$ and $\mu = p_k$ if and only if $f_k \le Q$.

If $\mu < p_K < p_k$, then $p_K \neq \mu$ when $f_K > Q$, which implies that $f_k > f_K > Q$ for all $1 \leq k \leq K$. Hence, substituting $\phi_q'(f_k) = -(Qf_k^{-1})^q$ into (A1) yields

$$f_k = Q\sqrt[q]{p_k\mu^{-1}} > Q > 0.$$

However, $\sum_{k=1}^K f_k > 0$, contradicting the sum-to-zero constraint. Therefore, $\mu = p_K < p_k$ for $k < K$ and the result follows. □

Appendix A.2. Proof of Theorem 1

Lemma A1. *Under (C1), $\mathcal{L}(\boldsymbol{\vartheta})$ exists, and it is convex on $\boldsymbol{\vartheta}$.*

Proof. The existence of $\mathcal{L}(\boldsymbol{\vartheta})$ will be satisfied if

$$\mathbb{E}_{\boldsymbol{X}|y}\{|\phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)| \mid y = k\} = \int_{\mathscr{X}} |\phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)| g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X} < \infty.$$

We divide $\mathscr{X}$ into two disjoint subsets. Defining $\mathscr{X}_k = \{\boldsymbol{X} \in \mathscr{X} \mid \boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\}$, it is clear that

$$\int_{\mathscr{X}_k} |\phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)| g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X} \leq (q+1)^{-1} \int_{\mathscr{X}_k} g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X} < \infty.$$

Note that $0 < \phi_q(u) < (1+q)^{-1} < 1$ when $u > Q$. On the other hand, for $\mathscr{X}_k^{\mathrm{c}} = \{\boldsymbol{X} \in \mathscr{X} \mid \boldsymbol{Z}^\top\boldsymbol{\theta}_k \leq Q\}$,

$$\int_{\mathscr{X}_k^{\mathrm{c}}} |\phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)| g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X} \leq |1 - a_k| + \sum_{j=1}^p b_{jk} \int_{\mathscr{X}} |x_j| g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X} < \infty,$$

if $\mathbb{E}_{\boldsymbol{X}|y}\{|x_j| \mid y = k\} < \infty$ for all $k \in \mathscr{Y}$. This completes the proof of the existence of $\mathcal{L}(\boldsymbol{\vartheta})$.

Recall that

$$\mathcal{L}(\boldsymbol{\vartheta}) = \sum_{k=1}^K \pi_k \int_{\mathscr{X}} \phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k) g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X},$$

where $\phi_q(u)$ is a convex function of u, so its composition with the affine mapping $u = \boldsymbol{Z}^\top\boldsymbol{\theta}_k$ is still convex in $\boldsymbol{\theta}_k$. Clearly, $g_k(\boldsymbol{X})$, $\pi_k > 0$, so the non-negatively-weighted integral and sum both preserve convexity. □

Lemma A2. *Existence of minimizers of $\mathcal{L}(\boldsymbol{\vartheta})$ on $\mathscr{C} = \{\boldsymbol{\vartheta} \in \mathbb{R}^{K(p+1)} \mid \mathbf{C}\boldsymbol{\vartheta} = \mathbf{0}_K\}$, where $\mathbf{C} = \mathbf{1}_K^\top \otimes \mathbf{I}_{p+1}$.*

Proof. By Jensen's inequality, for any $\boldsymbol{\vartheta} \in \mathscr{C}$, we have that

$$\mathcal{L}(\boldsymbol{\vartheta}) \geq \phi_q\Big(\sum_{k=1}^K \pi_k \mathbb{E}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k | y = k\}\Big).$$

Let $\boldsymbol{\mu} = \mathrm{vec}\{(\pi_k\mathbb{E}\{z_j | y = k\})_{jk}\}$, where $\|\boldsymbol{\mu}\|_2 \geq (\sum_{k=1}^K \pi_k^2)^{\frac{1}{2}} \geq K^{-\frac{1}{2}} > 0$. For some $C > 0$, we have that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\vartheta}) \geq & \phi_q(\boldsymbol{\mu}^\top \boldsymbol{\vartheta}) = \mathbb{1}\{\boldsymbol{\mu}^\top \boldsymbol{\vartheta} < Q\}(1 - \boldsymbol{\mu}^\top \boldsymbol{\vartheta}) + \mathbb{1}\{\boldsymbol{\mu}^\top \boldsymbol{\vartheta} \geq Q\}\varphi_q(\boldsymbol{\mu}^\top \boldsymbol{\vartheta}) \\
\geq & \mathbb{1}\{\boldsymbol{\mu}^\top \boldsymbol{\vartheta} < Q\}\big|1 - |\boldsymbol{\mu}^\top \boldsymbol{\vartheta}|\big| \\
= & \mathbb{1}\{\boldsymbol{\mu}^\top \boldsymbol{\vartheta} < -(C+1)\}(|\boldsymbol{\mu}^\top \boldsymbol{\vartheta}| - 1) + \mathbb{1}\{-(C+1) < \boldsymbol{\mu}^\top \boldsymbol{\vartheta} < -1\}(|\boldsymbol{\mu}^\top \boldsymbol{\vartheta}| - 1) \\
& + \mathbb{1}\{-1 < \boldsymbol{\mu}^\top \boldsymbol{\vartheta} < Q\}(1 - |\boldsymbol{\mu}^\top \boldsymbol{\vartheta}|) \\
> & \mathbb{1}\{\|\boldsymbol{\mu}\|_2 \|\boldsymbol{\vartheta}\|_2 > C + 1\}C \\
= & \mathbb{1}\Big\{\|\boldsymbol{\vartheta}\|_2 > \frac{C+1}{\|\boldsymbol{\mu}\|_2}\Big\}C.
\end{aligned}$$

Note that $1 - \boldsymbol{\mu}^\top \boldsymbol{\vartheta} > 1 - Q > 0$ when $\boldsymbol{\mu}^\top \boldsymbol{\vartheta} < Q$. By the Cauchy–Schwarz inequality, $-\boldsymbol{\mu}^\top \boldsymbol{\vartheta} = |\boldsymbol{\mu}^\top \boldsymbol{\vartheta}| \leq \|\boldsymbol{\mu}\|_2 \|\boldsymbol{\vartheta}\|_2$.

Hence, if $\|\boldsymbol{\vartheta}\|_2 > \dfrac{C+1}{\|\boldsymbol{\mu}\|_2} > 0$, then $\mathcal{L}(\boldsymbol{\vartheta}) > C > 0$. The contrapositive of this result implies the existence of a minimizer in the unconstrained problem. That is, the closed set $\big\{\boldsymbol{\vartheta} \in \mathscr{C} \;\big|\; \mathcal{L}(\boldsymbol{\vartheta}) \leq C\big\}$ is bounded for some large enough C. This guarantees the existence of a solution, as desired. □

Lemma A3. *Under (C1), $\mathcal{S}(\boldsymbol{\vartheta})$ exists and*

$$\frac{\partial \mathcal{L}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \mathcal{S}(\boldsymbol{\vartheta}).$$

Proof. The existence of $\mathcal{S}(\boldsymbol{\vartheta})$ will follow if

$$\int_{\mathscr{X}} |\phi_q'(\boldsymbol{Z}^\top \theta_k) z_j| \pi_k g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} \leq \pi_k \int_{\mathscr{X}} |z_j| g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} < \infty$$

for $j = 1, \ldots, p+1$. Note that $|\phi_q'(u)| \leq 1$ when $u > Q$.

For every $\theta_{kj} \in \mathbb{R}$, $\phi_q(\boldsymbol{Z}^\top \theta_k)$ is a Lebesgue integrable function of $\boldsymbol{X}$. For any $u \in \mathbb{R}$, $\phi_q'(u)$ exists and $|\phi_q'(u)| \leq 1$. Hence, by the Leibniz integral rule, we have that

$$\begin{aligned}
\frac{\partial}{\partial \theta_{jk}} \int_{\mathscr{X}} \phi_q(\boldsymbol{Z}^\top \theta_k) \pi_k g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} &= \int_{\mathscr{X}} \frac{\partial \phi_q(\boldsymbol{Z}^\top \theta_k)}{\partial \theta_{jk}} \pi_k g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} \\
&= \int_{\mathscr{X}} \phi_q'(\boldsymbol{Z}^\top \theta_k) z_j \pi_k g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X}
\end{aligned}$$

and for any $l \neq k$,

$$\frac{\partial}{\partial \theta_{jl}} \int_{\mathscr{X}} \phi_q(\boldsymbol{Z}^\top \theta_k) \pi_k g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} = 0,$$

which is sufficient to show that

$$\frac{\partial \mathcal{L}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \mathcal{S}(\boldsymbol{\vartheta}).$$

□

Lemma A4. *Suppose (C1) is satisfied. Then (C2) implies that $\boldsymbol{b}_k^* \neq \mathbf{0}$.*

Proof. We can rewrite $\phi_q(u)$ as

$$\begin{aligned}\phi_q(u) =&\mathbb{1}\{u \le Q\}(1-u) + \mathbb{1}\{u > Q\}(1-Q)\Big(\frac{Q}{u}\Big)^q \\ =&\Big\{-\mathbb{1}\{u \le Q\} - \mathbb{1}\{u > Q\}\Big(\frac{Q}{u}\Big)^{q+1}\Big\}u + \mathbb{1}\{u \le Q\} + \mathbb{1}\{u > Q\}\Big(\frac{Q}{u}\Big)^q \\ =&\phi_q'(u)u + \mathbb{1}\{u \le Q\} + \mathbb{1}\{u > Q\}\Big(\frac{Q}{u}\Big)^q.\end{aligned}$$

Then for any $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ and its corresponding $\mathscr{X}_k = \{\boldsymbol{X} \in \mathscr{X} | \boldsymbol{Z}^\top\boldsymbol{\gamma} > Q\}$, we have that

$$\begin{aligned}&\mathbb{E}\big\{\mathbb{1}\{y=k\}\phi_q(\boldsymbol{Z}^\top\boldsymbol{\gamma})\big\} \\ &\quad=\mathbb{E}\big\{\mathbb{1}\{y=k\}\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\gamma})\boldsymbol{Z}^\top\boldsymbol{\gamma}\big\} + \mathbb{E}\big\{\mathbb{1}\{y=k,\, \boldsymbol{Z}^\top\boldsymbol{\gamma} \le Q\}\big\} \\ &\qquad+\mathbb{E}\Big\{\mathbb{1}\{y=k,\, \boldsymbol{Z}^\top\boldsymbol{\gamma} > Q\}\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\gamma}}\Big)^q\Big\} \\ &\quad=\boldsymbol{S}_k^\top(\boldsymbol{\gamma})\boldsymbol{\gamma} + \Pr\{y=k,\, \boldsymbol{X} \notin \mathscr{X}_k\} + \mathbb{E}\Big\{\mathbb{1}\{y=k,\, \boldsymbol{X} \in \mathscr{X}_k\}\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\gamma}}\Big)^q\Big\} \\ &\quad=\boldsymbol{S}_k^\top(\boldsymbol{\gamma})\boldsymbol{\gamma} + \pi_k\Big(1 - \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k\}\Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\gamma}}\Big)^q\Big\}\Big|y=k\Big\}\Big).\end{aligned}$$

Let $\boldsymbol{\vartheta}^* \in \mathscr{C}$ be a local minimizer. It follows that $\mathbf{P}\mathcal{S}(\boldsymbol{\vartheta}^*) = \mathbf{0}$ and $\sum_{k=1}^K \boldsymbol{S}_k^\top(\boldsymbol{\theta}_k^*)\boldsymbol{\theta}_k^* = \mathcal{S}^\top(\boldsymbol{\vartheta}^*)\boldsymbol{\vartheta}^* = 0$ since $\boldsymbol{\vartheta}^* = \mathbf{P}\boldsymbol{\vartheta}^*$ and $\mathbf{P} = \big(\mathbf{I}_K - K^{-1}\mathbf{1}_K\mathbf{1}_K^\top\big) \otimes \mathbf{I}_{p+1}$. Therefore,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\vartheta}^*) =&\mathbb{E}\big\{\mathbb{1}\{y=k\}\phi_q(\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*)\big\} \\ =&\sum_{k=1}^K \pi_k\Big(1 - \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k^*\}\Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*}\Big)^q\Big\}\Big|y=k\Big\}\Big) \\ =&\sum_{k=1}^K \pi_k\Big(1 - \Pr\{\boldsymbol{X} \in \mathscr{X}_k^*|y=k\}\mathbb{E}\Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*}\Big)^q \,\Big|\, y=k,\, \boldsymbol{X} \in \mathscr{X}_k^*\Big\}\Big).\end{aligned} \tag{A2}$$

For any $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ and its corresponding $\mathscr{X}_k = \{\boldsymbol{X} \in \mathscr{X} | \boldsymbol{Z}^\top\boldsymbol{\gamma} > Q\}$, we always have that

$$0 < \mathbb{E}\Big\{\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\gamma}}\Big)^q \,\Big|\, y=k,\, \boldsymbol{X} \in \mathscr{X}_k\Big\} < 1.$$

If $\boldsymbol{\gamma} = \mathbf{0}_{p+1}$, then $\mathscr{X}_k = \varnothing$ so that $\Pr\{y=k,\, \boldsymbol{X} \notin \mathscr{X}_k\} = \pi_k$ and $\Pr\{y=k,\, \boldsymbol{X} \in \mathscr{X}_k\} = 0$. If $\gamma_1 \le Q$ and $\boldsymbol{\gamma}_{/1} = \mathbf{0}_p$, then $\mathscr{X}_k = \varnothing$, giving the same conclusions as the previous case. If $\gamma_1 > Q$ and $\boldsymbol{\gamma}_{/1} = \mathbf{0}_p$, then $\mathscr{X}_k = \mathscr{X}$ so that $\Pr\{y=k,\, \boldsymbol{X} \notin \mathscr{X}_k\} = 0$ and $\Pr\{y=k,\, \boldsymbol{X} \in \mathscr{X}_k\} = \pi_k$. Consequently, when $0 < \Pr\{\boldsymbol{X} \in \mathscr{X}_k|y=k\} < 1$, then neither $\mathscr{X}_k$ nor $\mathscr{X}$ equal $\varnothing$, so $\boldsymbol{b}_k \neq 0$ follows.

Note that $\Pr\{\boldsymbol{X} \notin \mathscr{X}_k|y=k\} > 0$ implies that $\Pr\{0 < \boldsymbol{Z}^\top\boldsymbol{\gamma} \le Q|y=k\} > 0$ or $\Pr\{\boldsymbol{Z}^\top\boldsymbol{\gamma} \le 0|y=k\} > 0$, and so special attention should be paid to bounded random variables. □

Lemma A5. *Under (C1), $\mathcal{H}(\boldsymbol{\vartheta})$ exists and*

$$\frac{\partial^2\mathcal{L}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top} = \mathcal{H}(\boldsymbol{\vartheta}).$$

Furthermore, $\mathcal{H}(\boldsymbol{\vartheta}^*) \succ \mathbf{O}_{K(p+1)}$ when (C2) and (C3) hold.

Proof. The existence of $\mathcal{H}(\boldsymbol{\vartheta})$ follows if its all entries are absolutely integrable, that is, for any $j,k=1,\ldots,p+1$,

$$\begin{aligned}\int_{\mathscr{X}} &|\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\}\varphi_q''(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)z_jz_l|\pi_k g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X}\\ &\leq (q+q^{-1}+2)\int_{\mathscr{X}_k^{\mathrm{c}}}|z_jz_l|g_k(\boldsymbol{X})\mathrm{d}\boldsymbol{X}\\ &<\infty.\end{aligned}$$

Equivalently, the result follows if $\mathbb{E}_{\boldsymbol{X}|y}\{|z_jz_l| \mid y=k\}<\infty$ for all $k\in\mathscr{Y}$. Note that $0<\varphi_q''(u)\leq q+q^{-1}+2$ when $u>Q$.

Let η be a test function belonging to the Schwartz space $\mathscr{D}$. Then $\eta'\in\mathscr{D}$ with some support denoted by $\mathrm{supp}(\eta')$.

Clearly, $\phi_q'(u)$ is not differentiable at Q but is Lipschitz continuous. Therefore, the measurable function $\boldsymbol{S}_k(\boldsymbol{\theta}_k)$ is a locally integrable function of $\boldsymbol{\theta}_k$. Then the (regular) generalized functions $\boldsymbol{S}_k(\boldsymbol{\theta}_k)$ belong to the dual space of $\mathscr{D}$.

For the distributional derivative of $\boldsymbol{S}_k(\boldsymbol{\theta}_k)$ with respect to θ_{jk}, we have that

$$\begin{aligned}\left|\left\langle\frac{\partial\boldsymbol{S}_k(\boldsymbol{\theta}_k)}{\partial\theta_{jk}},\eta(\theta_{jk})\right\rangle\right| &= \left|-\left\langle\boldsymbol{S}_k(\boldsymbol{\theta}_k),\frac{\mathrm{d}\eta(\theta_{jk})}{\mathrm{d}\theta_{jk}}\right\rangle\right|\\ &\leq \int_{\mathbb{R}}\left|\boldsymbol{S}_k(\boldsymbol{\theta}_k)\eta'(\theta_{jk})\right|\mathrm{d}\theta_{jk}\\ &\leq \max_{\theta_{jk}\in\mathrm{supp}(\eta')}|\eta'(\theta_{jk})|\int_{\mathrm{supp}(\eta')}|\boldsymbol{S}_k(\boldsymbol{\theta}_k)|\mathrm{d}\theta_{jk}\\ &<\infty\end{aligned}$$

implying that the function $f(\theta_{jk},\boldsymbol{X}) = \phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)\boldsymbol{Z}\pi_k g_k(\boldsymbol{X})\eta'(\theta_{jk})$ is integrable on $\mathbb{R}\times\mathscr{X}$. Therefore, by Fubini's Theorem,

$$\begin{aligned}\left\langle\frac{\partial\boldsymbol{S}_k(\boldsymbol{\theta}_k)}{\partial\theta_{jk}},\eta(\theta_{jk})\right\rangle &= -\left\langle\boldsymbol{S}_k(\boldsymbol{\theta}_k),\frac{\mathrm{d}\eta(\theta_{jk})}{\mathrm{d}\theta_{jk}}\right\rangle\\ &= \int_{\mathscr{X}}-\left\langle\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)\boldsymbol{Z}\pi_k g_k(\boldsymbol{X}),\frac{\mathrm{d}\eta(\theta jk)}{\mathrm{d}\theta_{jk}}\right\rangle\mathrm{d}\boldsymbol{X}\\ &= \int_{\mathscr{X}}\left\langle\frac{\partial\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)}{\partial\theta_{jk}}\boldsymbol{Z}\pi_k g_k(\boldsymbol{X}),\eta(\theta_{jk})\right\rangle\mathrm{d}\boldsymbol{X}\\ &= \left\langle\mathbb{E}\left\{\frac{\partial\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)}{\partial\theta_{jk}}\boldsymbol{Z}\mathbb{1}\{y=k\}\right\},\eta(\theta_{jk})\right\rangle,\end{aligned}$$

which implies that

$$\frac{\partial\boldsymbol{S}_k(\boldsymbol{\theta}_k)}{\partial\theta_{jk}} = \mathbb{E}\left\{\frac{\partial\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)}{\partial\theta_{jk}}\boldsymbol{Z}\mathbb{1}\{y=k\}\right\}.$$

Recall that ϕ_q' can be written as

$$\phi_q'(u)=\varphi_q'(u)\mathbb{1}\{u>Q\}+(-1)\mathbb{1}\{u\leq Q\}=(\varphi_q'(u)+1)\mathbb{1}\{u>Q\}-1,$$

which contains a Schwartz product between the differentiable function $\varphi_q'(u)$ and the generalized function $\mathbb{1}\{u > Q\}$. Note that

$$\begin{aligned}\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\} =& \mathbb{1}\{z_j > 0,\ \theta_{jk} > c_{jk}\} + \mathbb{1}\{z_j \le 0,\ \theta_{jk} \le c_{jk}\}\\ =& (2\mathbb{1}\{z_j > 0\} - 1)\mathbb{1}\{\theta_{jk} > c_{jk}\} + (1 - \mathbb{1}\{z_j > 0\})\\ =& \operatorname{sign}(z_j)\mathbb{1}\{\theta_{jk} > c_{jk}\} + \mathbb{1}\{z_j \le 0\},\end{aligned}$$

where $c_{jk} = (Q - \sum_{l\neq j} z_l\theta_{lk})/z_j$ and

$$\begin{aligned}\frac{\partial\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\}}{\partial\theta_{jk}} + 0 &= \operatorname{sign}(z_j)\delta(\theta_{jk} - c_{jk})\\ &= \operatorname{sign}(z_j)|z_j|\delta(\boldsymbol{Z}^\top\boldsymbol{\theta}_k - Q)\\ &= z_j\delta(\boldsymbol{Z}^\top\boldsymbol{\theta}_k - Q),\end{aligned}$$

where $\delta(x)$ is the Dirac delta function and the distributional derivative of $\mathbb{1}\{x > 0\}$. Recall that $\delta(cx) = \delta(x)/|c|$ and $f(x)\delta(x - c) = f(c)\delta(x - c)$ for some constant c and function f.

Thus, by the product rule for the distributional derivative of the Schwartz product,

$$\begin{aligned}\frac{\partial\phi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)}{\partial\theta_{jk}} =& \frac{\partial(\varphi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k) + 1)}{\partial\theta_{jk}}\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\} + (\varphi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k) + 1)\frac{\partial\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\}}{\partial\theta_{jk}}\\ =& \varphi_q''(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)z_j\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\} + (\varphi_q'(\boldsymbol{Z}^\top\boldsymbol{\theta}_k) + 1)z_j\delta(\boldsymbol{Z}^\top\boldsymbol{\theta}_k - Q)\\ =& \varphi_q''(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)z_j\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\}.\end{aligned}$$

Substituting the above expression, we obtain

$$\frac{\partial\boldsymbol{S}_k(\boldsymbol{\theta}_k)}{\partial\theta_{jk}} = \mathbb{E}\left\{\varphi_q''(\boldsymbol{Z}^\top\boldsymbol{\theta}_k)\boldsymbol{Z}z_j\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k > Q\}\mathbb{1}\{y = k\}\right\}.$$

Similarly, for $l \neq k$, we have the distributional derivative

$$\frac{\partial\boldsymbol{S}_k(\boldsymbol{\theta}_k)}{\partial\theta_{jl}} = 0.$$

Recall that the distributional derivative does not depend on the order of differentiation and agrees with the classical derivative whenever the latter exists. To summarize, we have that

$$\boldsymbol{H}_k(\boldsymbol{\theta}_k) = \frac{\partial^2\mathcal{L}(\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_k^\top} = \frac{\partial\boldsymbol{S}_k(\boldsymbol{\theta}_k)}{\partial\boldsymbol{\theta}_k^\top}, \quad \mathcal{H}(\boldsymbol{\vartheta}) = \bigoplus_{k=1}^K \boldsymbol{H}_k(\boldsymbol{\theta}_k).$$

The $\boldsymbol{H}_k(\boldsymbol{\theta}_k)$ are symmetric matrices, so $\mathcal{H}(\boldsymbol{\vartheta})$ is also symmetric.

In the sense of generalized functions, differentiation is a continuous operation with respect to convergence in $\mathscr{D}'$. Therefore, $\phi_0' = \lim_{q\to 0}\phi_q' = -\mathbb{1}\{u \le 0\}$ and $\phi_0'' = \lim_{q\to 0}\phi_q'' = \delta(u)$; $\phi_\infty' = \lim_{q\to\infty}\phi_q' = -\mathbb{1}\{u \le 1\}$ and $\phi_\infty'' = \lim_{q\to\infty}\phi_q'' = \delta(u - 1)$, which coincides with results from the hinge loss.

Next, $\mathcal{H}(\boldsymbol{\vartheta}) \succ \mathbf{O}_{K(p+1)}$ if and only if both $\boldsymbol{H}_1(\boldsymbol{\theta}_1)$ and its Schur complement $\bigoplus_{k=2}^K \boldsymbol{H}_k(\boldsymbol{\theta}_k)$ are both symmetric and positive definite. We can deduce that $\mathcal{H}(\boldsymbol{\vartheta}) \succ \mathbf{O}_{K(p+1)}$ if and only if $\boldsymbol{H}_k(\boldsymbol{\theta}_k) \succ \mathbf{O}_{p+1}$ for all k.

Note that there exists $c > 0$ such that $\varphi_q''(\boldsymbol{Z}^\top \boldsymbol{\theta}_k) \geq c$ on $\mathscr{X}_k$. Then for any $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$,

$$\begin{aligned}\boldsymbol{\gamma}^\top \boldsymbol{H}_k(\boldsymbol{\theta}_k)\boldsymbol{\gamma} =& \pi_k \int_{\mathscr{X}_k} \varphi_q''(\boldsymbol{Z}^\top \boldsymbol{\theta}_k)(\boldsymbol{Z}^\top \boldsymbol{\gamma})^2 g_k(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} \\ \geq& c \Pr\{\boldsymbol{X} \in \mathscr{X}_k,\, y = k\} \mathbb{E}\{(\boldsymbol{Z}^\top \boldsymbol{\gamma})^2 | \boldsymbol{X} \in \mathscr{X}_k,\, y = k\} \\ \geq& c \Pr\{\boldsymbol{X} \in \mathscr{X}_k,\, y = k\} \big(\gamma_0^2 + \boldsymbol{\gamma}_1^\top \mathrm{Var}\{\boldsymbol{X} | \boldsymbol{X} \in \mathscr{X}_k,\, y = k\} \boldsymbol{\gamma}_1\big),\end{aligned}$$

which implies that $\boldsymbol{\gamma}^\top \boldsymbol{H}_k(\boldsymbol{\theta}_k)\boldsymbol{\gamma} = 0$ if and only if $\boldsymbol{\gamma} = \mathbf{0}_{p+1}$ when $\mathrm{Var}\{\boldsymbol{X} | \boldsymbol{X} \in \mathscr{X}_k,\, y = k\}$ is assumed to be non-singular. Assuming that $\mathrm{Var}\{\boldsymbol{X} | y = k\} \succ \mathbf{O}$ implies that $\mathrm{Var}\{\boldsymbol{X} | \boldsymbol{X} \in \mathscr{X}_k,\, y = k\} \succeq \mathbf{O}$. □

Proof of Theorem 1. By Lemma A2, a minimizer $\boldsymbol{\vartheta}^* \in \mathscr{C}$ exists with $\boldsymbol{b}_k^* \neq \mathbf{0}_p$ (by Lemma A4) and $\mathcal{H}(\boldsymbol{\vartheta}^*) \succ \mathbf{O}_{K(p+1)}$ (by Lemma A5). By the second-order Lagrange condition and the convexity of $\mathcal{L}(\boldsymbol{\vartheta})$ (by Lemma A1), a minimizer of the population MgDWD loss is unique.

Recall from (A2) that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\vartheta}^*) =& \mathbb{E}\{\mathbb{1}\{y = k\} \phi_q(\boldsymbol{Z}^\top \boldsymbol{\theta}_k^*)\} \\ =& \sum_{k=1}^K \pi_k \Big(1 - \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k^*\} \Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^*}\Big)^q\Big\} \Big| y = k\Big\}\Big) \\ =& \sum_{k=1}^K A(k,q)\pi_k.\end{aligned}$$

It follows that

$$\begin{aligned}0 \leq& \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k^*\} \Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^*}\Big)^q\Big\} \Big| y = k\Big\} \\ <& \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{Z}^\top \boldsymbol{\gamma} > 1 + q^{-1}\} + \mathbb{1}\{Q < \boldsymbol{Z}^\top \boldsymbol{\gamma} \leq 1 + q^{-1}\} \Big\{1 - \Big(\frac{Q}{1 + q^{-1}}\Big)^q\Big\} \Big| y = m\Big\} \\ =& \Pr\{\boldsymbol{Z}^\top \boldsymbol{\gamma} > Q | y = m\} - \Pr\{Q < \boldsymbol{Z}^\top \boldsymbol{\gamma} \leq Q^{-1} | y = m\} Q^{2q} \\ \leq& 1\end{aligned}$$

and

$$\begin{aligned}1 \geq& \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X} \in \mathscr{X}_k^*\} \Big\{1 - \Big(\frac{Q}{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^*}\Big)^q\Big\} \Big| y = k\Big\} \\ >& \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* > 1 + \epsilon\} \Big\{1 - \Big(\frac{Q}{1 + \epsilon}\Big)^q\Big\} \Big| y = k\Big\} \\ \geq& \sup_{\epsilon > 0} \Big\{1 - \Big(\frac{Q}{1 + \epsilon}\Big)^q\Big\} \Pr\{\boldsymbol{Z}^\top \boldsymbol{\theta}_k^* > 1 + \epsilon \mid y = m\} \\ \geq& 0.\end{aligned}$$

Consequently, $0 \leq u(k,q) \leq A(k,q) \leq v(k,q) \leq 1$.

Note that $\lim_{q \to \infty} (1 + \epsilon)^{-q} Q^q = \mathrm{e}^{-1}$ when $\epsilon = 0$ and $\lim_{q \to \infty} (1 + \epsilon)^{-q} Q^q = 0$ when $\epsilon > 0$. The difference between these two results is attributed to pointwise convergence.

Let $f_m = 1 - A(k,m) \in \mathscr{D}'$ with $m = 1, 2, \ldots$ and $\eta \in \mathscr{D}$. By Fubini's theorem and the dominated convergence theorem,

$$\begin{aligned}\lim_{m\to\infty}\langle f_m,\eta\rangle &= \lim_{m\to\infty}\Big\langle \mathbb{E}\Big\{\mathbb{1}\{\boldsymbol{X}\in\mathscr{X}_k^*\}\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*}\Big)^q\Big|y=k\Big\},\eta(\boldsymbol{\gamma})\Big\rangle\\ &= \lim_{m\to\infty}\mathbb{E}\Big\{\Big\langle\mathbb{1}\{\boldsymbol{X}\in\mathscr{X}_k^*\}\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*}\Big)^q,\eta(\boldsymbol{\theta}_k^*)\Big\rangle\Big|y=k\Big\}\\ &=\mathbb{E}\Big\{\lim_{m\to\infty}\Big\langle\mathbb{1}\{\boldsymbol{X}\in\mathscr{X}_k^*\}\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*}\Big)^q,\eta(\boldsymbol{\theta}_k^*)\Big\rangle\Big|y=k\Big\}\\ &=0=\langle 0,\eta(\boldsymbol{\theta}_k^*)\rangle.\end{aligned}$$

Similarly,

$$\begin{aligned}\lim_{m\to 0}\langle f_m,\eta\rangle &=\mathbb{E}\Big\{\lim_{m\to 0}\Big\langle\mathbb{1}\{\boldsymbol{X}\in\mathscr{X}_k^*\}\Big(\frac{Q}{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*}\Big)^q,\eta(\boldsymbol{\gamma})\Big\rangle\Big|y=k\Big\}\\ &=\mathbb{E}\Big\{\Big\langle\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*>0\},\eta(\boldsymbol{\gamma})\Big\rangle\Big|y=k\Big\}\\ &=\Big\langle\mathbb{E}\{\mathbb{1}\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*>0\}|y=k\},\eta(\boldsymbol{\theta}_k^*)\Big\rangle\\ &=\Big\langle\Pr\{\boldsymbol{Z}^\top\boldsymbol{\theta}_k^*>0\mid y=k\},\eta(\boldsymbol{\theta}_k^*)\Big\rangle,\end{aligned}$$

hence

$$A(k,\infty)=\lim_{q\to\infty}A(k,q)=\Pr\{\boldsymbol{X}\notin\mathscr{X}_k^*\mid y=k\},\text{ and }A(k,0)=\lim_{q\to 0}A(k,q)=1.$$

As a result, $A(k,\infty)$ coincides with the population hinge/SVM loss and $A(k,0)$ is independent of $\boldsymbol{\theta}_k^*$. □

Appendix A.3. Proof of Lemma 2

Proof. By the definition of $\tilde{P}$,

$$\tilde{P}\{\mathbf{P}\boldsymbol{S}(\boldsymbol{\vartheta}^*)\}=\tau\|\mathbf{P}\boldsymbol{S}(\boldsymbol{\vartheta}^*)\|_\infty+(1-\tau)\max_j\Big\{\|\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\alpha}^*)\|_2,\|\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\beta}_j^*)\|_2\Big\},$$

where

$$\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\alpha}^*)=\mathbf{P}_K\big(\mathbf{E}\circ\phi_q'\{\mathbf{F}(\boldsymbol{\vartheta}^*)\}\big)^\top\mathbf{1}_K=\frac{1}{N}\sum_{i=1}^N\mathbf{P}_K\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*),$$

$$\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\beta}_j^*)=\mathbf{P}_K\big(\mathbf{E}\circ\phi_q'\{\mathbf{F}(\boldsymbol{\vartheta}^*)\}\big)^\top\boldsymbol{x}_j=\frac{1}{N}\sum_{i=1}^N x_{ij}\mathbf{P}_K\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*),$$

$\mathbf{P}_K=(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K)$ with $\boldsymbol{p}_k=(p_{lk})=\mathbb{1}\{l=k\}-K^{-1}$, and

$$\mathbb{E}\{\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\alpha}^*)\}=\mathbf{P}_K\mathcal{S}(\boldsymbol{\alpha}^*)=\mathbf{0}_K,\quad\mathbb{E}\{\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\beta}_j^*)\}=\mathbf{P}_K\mathcal{S}(\boldsymbol{\beta}_j^*)=\mathbf{0}_K.$$

Denoting

$$d_{ik}=\{\boldsymbol{p}_k^\top\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*)\}=\sum_{l=1}^K\Big(\mathbb{1}\{y_i=k\}-\frac{1}{K}\Big)e_{il}\phi_q'(f_{il}^*),$$

we have that $|d_{ik}|\le 1-K^{-1}$. Note that the d_{ik} are N i.i.d. random variables with

$$\frac{1}{N}\sum_{i=1}^N\mathbb{E}(d_{ik})=\boldsymbol{p}_k^\top\mathcal{S}(\boldsymbol{\alpha}^*)=0\text{ and }\frac{1}{N}\sum_{i=1}^N\mathbb{E}(d_{ik}x_{ij})=\boldsymbol{p}_k^\top\mathcal{S}(\boldsymbol{\beta}_j^*)=0.$$

By Hoeffding's inequality, we have that

$$\Pr\left\{|\boldsymbol{p}_k^\top \boldsymbol{S}(\boldsymbol{\alpha}^*)| > c_1\Big(1-\frac{1}{K}\Big)\sqrt{\frac{2\log(pK)}{N}}\right\} \le 2(pK)^{-c_1^2}, \tag{A3}$$

where $c_1 > 1$.

Regarding the $d_{ik}x_{ij}$, we have that

$$\mathbb{E}\exp\{d_{ik}x_{ij}\} \le \mathbb{E}\exp\{(1-K^{-1})|x_{ij}|\} \le \exp\{4(1-K^{-1})^2\varsigma_1^2\kappa^2\},$$

which implies that the $d_{ik}x_{ij}$ are N independent sub-Gaussian random variables with variance proxy $(1-K^{-1})^2\varsigma_1^2\kappa^2$. Taking $c_1 > 1$, we have that

$$\Pr\left\{|\boldsymbol{p}_k^\top \boldsymbol{S}(\boldsymbol{\beta}_j^*)| > c_1\varsigma_1\kappa\Big(1-\frac{1}{K}\Big)\sqrt{\frac{2\log(pK)}{N}}\right\} \le 2(pK)^{-c_1^2}. \tag{A4}$$

Then by (A3) and (A4),

$$\Pr\left\{\max_j\{|\boldsymbol{p}_k^\top \boldsymbol{S}(\boldsymbol{\alpha}^*)|, |\boldsymbol{p}_k^\top \boldsymbol{S}(\boldsymbol{\beta}_j^*)|\} > \Lambda_1\right\} \le 2(pK)^{-c_1^2} \tag{A5}$$

with

$$\Lambda_1 = \max\{\varsigma_1\kappa, 1\}c_1\Big(1-\frac{1}{K}\Big)\sqrt{\frac{2\log(pK)}{N}}.$$

Taking a union bound over the Kp entries of $\mathbf{P}\boldsymbol{S}(\boldsymbol{\beta}^*)$ yields that

$$\begin{aligned}\Pr\{\|\mathbf{P}\boldsymbol{S}(\boldsymbol{\vartheta}^*)\|_\infty \ge \Lambda_1\} =& \Pr\left\{\max_{j,k}\left\{\Big|\frac{1}{N}\sum_{i=1}^N \boldsymbol{p}_k^\top \boldsymbol{S}(\boldsymbol{\alpha}^*)\Big|, \Big|\frac{1}{N}\sum_{i=1}^N \boldsymbol{p}_k^\top \boldsymbol{S}(\boldsymbol{\beta}_j^*)\Big|\right\} \ge \Lambda_1\right\}\\ \le& 2K(p+1)(Kp)^{-c_1^2}.\end{aligned}$$

On one hand,

$$\|\mathbf{P}\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*)\|_2^2 = \|(\boldsymbol{E}_i - K^{-1}) \circ \phi_q'(\boldsymbol{F}_i^*)\|_2^2 \le \sum_{l=1}^K (e_{il} - K^{-1})^2 \cdot 1 = 1 - K^{-1},$$

so for any $\boldsymbol{\gamma} \in \mathbb{R}^K$,

$$|\boldsymbol{\gamma}^\top \mathbf{P}\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*)| \le \|\boldsymbol{\gamma}\|_2\sqrt{1-\frac{1}{K}}$$

and $\mathbb{E}\{\boldsymbol{\gamma}^\top \mathbf{P}\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*)\} = 0$. Applying Hoeffding's lemma,

$$\mathbb{E}\exp\{\boldsymbol{\gamma}^\top \mathbf{P}_K \boldsymbol{S}(\boldsymbol{\alpha}^*)\} = \prod_{i=1}^N \mathbb{E}\exp\left\{\frac{1}{N}\boldsymbol{\gamma}^\top \mathbf{P}_K \mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*)\right\} \le \exp\left\{\frac{\|\boldsymbol{\gamma}\|_2^2}{2N}\Big(1-\frac{1}{K}\Big)\right\}.$$

Applying a square root to Theorem 2.1 of [31] with $c_2 > 1$, we have that

$$\Pr\left\{\|\mathbf{P}\boldsymbol{S}(\boldsymbol{\alpha}^*)\|_2 \ge \sqrt{\frac{K-1}{N}} + c_2\sqrt{\Big(1-\frac{1}{K}\Big)\frac{2\log(p)}{N}}\right\} \le p^{-c_2^2}. \tag{A6}$$

On the other hand, since the x_{ij} are N independent sub-Gaussian random variables with variance proxy $\varsigma_1^2\kappa^2$,

$$\begin{aligned}\mathbb{E}\exp\{\gamma^\top \mathbf{P}\boldsymbol{S}(\boldsymbol{\beta}_j^*)\} &= \prod_{i=1}^{N}\mathbb{E}\exp\left\{\frac{x_{ij}}{N}\{\gamma^\top \mathbf{P}\mathrm{diag}\{\boldsymbol{E}_i\}\phi_q'(\boldsymbol{F}_i^*)\}\right\}\\ &\leq \prod_{i=1}^{N}\mathbb{E}\exp\left\{\sqrt{1-\frac{1}{K}}\frac{\|\gamma\|_2}{N}|x_{ij}|\right\}\\ &\leq = \exp\left\{\frac{\|\gamma\|_2^2}{2}\Big(1-\frac{1}{K}\Big)\frac{8\varsigma_1^2\kappa^2}{N}\right\}\end{aligned}$$

and $\mathbb{E}\{\mathbf{P}_K\boldsymbol{S}(\boldsymbol{\beta}_j^*)\} = \mathbf{0}_K$. Similarly, we have that

$$\Pr\left\{\|\mathbf{P}\boldsymbol{S}(\boldsymbol{\beta}_j^*)\|_2 \geq 2\sqrt{2}\varsigma_1\kappa\left\{\sqrt{\frac{K-1}{N}}+c_2\sqrt{\Big(1-\frac{1}{K}\Big)\frac{2\log(p)}{N}}\right\}\right\} \leq p^{-c_2^2} \tag{A7}$$

for a constant $c_2 > 1$.

Therefore, by (A6) and (A7),

$$\Pr\left\{\max_j\left\{\|\mathbf{P}\boldsymbol{S}(\boldsymbol{\alpha}^*)\|_2, \|\mathbf{P}\boldsymbol{S}(\boldsymbol{\beta}_j^*)\|_2\right\} \geq \Lambda_2\right\} \leq p^{-c_2^2}$$

with

$$\Lambda_2 = \max\{2\sqrt{2}\varsigma_1\kappa, 1\}\left\{\sqrt{\frac{K-1}{N}}+c_2\sqrt{\Big(1-\frac{1}{K}\Big)\frac{2\log(p)}{N}}\right\}.$$

Applying the union bound to (A5), it follows that

$$\Pr\left\{\tilde{P}\{\mathbf{P}\boldsymbol{S}(\boldsymbol{\vartheta}^*)\} \geq \tau\Lambda_1 + (1-\tau)\Lambda_2\right\} \leq 2K(p+1)(pK)^{1-c_1^2} + p^{1-c_2^2},$$

and the desired result follows. □

Appendix A.4. Proof of Theorem 2

Lemma A6. *Suppose that* $\lambda = c_0\sqrt{\frac{\log(pK)}{N}}$. *Then* $\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^* \in \mathscr{U}$, *where*

$$\mathscr{U} = \left\{\boldsymbol{\delta} \in \mathbb{R}^{K(p+1)} \;\middle|\; \frac{\tau}{1-\tau}\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + \sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 \geq C_0\Big(\frac{\tau}{1-\tau}\|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1 + \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}_j\|_2\Big)\right\},$$

$C_0 = \frac{(c_0-1)}{(c_0+1)}$, $\mathscr{E}^c$ *denotes the complement of* $\mathscr{E}$, $\mathscr{E}_+ = \mathscr{E}\cup\{l = 1+(k-1)(p+1)|k = 1,\ldots,K\}$, *and* $\mathscr{G}_+ = \mathscr{G}\cup\{0\}$.

Proof. Since $\hat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}^* + \boldsymbol{\delta}$ is the minimizer, we have that

$$\begin{aligned}L(\boldsymbol{\vartheta}^*) + \lambda P(\boldsymbol{\beta}^*) &\geq L(\hat{\boldsymbol{\vartheta}}) + \lambda P(\hat{\boldsymbol{\beta}})\\ \lambda\big\{P(\boldsymbol{\beta}^*) - P(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}})\big\} &\geq L(\boldsymbol{\vartheta}^* + \boldsymbol{\delta}) - L(\boldsymbol{\vartheta}^*),\end{aligned} \tag{A8}$$

where $\boldsymbol{\beta}^*$ is the vector $\boldsymbol{\vartheta}^*$ without the a_k components, replacing $\tilde{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$. Then

$$\begin{aligned}
P(\boldsymbol{\beta}^*) - P(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}}) =& \tau\big(\|\boldsymbol{\beta}^*_{\mathscr{E}}\|_1 - \|\boldsymbol{\beta}^*_{\mathscr{E}} + \tilde{\boldsymbol{\delta}}_{\mathscr{E}}\|_1 - \|\tilde{\boldsymbol{\delta}}_{\mathscr{E}^c}\|_1\big) \\
&+ (1-\tau)\Big(\sum_{j\in\mathscr{G}}\|\boldsymbol{\beta}^*_j\|_2 - \sum_{j\in\mathscr{G}}\|\boldsymbol{\beta}^*_j + \boldsymbol{\delta}_j\|_2 - \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}^*_j\|_2\Big) \\
\le& \tau\big(\|\tilde{\boldsymbol{\delta}}_{\mathscr{E}}\|_1 - \|\tilde{\boldsymbol{\delta}}_{\mathscr{E}^c}\|_1\big) + (1-\tau)\Big(\sum_{j\in\mathscr{G}}\|\boldsymbol{\delta}_j\|_2 - \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}^*_j\|_2\Big) \\
\le& \tau\big(\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 - \|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1\big) + (1-\tau)\Big(\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 - \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}_j\|_2\Big).
\end{aligned}$$

By the convexity of L,

$$L(\boldsymbol{\vartheta}^* + \boldsymbol{\delta}) - L(\boldsymbol{\vartheta}^*) \ge \langle \boldsymbol{S}(\boldsymbol{\vartheta}^*), \boldsymbol{\delta}\rangle \ge -\bar{P}\{\mathbf{P}\boldsymbol{S}(\boldsymbol{\vartheta}^*)\}P(\boldsymbol{\delta}) \ge -\frac{\lambda}{c_0}P(\boldsymbol{\delta}).$$

Note that

$$P(\boldsymbol{\delta}) = \tau\big(\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + \|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1\big) + (1-\tau)\Big(\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 + \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}_j\|_2\Big).$$

Combining the above results, we have that

$$\begin{aligned}
\lambda\big\{P(\boldsymbol{\vartheta}^*) - P(\boldsymbol{\vartheta}^* + \boldsymbol{\delta})\big\} &\ge \big\{L(\boldsymbol{\vartheta}^* + \boldsymbol{\delta}) - L(\boldsymbol{\vartheta}^*)\big\} \\
(c+1)\tau\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + (1-\tau)\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 &\ge (c-1)\tau\|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1 + (1-\tau)\sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}_j\|_2 \\
\frac{\tau}{1-\tau}\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + \sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 &\ge C_0\Big(\frac{\tau}{1-\tau}\|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1 + \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}_j\|_2\Big).
\end{aligned}$$

□

Lemma A7. *Assume that conditions (A1)-(A3) are satisfied. Then*

$$\sup_{\boldsymbol{v}\in\mathscr{V}}\frac{|\Delta L(\boldsymbol{u},\boldsymbol{v}) - \mathbb{E}\{\Delta L(\boldsymbol{u},\boldsymbol{v})\}|}{\|\boldsymbol{v}\|_2} > \Lambda_3$$

with probability at most $2(Kp)^{2(s_e+K)(1-c_3^2)}$*, where*

$$\Lambda_3 = (1+\sqrt{2}c_3)\varsigma_2\sqrt{\frac{2(s_e+K)\log(pK)}{N}}$$

and $\Delta L(\boldsymbol{u},\boldsymbol{v}) = L(\boldsymbol{u}+\boldsymbol{v}) - L(\boldsymbol{u})$ *for any* $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^{K(p+1)}$ *and for some constant* $c_3 > 1$*.*

Proof. Given any $\boldsymbol{u}\in\mathbb{R}^{K(p+1)}$ and $\boldsymbol{v}\in\mathscr{V}$ with $\mathscr{V} = \big\{\boldsymbol{v}\in\mathbb{R}^{K(p+1)}|0 < \|\boldsymbol{v}\|_0 \le s_e + K\big\}$,

$$\begin{aligned}
\Delta L(\boldsymbol{u},\boldsymbol{v}) =& \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{E}_i^\top\Big(\phi_q\big\{(\mathbf{U}+\mathbf{V})^\top\mathbf{Z}_i\big\} - \phi_q\big\{\mathbf{U}^\top\mathbf{Z}_i\big\}\Big) \\
=& \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}e_{ik}\Big(\phi_q\big\{\boldsymbol{Z}_i^\top(\boldsymbol{u}_k+\boldsymbol{v}_k)\big\} - \phi_q\big\{\boldsymbol{Z}_i^\top(\boldsymbol{u}_k)\big\}\Big) \\
=& \frac{1}{N}\sum_{i=1}^{N}d_i(\boldsymbol{u},\boldsymbol{v}),
\end{aligned}$$

where $\boldsymbol{u} = \mathrm{vec}\{\mathbf{U}\}$, $\boldsymbol{v} = \mathrm{vec}\{\mathbf{V}\}$.

The bounded gradient implies the Lipschitz continuity of ϕ_q so that $|\phi_q(u+v)-\phi_q(u)| \leq |v|$. Since $e_{ik} \in \{0,1\}$, we have that

$$\begin{aligned}|d_i(\boldsymbol{u},\boldsymbol{v})| \leq & \sum_{k=1}^{K}\left|e_{ik}\{\phi_q\{\boldsymbol{Z}_i^\top(\boldsymbol{u}_k+\boldsymbol{v}_k)\}-\phi_q(\boldsymbol{Z}_i^\top\boldsymbol{u}_k)\}\right| \\ \leq & \sum_{k=1}^{K}\left|e_{ik}\boldsymbol{Z}_i^\top\boldsymbol{v}_k\right| \leq \boldsymbol{E}_i^\top \mathrm{vec}\{\mathbf{V}^\top\boldsymbol{Z}_i\} \\ = & \boldsymbol{v}^\top(\boldsymbol{Z}_i\otimes\mathbf{I}_K)\boldsymbol{E}_i.\end{aligned}$$

Note that

$$\sum_{i=1}^{N}\left(\boldsymbol{v}^\top(\boldsymbol{Z}_i\otimes\mathbf{I}_K)\boldsymbol{E}_i\right)^2 = \|\mathrm{diag}\{\mathrm{vec}\{\mathbf{E}^\top\}\}(\mathbf{Z}\otimes\mathbf{I}_K)\boldsymbol{v}\|_2^2.$$

By Hoeffding's inequality, we have that

$$\begin{aligned}&\Pr\left\{\left|\frac{1}{N}\sum_{i=1}^{N}d_i(\boldsymbol{u},\boldsymbol{v})-\mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^{N}d_i(\boldsymbol{u},\boldsymbol{v})\Big)\right|>t\right\} \\ \leq & 2\exp\left\{-\frac{2N^2t^2}{4\|\mathrm{diag}\{\mathrm{vec}\{\mathbf{E}^\top\}\}(\mathbf{Z}\otimes\mathbf{I}_K)\boldsymbol{v}\|_2^2}\right\} \\ \leq & 2\exp\left\{-\frac{Nt^2}{2\varsigma_2^2\|\boldsymbol{v}\|_2^2}\right\}.\end{aligned}$$

Thus $\Pr\{R(\boldsymbol{v})>\Lambda_3\} \leq 2(Kp)^{-(s_e+K)c_3^2}$ with

$$R(\boldsymbol{v}) = \frac{|\Delta L(\boldsymbol{u},\boldsymbol{v})-\mathbb{E}\{\Delta L(\boldsymbol{u},\boldsymbol{v})\}|}{\|\boldsymbol{v}\|_2} \text{ and } \Lambda_3 = c_3\varsigma_2\sqrt{\frac{2(s_e+K)\log(pK)}{N}}.$$

Next, we consider covering $\mathscr{V}$ with ϵ-balls such that for any $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ in the same ball, $\left|\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}-\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}\right| \leq \epsilon$, where ϵ is a small positive number. The number of ϵ-balls required to cover a m-dimensional unit ball is bounded by $(\frac{2}{\epsilon}+1)^m$. Then for those $\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}$, we require a covering number of at most $(3(Kp)/\epsilon)^{s_e+K}$. Let $\mathscr{N}$ denote such an ϵ-net. We have that

$$\Pr\left\{\sup_{\boldsymbol{v}\in\mathscr{N}}R(\boldsymbol{v})>\Lambda_3\right\} \leq \Big(\frac{3Kp}{\epsilon}\Big)^{s_e+K}2(Kp)^{-(s_e+K)c_3^2} = 2\Big\{\frac{3}{\epsilon}(Kp)^{1-c_3^2}\Big\}^{s_e+K}.$$

Furthermore, for any $\boldsymbol{v}_1,\boldsymbol{v}_2 \in \mathscr{V}$,

$$\begin{aligned}|R(\boldsymbol{v}_1)-R(\boldsymbol{v}_2)| \leq & \frac{2}{N}\Big\|\mathrm{diag}\{\mathrm{vec}\{\mathbf{E}^\top\}\}(\mathbf{Z}\otimes\mathbf{I}_K)\Big(\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}-\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}\Big)\Big\|_1 \\ \leq & \frac{2}{\sqrt{N}}\Big\|\mathrm{diag}\{\mathrm{vec}\{\mathbf{E}^\top\}\}(\mathbf{Z}\otimes\mathbf{I}_K)\Big(\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}-\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}\Big)\Big\|_2 \\ \leq & 2\varsigma_2\epsilon.\end{aligned}$$

Therefore $\sup_{\boldsymbol{v}\in\mathscr{V}}R(\boldsymbol{v}) \leq \sup_{\boldsymbol{v}\in\mathscr{N}}R(\boldsymbol{v})+2\varsigma_2\epsilon$. Taking $\epsilon = \sqrt{\frac{(s_e+K)\log(pK)}{2N}}$, we have that

$$\begin{aligned}\Pr\left\{\sup_{\boldsymbol{v}\in\mathscr{V}} R(\boldsymbol{v}) > \Lambda_3\right\} \leq& \Pr\left\{\sup_{\boldsymbol{v}\in\mathscr{N}} R(\boldsymbol{v}) > (c_3-1)\varsigma_1\sqrt{\frac{2(s_e+K)\log(pK)}{N}}\right\}\\ \leq& 2\left\{\sqrt{\frac{2N}{(s_e+K)\log(pK)}}3(Kp)^{1-(c_3-1)^2}\right\}^{s_e+K}\\ \leq& 2\left\{(Kp)^{2-(c_3-1)^2}\right\}^{s_e+K}.\end{aligned}$$

Setting $c_3 = 1+\sqrt{2}c_4$ and $c_4 > 1$, we obtain the desired result that

$$\Pr\left\{\sup_{\boldsymbol{v}\in\mathscr{V}} R(\boldsymbol{v}) > (1+\sqrt{2}c_4)\varsigma_2\sqrt{\frac{2(s_e+K)\log(pK)}{N}}\right\} \leq 2(Kp)^{2(s_e+K)(1-c_4^2)}.$$

□

Proof of Theorem 2. Consider a disjoint partition on the coordinate set $\boldsymbol{\delta} = \hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*$, that is, $\boldsymbol{\delta} = \sum_{m=1}^{M} \boldsymbol{v}_m$ with $\boldsymbol{v}_m \in \mathscr{V}$. Note that, each subvector $\boldsymbol{v}_m$ has at most $s_e + K$ non-zero coordinates. Denote $\boldsymbol{v}_0 = \mathbf{0}$ and $\boldsymbol{u}_m = \boldsymbol{\vartheta}^* + \sum_{l=0}^{m-1} \boldsymbol{v}_l$ so that $\boldsymbol{u}_1 = \boldsymbol{\vartheta}^*$ and $\boldsymbol{u}_M + \boldsymbol{v}_M = \boldsymbol{\vartheta}^* + \boldsymbol{\delta}$. We have the decomposition

$$\begin{aligned}\Delta L(\boldsymbol{\vartheta}^*, \boldsymbol{\delta}) =& L\Big(\boldsymbol{\vartheta}^* + \sum_{m=1}^{M} \boldsymbol{v}_m\Big) - L(\boldsymbol{\vartheta}^*) = \sum_{m=1}^{M} L\Big(\boldsymbol{\vartheta}^* + \sum_{l=0}^{m} \boldsymbol{v}_l\Big) - L\Big(\boldsymbol{\vartheta}^* + \sum_{l=0}^{m-1} \boldsymbol{v}_l\Big)\\ =& \sum_{m=1}^{M} L(\boldsymbol{u}_m + \boldsymbol{v}_m) - L(\boldsymbol{u}_m) = \sum_{m=1}^{M} \Delta L(\boldsymbol{u}_m, \boldsymbol{v}_m).\end{aligned}$$

By Lemma A7,

$$\sum_{m=1}^{M} \Delta L(\boldsymbol{u}_m, \boldsymbol{v}_m) \geq \sum_{m=1}^{M} \mathbb{E}\big\{\Delta L(\boldsymbol{u}_m, \boldsymbol{v}_m)\big\} - \Lambda_3\|\boldsymbol{v}_m\|_2 = \mathbb{E}\big\{\Delta L(\boldsymbol{\vartheta}^*, \boldsymbol{\delta})\big\} - \Lambda_3\|\boldsymbol{\delta}\|_2$$

with high probability. By Lemma A5, $\mathcal{L}$ is twice differentiable so that

$$\begin{aligned}\mathbb{E}\big\{\Delta L(\boldsymbol{\vartheta}^*, \boldsymbol{\delta})\big\} =& \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Big(\boldsymbol{E}_i^\top \phi_q\big\{\boldsymbol{F}_i(\boldsymbol{\vartheta}^* + \boldsymbol{\delta})\big\}\Big) - \mathbb{E}\Big(\boldsymbol{E}_i^\top \phi_q\big\{\boldsymbol{F}_i(\boldsymbol{\vartheta}^*)\big\}\Big)\\ =& \mathcal{L}(\boldsymbol{\vartheta}^* + \boldsymbol{\delta}) - \mathcal{L}(\boldsymbol{\vartheta}^*)\\ =& \mathcal{S}(\boldsymbol{\vartheta}^*)^\top \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top \mathcal{H}(\boldsymbol{\vartheta}^*)\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2^2)\\ \geq& 0 + \frac{\varsigma_3^2}{2}\|\boldsymbol{\delta}\|_2^2 + o(\|\boldsymbol{\delta}\|_2^2).\end{aligned}$$

Consequently, $\Delta L(\boldsymbol{\vartheta}^*, \boldsymbol{\delta})$ is bounded below by $\frac{\varsigma_3^2}{2}\|\boldsymbol{\delta}\|_2^2 - \Lambda_3\|\boldsymbol{\delta}\|_2$ with high probability. Note that

$$\begin{aligned}P(\boldsymbol{\beta}^*) - P(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}}) \leq& \tau\big(\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 - \|\boldsymbol{\delta}_{\mathscr{E}^c}\|_1\big) + (1-\tau)\Big(\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 - \sum_{j\notin\mathscr{G}}\|\boldsymbol{\delta}_j\|_2\Big)\\ \leq& \Big(\tau\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + (1-\tau)\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2\Big).\end{aligned}$$

From (A8),

$$\begin{aligned} L(\boldsymbol{\vartheta}^*) + \lambda P(\boldsymbol{\beta}^*) \geq & L(\hat{\boldsymbol{\vartheta}}) + \lambda P(\hat{\boldsymbol{\beta}}) \\ \lambda\{P(\boldsymbol{\beta}^*) - P(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}})\} \geq & L(\boldsymbol{\vartheta}^* + \boldsymbol{\delta}) - L(\boldsymbol{\vartheta}^*) \\ \lambda\Big(\tau\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + (1-\tau)\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2\Big) \geq & \frac{\varsigma_3^2}{2}\|\boldsymbol{\delta}\|_2^2 - \Lambda_3\|\boldsymbol{\delta}\|_2. \end{aligned}$$

Clearly, $\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 \leq \sqrt{s_e+K}\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_2 \leq \sqrt{s_e+K}\|\boldsymbol{\delta}\|_2$ and $\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2 \leq \sqrt{s_g+1}\|\boldsymbol{\delta}\|_2$. We conclude that

$$\begin{aligned} \frac{\varsigma_3^2}{2}\|\boldsymbol{\delta}\|_2^2 \leq & \lambda\Big(\tau\|\boldsymbol{\delta}_{\mathscr{E}_+}\|_1 + (1-\tau)\sum_{j\in\mathscr{G}_+}\|\boldsymbol{\delta}_j\|_2\Big) + \Lambda_3\|\boldsymbol{\delta}\|_2 \\ \|\boldsymbol{\delta}\|_2^2 \leq & 2\varsigma_3^{-2}\Big\{\lambda\Big(\tau\sqrt{s_e+K} + (1-\tau)\sqrt{s_g+1}\Big) + \Lambda_3\Big\}\|\boldsymbol{\delta}\|_2, \end{aligned}$$

after which the desired result follows from straightforward algebraic manipulation. □

Appendix A.5. Proof of Lemma 3

Proof. Since

$$\mathrm{vec}(\mathbf{F}^\top)^\top = \mathrm{vec}\big\{(\mathbf{1}_N\boldsymbol{\alpha}^\top + \mathbf{XB})^\top\big\}^\top = \boldsymbol{\alpha}^\top(\mathbf{1}_N^\top \otimes \mathbf{I}_K) + \mathrm{vec}(\mathbf{B}^\top)^\top(\mathbf{X}^\top \otimes \mathbf{I}_K),$$

we have that

$$\begin{cases} \dfrac{\partial \mathrm{vec}(\mathbf{F}^\top)^\top}{\partial \boldsymbol{\alpha}} = \dfrac{\partial \boldsymbol{\alpha}^\top(\mathbf{1}_N^\top \otimes \mathbf{I}_K)}{\partial \boldsymbol{\alpha}} = \dfrac{\partial \boldsymbol{\alpha}^\top}{\partial \boldsymbol{\alpha}}(\mathbf{1}_N^\top \otimes \mathbf{I}_K) = \mathbf{I}_K(\mathbf{1}_N^\top \otimes \mathbf{I}_K) = \mathbf{1}_N^\top \otimes \mathbf{I}_K \\ \dfrac{\partial \mathrm{vec}(\mathbf{F}^\top)^\top}{\partial \mathrm{vec}(\mathbf{B}^\top)} = \dfrac{\partial \mathrm{vec}(\mathbf{B}^\top)^\top(\mathbf{X}^\top \otimes \mathbf{I}_K)}{\partial \mathrm{vec}(\mathbf{B}^\top)} = \mathbf{I}_{pK}(\mathbf{X}^\top \otimes \mathbf{I}_K) = \mathbf{X}^\top \otimes \mathbf{I}_K. \end{cases}$$

The derivative with respect to $\boldsymbol{\alpha}$ is

$$\begin{aligned} N\boldsymbol{S}(\boldsymbol{\alpha}) = & N\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{\alpha}}\mathrm{vec}\{\mathbf{E}^\top\}^\top \mathrm{vec}\{\phi_q(\mathbf{F}^\top)\} \\ = & \frac{\partial \mathrm{vec}(\mathbf{F}^\top)^\top}{\partial \boldsymbol{\alpha}}\frac{\partial \phi_q\{\mathrm{vec}(\mathbf{F}^\top)^\top\}}{\partial \mathrm{vec}(\mathbf{F}^\top)}\mathrm{vec}\{\mathbf{E}^\top\} \\ = & (\mathbf{1}_N^\top \otimes \mathbf{I}_K)\mathrm{diag}\big(\mathrm{vec}\{\phi_q'(\mathbf{F}^\top)\}\big)\mathrm{vec}\{\mathbf{E}^\top\} \\ = & \mathrm{vec}\big(\mathbf{I}_K\left\{\mathbf{E}\circ\phi_q'(\mathbf{F})\right\}^\top \mathbf{1}_N\big) \\ = & \left\{\mathbf{E}\circ\phi_q'(\mathbf{F})\right\}^\top \mathbf{1}_N. \end{aligned}$$

Thus,

$$\begin{aligned} \big\|\boldsymbol{S}(\boldsymbol{\alpha})|_{\boldsymbol{v}}^{\boldsymbol{u}}\big\|_2^2 = & \big\|\boldsymbol{S}(\boldsymbol{u}) - \boldsymbol{S}(\boldsymbol{v})\big\|_2^2 = N^{-2}\big\|(\mathbf{1}_N^\top \otimes \mathbf{I}_K)\mathrm{vec}\big\{\big(\mathbf{E}\circ\phi_q'\{\mathbf{F}(\boldsymbol{\alpha})\}|_{\boldsymbol{v}}^{\boldsymbol{u}}\big)^\top\big\}\big\|_2^2 \\ \leq & N^{-2}\big\|\mathbf{1}_N^\top \otimes \mathbf{I}_K\big\|_2^2\big\|\mathrm{vec}\big\{\big(\mathbf{E}\circ\phi_q'\{\mathbf{F}(\boldsymbol{\alpha})\}|_{\boldsymbol{v}}^{\boldsymbol{u}}\big)^\top\big\}\big\|_2^2 \\ = & N^{-1}\sum_{k=1}^{K}\sum_{i=1}^{N} e_{ik}^2\big(\phi_q'\{f_{ik}(u_k)\} - \phi_q'\{f_{ik}(v_k)\}\big)^2 \\ \leq & N^{-1}\sum_{k=1}^{K}\Big(\sum_{i=1}^{N} e_{ik}\Big)L_q^2(u_k - v_k)^2 \\ \leq & N^{-1}n_{\max}L_q^2\|\boldsymbol{u} - \boldsymbol{v}\|_2^2, \end{aligned}$$

where $L_q = \frac{(q+1)^2}{q}$ is the Lipschitz constant of ϕ_q'. We have that $L_{\boldsymbol{\alpha}} = \sqrt{\frac{n_{\max}}{N}} L_q$.

The derivative with respect to $\mathrm{vec}(\mathbf{B}^\top)$ is

$$\begin{aligned} N\frac{\partial L(\boldsymbol{\theta})}{\partial \mathrm{vec}(\mathbf{B}^\top)} =& \frac{\partial}{\partial \mathrm{vec}(\mathbf{B}^\top)} \mathrm{vec}\{\mathbf{E}^\top\}^\top \mathrm{vec}\{\phi_q(\mathbf{F}^\top)\} \\ =& \frac{\partial \mathrm{vec}(\mathbf{F}^\top)^\top}{\partial \mathrm{vec}(\mathbf{B}^\top)} \frac{\partial \phi_q\{\mathrm{vec}(\mathbf{F}^\top)^\top\}}{\partial \mathrm{vec}(\mathbf{F}^\top)} \mathrm{vec}\{\mathbf{E}^\top\} \\ =& (\mathbf{X}^\top \otimes \mathbf{I}_K)\mathrm{diag}\big(\mathrm{vec}\{\phi_q'(\mathbf{F}^\top)\}\big)\mathrm{vec}\{\mathbf{E}^\top\} \\ =& \mathrm{vec}\Big(\mathbf{I}_K\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}^\top \mathbf{X}\Big) \\ =& \mathrm{vec}\Big(\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}^\top \mathbf{X}\Big). \end{aligned}$$

Therefore, the derivative with respect to $\mathbf{B}$ is $\boldsymbol{S}(\mathbf{B}) = N^{-1}\mathbf{X}^\top\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}$. Note that

$$\begin{aligned} \mathrm{vec}\Big(\mathbf{X}^\top\{\mathbf{E}\circ\phi_q'(\mathbf{F})\}\Big) =& (\mathbf{I}_K \otimes \mathbf{X}^\top)\mathrm{diag}\{\mathrm{vec}(\mathbf{E})\}\mathrm{vec}\{\phi_q'(\mathbf{F})\} \\ =& \left\{\bigoplus_{k=1}^K \mathbf{X}^\top \mathrm{diag}(\boldsymbol{e}_k)\right\}\mathrm{vec}\{\phi_q'(\mathbf{F})\} \end{aligned}$$

and

$$\sum_{i=1}^N \{e_{ik}\boldsymbol{X}_i^\top(\boldsymbol{u}_k - \boldsymbol{v}_k)\}^2 = \|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}(\boldsymbol{u}_k - \boldsymbol{v}_k)\|_2^2 \le \|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2\|\boldsymbol{u}_k - \boldsymbol{v}_k\|_2^2;$$

thus

$$\begin{aligned} N^2\big\|\mathrm{vec}\{\boldsymbol{S}(\mathbf{U}) - \boldsymbol{S}(\mathbf{V})\}\big\|_2^2 =& \sum_{k=1}^K \Big\|\mathbf{X}^\top \mathrm{diag}(\boldsymbol{e}_k)\phi_q\{\boldsymbol{f}_k(\boldsymbol{b}_k)\}\big|_{\boldsymbol{v}_k}^{\boldsymbol{u}_k}\Big\|_2^2 \\ \le& \sum_{k=1}^K \|\mathbf{X}^\top \mathrm{diag}(\boldsymbol{e}_k)\|_2^2 \|\mathrm{diag}(\boldsymbol{e}_k)\phi_q\{\boldsymbol{f}_k(\boldsymbol{b}_k)\}\big|_{\boldsymbol{v}_k}^{\boldsymbol{u}_k}\|_2^2 \\ \le& \sum_{k=1}^K \|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2 \sum_{i=1}^N e_{ik}\big(\phi_q\{\boldsymbol{f}_{ik}(\boldsymbol{u}_k)\} - \phi_q\{\boldsymbol{f}_{ik}(\boldsymbol{v}_k)\}\big)^2 \\ \le& L_q^2 \sum_{k=1}^K \|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2 \sum_{i=1}^N \{e_{ik}\boldsymbol{X}_i^\top(\boldsymbol{u}_k - \boldsymbol{v}_k)\}^2 \\ \le& L_q^2 \sum_{k=1}^K \|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^4 \|\boldsymbol{u}_k - \boldsymbol{v}_k\|_2^2 \\ \le& \max_k \Big\{\|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2\Big\}^2 \|\mathrm{vec}(\mathbf{U} - \mathbf{V})\|_2^2. \end{aligned}$$

We conclude that $L_{\mathbf{B}} = L_q N^{-1} \max_k \|\mathrm{diag}(\boldsymbol{e}_k)\mathbf{X}\|_2^2$. □

Appendix A.6. Proof of Theorem 3

Lemma A8. *The indicator function*

$$\delta_{\mathcal{R}}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in \mathcal{R} \\ \infty, & \text{if } \boldsymbol{x} \notin \mathcal{R}, \end{cases}$$

where $\mathcal{R} = \{\boldsymbol{x} \in \mathbb{R}^p \mid \mathbf{1}_p^\top \boldsymbol{x} = 0\}$, *has subdifferential*

$$\partial \delta_{\mathcal{R}}(\boldsymbol{x}) = \begin{cases} \{\boldsymbol{g} \in \mathbb{R}^p \mid \boldsymbol{g} = s\mathbf{1}_p, s \in \mathbb{R}\}, & \text{if } \boldsymbol{x} \in \mathcal{R} \\ \varnothing, & \text{if } \boldsymbol{x} \notin \mathcal{R}. \end{cases}$$

Proof. Suppose that $\boldsymbol{x} \in \mathcal{R}$. Then $\boldsymbol{g} \in \partial \delta_{\mathcal{R}}(\boldsymbol{x})$ if and only if both

$$\delta_{\mathcal{R}}(\boldsymbol{y}) \geq \delta_{\mathcal{R}}(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle \text{ for all } \boldsymbol{y} \in \mathcal{R} \text{ and}$$
$$\boldsymbol{\omega}^\top (\boldsymbol{y} - \boldsymbol{x}) \leq 0.$$

Let $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{x}$. Then $\boldsymbol{z} \in \mathcal{R}$ since $\mathbf{1}_p^\top (\boldsymbol{y} - \boldsymbol{x}) = 0$. Thus, $\boldsymbol{g}^\top \boldsymbol{z} \leq 0$. If $\boldsymbol{g}^\top \boldsymbol{z} = 0$, then $\boldsymbol{g} \in \{\boldsymbol{g} \in \mathbb{R}^p \mid \boldsymbol{g} = s\mathbf{1}_p,\ s \in \mathbb{R}\}$. If there exists $\boldsymbol{g} \in \partial \delta_{\mathcal{R}}(\boldsymbol{x})$ satisfying $\boldsymbol{g}^\top \boldsymbol{z} < 0$ for some $\boldsymbol{z} \in \mathcal{R}$, then $-\boldsymbol{z} \in \mathcal{R}$, so we must have that $\boldsymbol{g}^\top \boldsymbol{z} > 0$. This is a contradiction.

Now, for any $\boldsymbol{x} \notin \mathcal{R}$, we have that $\boldsymbol{g} \in \partial \delta_{\mathcal{R}}(\boldsymbol{x})$ if and only if both

$$\delta_{\mathcal{R}}(\boldsymbol{y}) \geq \delta_{\mathcal{R}}(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle \text{ for all } \boldsymbol{y} \in \mathcal{R} \text{ and}$$
$$\boldsymbol{\omega}^\top (\boldsymbol{x} - \boldsymbol{y}) \geq \infty.$$

For $\boldsymbol{x} \notin \mathcal{R}$ and $\boldsymbol{y} \in \mathcal{R}$, since $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y} \in \mathbb{R}^p$ and $\boldsymbol{g}^\top \boldsymbol{z} \geq \infty$, it must be that $\boldsymbol{g} \in \varnothing$. □

Proof of Theorem 3. It is sufficient to minimize the objective function

$$G(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2^2 + \rho_1 \|\boldsymbol{\beta}\|_1 + \rho_2 \|\boldsymbol{\beta}\|_2 + \delta_{\mathcal{R}}(\boldsymbol{\beta}),$$

where $\mathcal{R} = \{\boldsymbol{x} \in \mathbb{R}^K \mid \mathbf{1}_K^\top \boldsymbol{\beta} = 0\}$. Then the subdifferential of $G(\boldsymbol{\beta})$ is

$$\partial G(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}_* + \rho_1 \partial \|\boldsymbol{\beta}\|_1 + \rho_2 \partial \|\boldsymbol{\beta}\|_2 + \partial \delta_{\mathcal{R}}(\boldsymbol{\beta}).$$

For an optimal solution $\boldsymbol{\beta}^* \in \mathcal{R}$, we have that $\mathbf{0}_p \in \partial G(\boldsymbol{\beta}^*)$ if and only if there exist $\boldsymbol{u} \in \partial \|\boldsymbol{\beta}\|_1$, $\boldsymbol{v} \in \partial \|\boldsymbol{\beta}\|_2$ and $s \in \mathbb{R}$ such that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_* - \rho_1 \boldsymbol{u} - \rho_2 \boldsymbol{v} - s\mathbf{1}_p$. Since $\mathbf{1}^\top \boldsymbol{\beta}^* = 0$, we have that $s = p^{-1}\mathbf{1}_p^\top (\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u} - \rho_2 \boldsymbol{v})$, so

$$\boldsymbol{\beta}^* = \mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u} - \rho_2 \boldsymbol{v}).$$

If $\boldsymbol{\beta}^* = \mathbf{0}_p$, then $|u_j| < 1$ for $j = 1, \ldots, p$, $\|\boldsymbol{v}\|_2 \leq 1$ and

$$\|\mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u})\|_2 = \rho_2 \|\mathbf{P}_K \boldsymbol{v}\|_2 \leq \rho_2 \|\mathbf{P}_K\|_2 \|\boldsymbol{v}\|_2 = \rho_2 \|\boldsymbol{v}\|_2 \leq \rho_2;$$

If $\boldsymbol{\beta}^* \neq \mathbf{0}_K$, then $\boldsymbol{u} \in \partial \|\boldsymbol{x}\|_1$, $\boldsymbol{v} = \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_2}$ and

$$\boldsymbol{\beta}^* = \mathbf{P}_K\Big(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u} - \rho_2 \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_2}\Big)$$
$$\Big(1 + \frac{\rho_2}{\|\boldsymbol{\beta}^*\|_2}\Big)\boldsymbol{\beta}^* = \mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u}).$$

Note that $\boldsymbol{\beta}^* = \mathbf{P}_K \boldsymbol{\beta}^* \in \mathcal{R}$. Taking the norm of both sides, we see that

$$\Big(1 + \frac{\rho_2}{\|\boldsymbol{\beta}^*\|_2}\Big)\|\boldsymbol{\beta}^*\|_2 = \|\mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u})\|_2$$
$$\|\boldsymbol{\beta}^*\|_2 = \|\mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u})\|_2 - \rho_2 > 0.$$

Substituting this result back into the $\boldsymbol{\beta}^* \neq \mathbf{0}_K$ case, we have that

$$\boldsymbol{\beta}^* = \Big\{1 - \frac{\rho_2}{\|\mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u})\|_2}\Big\}\mathbf{P}_K(\boldsymbol{\beta}_* - \rho_1 \boldsymbol{u}).$$

Combining the above two cases gives the desired result. □

Appendix A.7. Proof of Theorem 4

Proof. Denote the objective function by

$$G(b) = \frac{1}{2}(b-t)^2 + \varrho\{|b| + |b+s|\}.$$

When $s = 0$, we obtain a lasso problem with

$$b^* = \underset{b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}(b-t)^2 + 2\varrho|x| = \mathcal{S}(t, 2\varrho).$$

When $s \neq 0$, the subdifferential of $G(b)$ is

$$\partial G(b) = b - t + \varrho\{\partial|x| + \partial|x+s|\}.$$

We see that $0 \in \partial G(b^*)$ if and only if there exist $u \in \partial|b|$ and $v \in \partial|b+s|$ with

$$b^* = b - \varrho(u+v).$$

If $b^* = 0$, then $|u| < 1$ and $v = \operatorname{sign}(s)$, hence

$$b^* = 0 \text{ if } |t - \varrho\operatorname{sign}(s)| \leq \varrho.$$

If $s > 0$, then $\operatorname{sign}(s) = 1$ and $0 \leq t \leq 2\varrho$. If $s < 0$, then $\operatorname{sign}(s) = -1$, and $-2\varrho \leq t \leq 0$. Note that if $t \neq 0$, then $\operatorname{sign}(s) = \operatorname{sign}(t)$ or $\operatorname{sign}(s)\operatorname{sign}(t) = 1$.

When $b^* = -s$, then $u = -\operatorname{sign}(s)$ and $|v| < 1$, hence

$$b^* = -s \text{ if } |t + s + \varrho\operatorname{sign}(s)| \leq \varrho.$$

If $s > 0$, then $\operatorname{sign}(s) = 1$ and $-(s+2\lambda) \leq t \leq -s < 0$. If $s < 0$, then $\operatorname{sign}(s) = -1$ and $0 < -s \leq t \leq -(s-2\lambda)$. Note that $\operatorname{sign}(s) = -\operatorname{sign}(t)$ is equivalent to $\operatorname{sign}(s)\operatorname{sign}(t) = -1$.

Let $C(s,t) = \dfrac{1 - \operatorname{sign}(s)\operatorname{sign}(t)}{2}|s| \geq 0$. We can summarize the two cases above as

$$b^* = -C(s,t) \text{ if } 0 \leq C(s,t) \leq |t| \leq C(s,t) + 2\varrho. \tag{A9}$$

If $b^* \neq 0, -s$, then $u = \operatorname{sign}(b^*)$ and $v = \operatorname{sign}(b^* + s)$, thus

$$b^* = t - \varrho\big\{\operatorname{sign}(b^*) + \operatorname{sign}(b^*+s)\big\}$$
$$b^* + s = t + s - \varrho\big\{\operatorname{sign}(b^*) + \operatorname{sign}(b^*+s)\big\}.$$

If $\operatorname{sign}(b^*) = -\operatorname{sign}(b^*+s) = 1$, then $b^*(b^*+s) < 0$ or $0 < t < -s$. Thus $b^* = t > 0$ if $0 < t < -s$. If $\operatorname{sign}(b^*) = -\operatorname{sign}(b^*+s) = -1$, then $b^*(b^*+s) < 0$ or $-s < t < 0$. Thus $b^* = t < 0$ if $-s < t < 0$. Rewriting the two cases above, we have that

$$b^* = t \quad \text{if } 0 < |t| < C(s,t). \tag{A10}$$

If $\text{sign}(b^*) = \text{sign}(b^* + s) = 1$, then

$$\begin{aligned} \min\{b^*, b^* + s\} &> 0 \\ t - 2\varrho + \frac{s - |s|}{2} &> 0 \\ \text{sign}(t)|t| > \text{sign}(t)\big(\frac{|s|}{2} + 2\varrho\big) - \frac{s}{2} &> 0. \end{aligned}$$

Note that $t > 0$ and $\text{sign}(x) = \text{sign}(t)$. If $\text{sign}(b^*) = \text{sign}(b^* + s) = -1$, then

$$\begin{aligned} \max\{b^*, b^* + s\} &< 0 \\ t + 2\varrho + \frac{s + |s|}{2} &> 0 \\ \text{sign}(t)|t| < \text{sign}(t)\big(\frac{|s|}{2} + 2\varrho\big) - \frac{s}{2} &< 0. \end{aligned}$$

Note that $t < 0$ and $\text{sign}(x) = \text{sign}(t)$. Rewriting the two cases above, we have that

$$b^* = t - 2\varrho\text{sign}(t) \text{ if } |t| > 2\varrho + C(s,t). \tag{A11}$$

Summarizing (A9)–(A11),

$$b^* = \begin{cases} t, & |t| < C(s,t), \\ -C(s,t), & C(s,t) \le |t| \le C(s,t) + 2\varrho, \\ \text{sign}(t)(|t| - 2\varrho), & |t| > C(s,t) + 2\varrho, \end{cases}$$

with $C(s,t) = \dfrac{1 - \text{sign}(s)\text{sign}(t)}{2}|s| \ge 0$. On one hand, when $s \neq 0$,

$$b^* = t - \mathcal{S}\big(t, C(s,t)\big) + \mathcal{S}\big\{\mathcal{S}\big(t, C(s,t)\big), 2\varrho\big\}.$$

On the other hand, when $s = 0$, it follows that $b^* = \mathcal{S}(t, 2\varrho)$ since $\mathcal{S}(z, 0) = z$. □

References

1. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
2. Wang, X.; Zhang, H.H.; Wu, Y. Multiclass probability estimation with support vector machines. *J. Comput. Graph. Stat.* **2019**, *28*, 586–595. [CrossRef]
3. Hansen, J.H.; Hasan, T. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Process. Mag.* **2015**, *32*, 74–99. [CrossRef]
4. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*; John Wiley & Sons: New York, NY, USA 2012.
5. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: New York, NY, USA, 2009.
6. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
7. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
8. Marron, J.S.; Todd, M.J.; Ahn, J. Distance-weighted discrimination. *J. Am. Stat. Assoc.* **2007**, *102*, 1267–1271. [CrossRef]
9. Qiao, X.; Zhang, H.H.; Liu, Y.; Todd, M.J.; Marron, J.S. Weighted distance weighted discrimination and its asymptotic properties. *J. Am. Stat. Assoc.* **2010**, *105*, 401–414. [CrossRef]
10. Marron, J. Distance-weighted discrimination. *Wiley Interdiscip. Rev. Comput. Stat.* **2015**, *7*, 109–114. [CrossRef]

11. Zhang, L.; Lin, X. Some considerations of classification for high dimension low-sample size data. *Stat. Methods Med. Res.* **2013**, *22*, 537–550. [CrossRef]
12. Wang, B.; Zou, H. Another look at distance-weighted discrimination. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2018**, *80*, 177–198. [CrossRef]
13. Liu, Y.; Zhang, H.H.; Wu, Y. Hard or soft classification? Large-margin unified machines. *J. Am. Stat. Assoc.* **2011**, *106*, 166–177. [CrossRef]
14. Huang, H.; Liu, Y.; Du, Y.; Perou, C.M.; Hayes, D.N.; Todd, M.J.; Marron, J.S. Multiclass distance-weighted discrimination. *J. Comput. Graph. Stat.* **2013**, *22*, 953–969. [CrossRef]
15. Wang, B.; Zou, H. A multicategory kernel distance weighted discrimination method for multiclass classification. *Technometrics* **2019**, *61*, 396–408. [CrossRef]
16. Wang, B.; Zou, H. Sparse distance weighted discrimination. *J. Comput. Graph. Stat.* **2016**, *25*, 826–838. [CrossRef]
17. Wang, L.; Shen, X. On L1-norm multiclass support vector machines: Methodology and theory. *J. Am. Stat. Assoc.* **2007**, *102*, 583–594. [CrossRef]
18. Zhang, X.; Wu, Y.; Wang, L.; Li, R. Variable selection for support vector machines in moderately high dimensions. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2016**, *78*, 53–76. [CrossRef]
19. Peng, B.; Wang, L.; Wu, Y. An error bound for L1-norm support vector machine coefficients in ultra-high dimension. *J. Mach. Learn. Res.* **2016**, *17*, 8279–8304.
20. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [CrossRef]
21. Friedman, J.; Hastie, T.; Tibshirani, R. A note on the group lasso and a sparse group lasso. *arXiv* **2010**, arXiv:1001.0736.
22. Cai, T.T.; Zhang, A.; Zhou, Y. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *arXiv* **2019**, arXiv:1909.09851.
23. Yu, D.; Zhang, L.; Mizera, I.; Jiang, B.; Kong, L. Sparse wavelet estimation in quantile regression with multiple functional predictors. *Comput. Stat. Data Anal.* **2019**, *136*, 12–29. [CrossRef]
24. He, Q.; Kong, L.; Wang, Y.; Wang, S.; Chan, T.A.; Holland, E. Regularized quantile regression under heterogeneous sparsity with application to quantitative genetic traits. *Comput. Stat. Data Anal.* **2016**, *95*, 222–239. [CrossRef]
25. Huang, H. Large dimensional analysis of general margin based classification methods. *arXiv* **2019**, arXiv:1901.08057.
26. Huang, H.; Yang, Q. Large scale analysis of generalization error in learning using margin based classification methods. *arXiv* **2020**, arXiv:2007.10112.
27. Lam, X.Y.; Marron, J.; Sun, D.; Toh, K.C. Fast algorithms for large-scale generalized distance weighted discrimination. *J. Comput. Graph. Stat.* **2018**, *27*, 368–379. [CrossRef]
28. Sun, D.; Toh, K.C.; Yang, L. A convergent 3-block semiproximal alternating direction method of multipliers for conic programming with 4-type constraints. *SIAM J. Optim.* **2015**, *25*, 882–915. [CrossRef]
29. Parikh, N.; Boyd, S. Proximal algorithms. *Found. Trends Optim.* **2014**, *1*, 127–239. [CrossRef]
30. Asahchop, E.L.; Branton, W.G.; Krishnan, A.; Chen, P.A.; Yang, D.; Kong, L.; Zochodne, D.W.; Brew, B.J.; Gill, M.J.; Power, C. HIV-associated sensory polyneuropathy and neuronal injury are associated with miRNA–455-3p induction. *JCI Insight* **2018**, *3*, e122450. [CrossRef]
31. Hsu, D.; Kakade, S.; Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **2012**, *17*, 52. [CrossRef]