Journal of Machine Learning Research 24 (2023) 1-35 Submitted 4/22; Revised 1/23; Published 4/23

Generalization Error Bounds for Multiclass Sparse Linear Classifiers

Tomer Levy TMRLVI@GMAIL.COM
Department of Statistics and Operation Research
Tel Aviv University

Felix Abramovich FELIX@TAUEX.TAU.AC.IL
Department of Statistics and Operation Research
Tel Aviv University

Editor: Pierre Alquier

Abstract

We consider high-dimensional multiclass classification by sparse multinomial logistic regression. Unlike binary classification, in the multiclass setup one can think about an entire spectrum of possible notions of sparsity associated with different structural assumptions on the regression coefficients matrix. We propose a computationally feasible feature selection procedure based on penalized maximum likelihood with convex penalties capturing a specific type of sparsity at hand. In particular, we consider global row-wise sparsity, double row-wise sparsity, and low-rank sparsity, and show that with the properly chosen tuning parameters the derived plug-in classifiers attain the minimax generalization error bounds (in terms of misclassification excess risk) within the corresponding classes of multiclass sparse linear classifiers. The developed approach is general and can be adapted to other types of sparsity as well.

Keywords: Feature selection, high-dimensionality, minimaxity, misclassification excess risk, sparsity

1. Introduction

Classification is a core problem of statistical and machine learning. One of its main challenges nowadays is high-dimensionality of the data, where the number of features d is of the same order or even larger than the available sample size n ("large d, small n" setup) that causes a severe "curse of dimensionality" problem. Moreover, the number of classes L may also be large ("large L, large d, small n" model). A key assumption to handle the "curse of dimensionality" is *sparsity*. Dimension reduction of the feature domain by selecting a sparse subset of significant features becomes crucial. Bickel and Levina (2004) and Fan and Fan (2008) showed that even binary classification in high-dimensional setup without a proper feature selection procedure might be as bad as pure guessing. Feature selection and classification procedures should also be computationally feasible to deal with high-dimensional data.

Although there exists a large amount of statistical and machine learning literature on feature selection in classification, the rigorous theory on the accuracy of resulting classifiers has been mostly developed for the simplest binary case. See Vapnik (2000), Shalev-Shwartz and Ben-David (2014), Mohri et al. (2018).

One common strategy for multiclass classification is its reduction to a series of binary classifications. The two most well-known methods are One-vs-All (OvA), where each class is compared against all others, and One-vs-One (OvO), where all pairs of classes are compared to each other. A more direct and appealing strategy is extending binary classification methods to a multiclass setup. One approach is based on empirical risk minimization (ERM) (e.g., Koltchinskii and Panchenko, 2002; Daniely et al., 2012). A general crucial drawback of ERM is in minimization of the non-convex 0-1 loss and a common remedy is to replace it by some convex surrogate. Zhang (2004), Chen and Sun (2006), Daniely et al. (2015), Maximov and Reshetova (2016), Lei et al. (2019) and Reeve and Kaban (2020) (see also references therein) investigated the error bounds for various surrogate losses in terms of Rademacher complexity, covering numbers, or Natarajan/graph dimensions. Daniely et al. (2015) compared these results with those for OvA and OvO. However, all the above works do not consider feature selection and to the best of our knowledge, there are no theoretical results for the ERM-based approach in high-dimensional sparse multiclass setups.

An alternative approach to ERM is plug-in classifiers, where one assumes some model for the underlying unknown probabilities of outcome classes, estimates them from the data and plugs-in estimated probabilities to derive a classification rule. It may be especially useful when one is interested not only in prediction but also in interpretability and inference. In particular, in this paper we consider multinomial logistic (linear) classifiers – one of the mostly used classification tools. We investigate feature selection in high-dimensional multinomial logistic regression model and the accuracy of the resulting plug-in classifiers under various sparsity scenarios.

For binary classification the notion of sparsity is naturally associated with the number of significant features. For linear classifiers it is the number of non-zero entries of a vector of coefficients $\boldsymbol{\beta}$. For multiclass case, in contrast, there is a matrix of coefficients B that allows one to consider the entire spectrum of various types of sparsity associated with different structural assumptions on B. Abramovich et al. (2021) studied the most straightforward extension of multiclass sparsity measured by the number of non-zero rows of B. Such *global row-wise* sparsity corresponds to the assumption that most of features do not affect any class predictions at all. In this paper we present other possible extensions. In particular, we consider *double row-wise* sparsity, where it is still assumed that B has a sparse subset of non-zero rows (global sparsity) but, in addition, its non-zero rows are also sparse (*local row-wise* sparsity), and the *low-rank* sparsity, where B is assumed to be of a low rank. The latter assumption is associated with the existence of a smaller number of latent variables defining the outcome classes.

For each considered type of sparsity we propose penalized maximum likelihood feature selection procedures with the corresponding convex penalties and establish the bounds for generalization errors in terms of misclassification excess risk of the resulting multinomial logistic regression classifiers. The penalties are variations of a celebrated Lasso and its recently developed more general and flexible version Slope (Bogdan et al., 2015). We show that for the proper choice of tuning parameters the derived classifiers attain the optimal (in the minimax sense) generalization errors within the corresponding classes of sparse linear classifiers. The errors can be improved under the additional low-noise condition.

The proposed approach is general and can be adapted to other types of sparsity. The machinery for a general form of a sparse multinomial logistic regression classifier is developed in Appendix A.1.

The paper is organized as follows. Section 2 presents sparse multinomial logistic regression model and some preliminaries. Section 3 contains the main theoretical results, where we introduce feature selection procedures for various types of sparsity and derive the error bounds for the resulting misclassification excess risks. In Section 4 we illustrate the performance of the developed procedures on a real-data example and compare them with other existing classifiers. Some concluding remarks are given in Section 5. All the proofs are left to the Appendix.

2. Sparse multinomial logistic regression

Consider a d-dimensional L-class classification model:

$$Y|(\boldsymbol{X}=\boldsymbol{x}) \sim Mult(p_1(\boldsymbol{x}),\ldots,p_L(\boldsymbol{x})), \qquad \sum_{j=1}^{L} p_j(\boldsymbol{x})=1, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^d$ is a vector of linearly independent features with a marginal probability distribution $\mathbb{P}_X$ on a bounded support $\mathcal{X} \subseteq \mathbb{R}^d$. Let $V = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T)$ be the second moment matrix of $\boldsymbol{X}$.

We consider a multinomial logistic regression model, where it is assumed that

$$p_l(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_l^T \boldsymbol{x})}{\sum_{k=1}^L \exp(\boldsymbol{\beta}_k^T \boldsymbol{x})}, \quad |\boldsymbol{\beta}_l|_2 \leq R, \quad l = 1, \ldots, L, \tag{2}$$

Let $B \in \mathbb{R}^{d \times L}$ be the corresponding matrix of regression coefficients with columns $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L$. The model (2) is not identifiable without an extra constraint on B since shifting each $\boldsymbol{\beta}_l$ by the same vector $\boldsymbol{c}$ does not affect the probabilities $p_l(\boldsymbol{x})$. In this paper we adopt a symmetric zero mean rows constraint $\sum_{l=1}^L \boldsymbol{\beta}_l = \boldsymbol{0}$ or, equivalently, $B\boldsymbol{1} = \boldsymbol{0}_d$. Hence, $\boldsymbol{\beta}_l$ represents the effects of $\boldsymbol{x}$ in the l-th class w.r.t. the mean response over all classes on the log-scale:

$$\boldsymbol{\beta}_l^T \boldsymbol{x} = \ln\left(\frac{p_l(\boldsymbol{x})}{\sqrt[L]{\prod_{k=1}^L p_k(\boldsymbol{x})}}\right) = \ln p_l(\boldsymbol{x}) - \overline{\ln p(\boldsymbol{x})}.$$

One can evidently choose any other constraint (e.g., $\boldsymbol{\beta}_L = \boldsymbol{0}$, where the L-class is used as the reference one) – the models will be equivalent but the vectors of coefficients $\boldsymbol{\beta}_l$ will have different interpretation. In particular, the symmetric constraint implies that the model is invariant to permutations of the classes.

For the considered multinomial logistic regression model (1)-(2) the Bayes classifier that minimizes generalized misclassification error (risk) $R(\eta) = P(Y \neq \eta(\boldsymbol{X}))$ is a linear classifier $\eta^*(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} p_l(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} \boldsymbol{\beta}_l^T \boldsymbol{x}$ with the (oracle) misclassification risk $R(\eta^*) = 1 - E_{\boldsymbol{X}} \max_{1 \leq l \leq L} p_l(\boldsymbol{x})$.

Given a random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, we estimate the unknown matrix B from the data and consider the resulting plug-in classifier $\widehat{\eta}_{\widehat{B}}(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} \widehat{\boldsymbol{\beta}}_l^T \boldsymbol{x}$. Its conditional misclassification error is $R(\widehat{\eta}_{\widehat{B}}) = P\left(Y \neq \widehat{\eta}_{\widehat{B}}(\boldsymbol{X}) | (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\right)$ and its goodness w.r.t. η^* is measured by the misclassification excess risk

$$\mathcal{E}(\widehat{\eta}_{\widehat{B}}, \eta^*) = \mathbb{E}R(\widehat{\eta}_{\widehat{B}}) - R(\eta^*).$$

The goal is to find $\widehat{B}$ that yields the minimal $\mathcal{E}(\widehat{\eta}_{\widehat{B}}, \eta^*)$.

Consider the log-likelihood function for the multinomial logistic regression model (1)-(2):

$$\ell(B) = \sum_{i=1}^n \left\{ \boldsymbol{X}_i^T B \boldsymbol{\xi}_i - \ln \sum_{l=1}^L \exp(\boldsymbol{\beta}_l^T \boldsymbol{X}_i) \right\}, \tag{3}$$

where $\boldsymbol{\xi}_i \in \mathbb{R}^L$ is the indicator vector corresponding to Y_i with $\xi_{il} = I\{Y_i = l\}$. One can find the maximal likelihood estimator (MLE) for B by maximizing $\ell(B)$ under the identifiability symmetric constraint $B\mathbf{1} = \mathbf{0}_d$. Although the solution is not available in closed form, it can be nevertheless obtained numerically by the fast iteratively reweighted least squares algorithm (McCullagh and Nelder, 1989).

As we have mentioned in the introduction, feature selection is essential for high-dimensional classification. To perform feature selection we consider a *penalized* maximum likelihood estimator of the form:

$$
\begin{aligned}
\widehat{B} &= \arg\min_{\widetilde{B}:\widetilde{B}\mathbf{1}=\mathbf{0}_d} \{-\ell(\widetilde{B}) + Pen(\widetilde{B})\} \\
&= \arg\min_{\widetilde{B}:\widetilde{B}\mathbf{1}=\mathbf{0}_d} \left\{ \sum_{i=1}^{n} \left(\ln \sum_{l=1}^{L} \exp(\boldsymbol{\beta}_l^T \boldsymbol{X}_i) - \boldsymbol{X}_i^T \widetilde{B} \boldsymbol{\xi}_i \right) + Pen(\widetilde{B}) \right\}
\end{aligned}
\tag{4}
$$

with a penalty $Pen(\cdot)$ capturing specific sparsity assumptions on B.

3. Main results

For binary classification, where a matrix B reduces to a single vector $\boldsymbol{\beta} \in \mathbb{R}^d$, the sparsity is naturally characterized by the l_0 (quasi)-norm $||\boldsymbol{\beta}||_0$ – the number of non-zero entries of $\boldsymbol{\beta}$ (see, e.g., Abramovich and Grinshtein, 2019; Chen and Lee, 2021). For the multiclass case there is a wide spectrum of possible ways to extend the notion of sparsity associated with different assumptions on the regression coefficients matrix B. In this section we consider several of them and derive misclassification excess risk bounds for the resulting multiclass sparse linear classifiers.

The straightforward approach in (4) is to use complexity-type penalties that mimic sparsity directly. However, despite strong theoretical ground (see, e.g., Abramovich et al., 2021), it is computationally infeasible for high-dimensional data since solving (4) requires in this case a combinatorial search over all possible models. The goal then is to find convex surrogates for complexity penalties while preserving their theoretical properties.

3.1 Global row-wise sparsity

The most straightforward extension of notion of sparsity for multiclass classification is *global* sparsity, where it is assumed that part of features do not affect any class predictions at all. In terms of the matrix B global sparsity corresponds to the assumption that B has a "small" number of non-zero rows (global row-wise sparsity). Such type of sparsity was studied in Abramovich et al. (2021) and in this subsection we review their main results (generalizing for a anisotropic $\boldsymbol{X}$) in order to extend them afterwards to other, finer types of sparsity.

Let r_B be the number of non-zero rows of B. To capture the global sparsity Abramovich et al. (2021) proposed to use a complexity penalty on the number of non-zero rows of B in (4).

Let $\mathcal{M} = \{B \in \mathbb{R}^{d\times L} : B\mathbf{1} = \mathbf{0}\}$ be the set of regression matrices satisfying the symmetric constraint, $\mathcal{M}(d_0) = \{B \in \mathcal{M} : r_B \leq d_0\}$ be its subset of d_0-globally row-wise sparse matrices and $\mathcal{C}_L(d_0) = \{\eta(\boldsymbol{x}) = \arg\max_{1\leq l\leq L} \boldsymbol{\beta}_l^T \boldsymbol{x} : B \in \mathcal{M}(d_0)\}$ be the set of d_0-sparse linear L-class classifiers. Define the penalized maximum likelihood estimator $\hat{B}$

$$\widehat{B} = \arg\min_{\widetilde{B}\in\mathcal{M}} \left\{-\ell(\widetilde{B}) + Pen(r_{\widetilde{B}})\right\} \tag{5}$$

with the complexity penalty of the form

$$Pen(r_{\widetilde{B}}) = C_1\ r_{\widetilde{B}}(L-1) + C_2\ r_{\widetilde{B}} \ln\left(\frac{de}{r_{\widetilde{B}}}\right) \tag{6}$$

for some positive constants C_1 and C_2.

Abramovich et al. (2021) showed that for the bounded $\mathcal{X}$,

$$\sup_{\eta^*\in\mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{\widehat{B}}, \eta^*) \leq C\sqrt{\frac{d_0(L-1) + d_0\ln\left(\frac{de}{d_0}\right)}{n}} \tag{7}$$

for some $C > 0$ simultaneously for all $1 \leq d_0 \leq \min(d, n/(L-1))$, and that the bound in (7), up to a probably different constant, is also the minimax for $\mathcal{C}_L(d_0)$.

Misclassification excess risk bounds (7) show that there is a phase transition between small and large number of classes. For $L \leq 2 + \ln(d/d_0)$, the multiclass effect is not yet manifested and the minimax misclassification excess risk over the set of d_0-sparse linear classifiers is of the order $\sqrt{\frac{d_0}{n}\ln\left(\frac{de}{d_0}\right)}$ regardless of L. Note that $d_0\ln\left(\frac{de}{d_0}\right) \sim \ln\binom{d}{d_0}$ which is the log of the number of all possible models of size d_0. For larger L, the risk is of the order $\sqrt{\frac{d_0(L-1)}{n}}$, where $d_0(L-1)$ is the overall number of estimated parameters in the given model of size d_0, and does not depend on d. For $L > n/d_0$ the number of parameters in the (true) model becomes larger than the sample size and consistent classification is evidently impossible.

Classification is mostly challenging at points, where it is difficult to distinguish the most likely class from others, that is, at those $\boldsymbol{x} \in \mathcal{X}$, where the largest probability $p_{(1)}(\boldsymbol{x})$ is close to the second largest $p_{(2)}(\boldsymbol{x})$. The misclassification error bounds (7) may be then improved under the additional multiclass extension of the low-noise (aka Tsybakov) condition (Tsybakov, 2004):

Assumption A *Consider the multinomial logistic regression model (1)-(2) and assume that there exist* $C>0, \alpha \geq 0$ *and* $h^*>0$ *such that for all* $0<h \leq h^*$,

$$P\left(p_{(1)}(\boldsymbol{X})-p_{(2)}(\boldsymbol{X}) \leq h\right) \leq C h^\alpha .$$

Assumption A implies that with high probability (depending on the parameter α) the most likely class is sufficiently distinguished from others. The two extreme cases are $\alpha=0$ and $\alpha=\infty$. The former does not impose any assumption on the noise, while the latter assumes the existence of a hard margin of size h^* separating $p_{(1)}(\boldsymbol{x})$ and $p_{(2)}(\boldsymbol{x})$.

Abramovich et al. (2021) proved that under the additional low-noise Assumption A the misclassification excess risk bound (7) of $\widehat{\eta}_{\widehat{B}}$ can be indeed improved:

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{\widehat{B}}, \eta^*) \leq \left(C\, \frac{d_0(L-1)+d_0 \ln\left(\frac{de}{d_0}\right)}{n}\right)^{\frac{\alpha+1}{\alpha+2}} \tag{8}$$

for all $1 \leq d_0 \leq \min(d, n/(L-1))$ and all $\alpha \geq 0$. Note that the proposed classifier $\widehat{\eta}_{\widehat{B}}$ is inherently adaptive to both sparsity d_0 and noise level α.

As we have mentioned above, solving for $\widehat{B}$ in (5) requires a combinatorial search over all possible 2^d models that makes it computationally infeasible for large d. One should apply convex relaxation techniques to replace the original complexity penalty (6) by some convex surrogate.

The well-known examples of convex surrogates are celebrated Lasso, where the l_0-norm in the complexity penalty is replaced by the l_1-norm norm, and its recently developed more general variation Slope that uses a *sorted* l_1-type norm (Bogdan et al., 2015). Lasso and Slope estimators have been intensively studied in the last decade in various regression setups (see e.g., van de Geer, 2008; Bickel et al., 2009; Su and Candès, 2016; Bellec et al., 2018; Abramovich and Grinshtein, 2019; Alquier et al., 2019). Abramovich and Grinshtein (2019) and Abramovich et al. (2021) applied logistic Lasso and Slope classifiers for classification.

To capture a global row-wise sparsity for multinomial logistic regression, Abramovich et al. (2021) considered a *group* version of multinomial logistic Slope defined as follows. Let

$$\widehat{B}_{gS} = \arg\min_{\widetilde{B} \in \mathcal{M}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\ln\left(\sum_{l=1}^L \exp(\widetilde{\boldsymbol{\beta}}_l^T \boldsymbol{X}_i) \right) - \boldsymbol{X}_i^T \widetilde{B} \boldsymbol{\xi}_i \right) + \sum_{j=1}^d \lambda_j |\widetilde{B}|_{(j)} \right\}, \tag{9}$$

where $|\widetilde{B}|_{(1)} \geq \ldots \geq |\widetilde{B}|_{(d)}$ are the descendingly ordered l_2-norms of rows of $\widetilde{B}$ and $\lambda_1 \geq \ldots \geq \lambda_d > 0$ are tuning parameters, and define $\widehat{\eta}_{gS}(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} \widehat{\boldsymbol{\beta}}_{gS,l}^T \boldsymbol{x}$. Multinomial logistic group Lasso classifier $\widehat{\eta}_{gL}$ is a particular case of $\widehat{\eta}_{gS}$ corresponding to equal λ_j's in (9).

The identifiability symmetric constraint $\widetilde{B} \in \mathcal{M}$ is, in fact, unnecessary in (9) since unlike the complexity penalty in (5), the solution of (9) is identifiable without any additional constraint. Moreover, since the unconstrained log-likelihood (3) satisfies $\ell(\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_L) = \ell(\widetilde{\boldsymbol{\beta}}_1 - \mathbf{c}, \ldots, \widetilde{\boldsymbol{\beta}}_L - \mathbf{c})$ for any vector $\mathbf{c} \in \mathbb{R}^d$, it can be always improved by taking $\hat{c}_j = \arg\min_{c_j} \sum_{l=1}^L (\widetilde{B}_{jl} - c_j)^2$, that is, for $\hat{c}_j = \bar{B}_{j\cdot}$. Hence, the unconstrained solution of (9) will inherently have zero mean rows.

As usual for convex relaxation, one needs some (mild) extra conditions on the design. Assume that all X_j are scaled, i.e. $EX_j^2 = 1,\ j = 1, \ldots, d$. For a given matrix $A \in \mathcal{M}$ let $\Pi_{d_0}(A)$ be its d_0-sparse projection, i.e. the matrix with at most d_0 nonzero rows closest to A in the Frobenius norm.

Assumption $\mathbf{B_1}$ *Assume that*

$$\nu_{gS}(d_0) = \inf_{A \in \mathcal{M}: A \neq 0_{d\times L}} \frac{\|V^{\frac{1}{2}} A\|_F^2}{\|\Pi_{d_0}(A)\|_F^2} > 0,$$

In fact, one immediately realizes that $\Pi_{d_0}(A)$ keeps d_0 rows of A with the largest l_2-norms and zeroes other rows. Hence, $\|\Pi_{d_0}(A)\|_F^2 = \sum_{j=1}^{d_0} |A|_{(j)}^2$.

Such or similar types of conditions are common for convex relaxation methods (see Bellec et al., 2018, Section 8 for discussion).

Let $||A||_{gS} = \sum_{j=1}^d \lambda_j |A|_{(j)}$ be the group Slope norm of a matrix A. The following theorem provides an upper bound for misclassification excess risk of the group Slope classifier extending the results of Abramovich et al. (2021) to anisotropic design. In addition, it provides also the upper bounds for the integrated prediction error $\sum_{l=1}^L \mathbb{E}\|(\widehat{\boldsymbol{\beta}}_{gS,l} - \boldsymbol{\beta}_l)^T \boldsymbol{x}\|_{L_2(\mathbb{P}_X)}^2 = E\|V^{\frac{1}{2}}(\widehat{B}_{gS} - B)\|_F^2$ and the estimation error of the regression coefficients matrix B w.r.t. the group Slope norm $\mathbb{E}||\widehat{B}_{gS} - B||_{gS}$:

Theorem 1 *Consider a d_0-globally row-sparse multinomial logistic regression (1)-(2), where $d_0 \leq \min(d, n/(L-1))$ and $\boldsymbol{X}_j$'s are scaled to have $EX_j^2 = 1,\ j = 1, \ldots, d$. Apply the multinomial logistic sparse group Slope classifier (12) with λ_j's satisfying*

$$\max_{1 \leq j \leq d} \frac{\sqrt{L + \ln(d/j)}}{\lambda_j} \leq C_0 \sqrt{n}, \tag{10}$$

where the constant C_0 is derived from Abramovich et al. (2021). Then, under Assumptions A-B_1,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{gS}, \eta^*) \leq \left(\frac{C}{\nu_{gS}(d_0)} \sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right)^{\frac{2(\alpha+1)}{\alpha+2}}.$$

In addition,

$$\sup_{B \in \mathcal{M}(d_0)} \mathbb{E}\|V^{\frac{1}{2}}(\widehat{B}_{gS} - B)\|_F^2 \leq \frac{C_1}{\nu_{gS}(d_0)} \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right)^2$$

and

$$\sup_{B \in \mathcal{M}(d_0)} \mathbb{E}||\widehat{B}_{gS} - B||_{gS} \leq \frac{C_2}{\nu_{gS}(d_0)} \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right)^2$$

In particular, setting

$$\lambda_j = \frac{1}{C_0}\sqrt{\frac{L + \ln(d/j)}{n}},$$

the misclassification excess risk of the multinomial logistic group Slope classifier $\widehat{\eta}_{gS}$ is of the minimax order (8):

Corollary 2 *Apply Theorem 1 with*

$$\lambda_j = \frac{1}{C_0}\sqrt{\frac{L + \ln(d/j)}{n}}.$$

Then, under Assumptions A-B$_1$,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{gS}, \eta^*) \leq \left(\frac{C}{\nu_{gS}(d_0)} \; \frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n} \right)^{\frac{\alpha+1}{\alpha+2}}. \tag{11}$$

Furthermore,

$$\sup_{B \in \mathcal{M}(d_0)} \mathbb{E}\|V^{\frac{1}{2}}(\widehat{B}_{gS} - B)\|_F^2 \leq \frac{C_1}{\nu_{gS}(d_0)} \; \frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}$$

and

$$\sup_{B \in \mathcal{M}(d_0)} \mathbb{E}||\widehat{B}_{gS} - B||_{gS} \leq \frac{C_2}{\nu_{gS}(d_0)} \; \frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}$$

Note that $\widehat{\eta}_{gS}$ is inherently adaptive to d_0 and α.

Similarly, the multinomial logistic group Lasso classifier $\widehat{\eta}_{gL}$ with a (constant) $\lambda = \frac{1}{C_0}\sqrt{\frac{L+\ln d}{n}}$ is sub-optimal (up to the log-factor):

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{gL}, \eta^*) \leq \left(\frac{C}{\nu_{gS}(d_0)} \; \frac{d_0(L-1) + d_0 \ln d}{n} \right)^{\frac{\alpha+1}{\alpha+2}}.$$

We consider now other possible types of sparsity for multiclass case and derive the corresponding generalization error bounds.

3.2 Double row-wise sparsity

We show that the misclassification excess risks bounds for a global row-wise sparsity can be improved under a finer row-wise sparsity structure. Namely, assume that even each significant feature is involved in only part of probabilities p_l's. In terms of the matrix B it implies the additional sparsity assumption on its non-zero rows in the usual l_0-norm sense, i.e., local row-wise sparsity.

For a given matrix B, let $\mathcal{J}(B) = \{j_1, \ldots, j_{r_B}\}$ be the set of indices of its non-zero rows. Consider a set of *double* (global and local) row-wise sparse matrices $\mathcal{M}(d_0, \mathbf{m}) = \{B \in \mathcal{M} : |\mathcal{J}(B)| \leq d_0;\ ||B_{j\cdot}||_0 \leq m_j,\ j \in \mathcal{J}(B)\}$ and the corresponding set of $(d_0, \mathbf{m})$-sparse linear L-class classifiers $\mathcal{C}_L(d_0, \mathbf{m}) = \{\eta(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} \boldsymbol{\beta}_l^T \boldsymbol{x} : B \in \mathcal{M}(d_0, \mathbf{m})\}$.

To capture a double row-wise sparsity one should impose penalties on both the number of non-zero rows of B and on the numbers of their non-zero entries. A natural convex surrogate in this case is a multinomial logistic *sparse* group Slope estimator of B defined as follows:

$$\begin{aligned} \widehat{B}_{sgS} = \arg\min_{\widetilde{B} \in \mathcal{M}} \Bigg\{ \frac{1}{n} \sum_{i=1}^n \left(\ln \left(\sum_{l=1}^L \exp(\widetilde{\boldsymbol{\beta}}_l^T \boldsymbol{X}_i) \right) - \boldsymbol{X}_i^T \widetilde{B} \boldsymbol{\xi}_i \right) \\ + \sum_{j=1}^d \lambda_j |\widetilde{B}|_{(j)} + \sum_{j=1}^d \sum_{l=1}^L \kappa_l |\widetilde{B}|_{j(l)} \Bigg\}, \end{aligned} \tag{12}$$

where $|\widetilde{B}|_{(1)} \geq \ldots \geq |\widetilde{B}|_{(d)}$ are the descendingly ordered l_2-norms of the rows of $\widetilde{B}$, $|\widetilde{B}|_{j(1)} \geq \ldots \geq |\widetilde{B}|_{j(L)}$ are the descendingly ordered absolute values of entries of its j-th row, and $\lambda_1 \geq \ldots \geq \lambda_d > 0$ and $\kappa_1 \geq \ldots \geq \kappa_L > 0$ are tuning parameters. The additional last term in the penalty in (12) yields sparsity of non-zero rows. Sparse group Slope essentially combines group Slope on the row's norms with usual Slope within each row.

The multinomial logistic sparse group Slope classifier is $\widehat{\eta}_{sgS}(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} \widehat{\boldsymbol{\beta}}_{sgS,l}^T \boldsymbol{x}$. Multinomial logistic sparse group Lasso classifier $\widehat{\eta}_{sgL}$ (see Friedman et al., 2010; Vincent and Hansen, 2014) is its particular case with identical λ_j's and κ_l's in (12).

Let $||A||_{sgS} = \sum_{j=1}^d \lambda_j |A|_{(j)} + \sum_{j=1}^d \sum_{l=1}^L \kappa_l |A|_{j(l)}$ be the sparse group Slope norm of a matrix $A \in \mathbb{R}^{d \times L}$. The following theorem provides an upper bound for misclassification excess risk of $\widehat{\eta}_{sgS}$:

Theorem 3 *Consider a $(d_0, \mathbf{m})$-sparse multinomial logistic regression (1)-(2) with $d_0 \leq \min(d, n/(L-1))$ and scaled X_j's. Apply the multinomial logistic sparse group classifier*

(12) with λ_j's and κ_l's satisfying $\kappa_L \geq \frac{1}{C\sqrt{n}}$ and

$$\max_{1\leq j\leq d} \frac{\sqrt{2\sum_{l=1}^{L}\frac{1}{l}\left(\frac{Le}{l}\right)^l e^{-C^2 nl\kappa_l^2} + 2\log\left(\frac{de}{j}\right)}}{\lambda_j} \leq \frac{2C}{C_0}\sqrt{n}, \tag{13}$$

where $C = \sqrt{\frac{2}{\pi}}\frac{7}{2880}$ and C_0 is derived in the proof. Then, under Assumptions A-B$_1$,

$$\sup_{\eta^*\in\mathcal{C}_L(d_0,\mathbf{m})} \mathcal{E}(\widehat{\eta}_{sgS},\eta^*) \leq \left(\frac{C}{\nu_{gS}(d_0)}\left(\sum_{j=1}^{d_0}\frac{\lambda_j}{\sqrt{j}} + \sqrt{\sum_{j=1}^{d_0}\left(\sum_{l=1}^{m_j}\frac{\kappa_l}{\sqrt{l}}\right)^2}\right)\right)^{\frac{2(\alpha+1)}{\alpha+2}}.$$

In addition,

$$\sup_{B\in\mathcal{M}(d_0,\boldsymbol{m})} \mathbb{E}\|V^{\frac{1}{2}}(\widehat{B}_{sgS}-B)\|_F^2 \leq \frac{C_1}{\nu_{gS}(d_0)}\left(\sum_{j=1}^{d_0}\frac{\lambda_j}{\sqrt{j}} + \sqrt{\sum_{j=1}^{d_0}\left(\sum_{l=1}^{m_j}\frac{\kappa_l}{\sqrt{l}}\right)^2}\right)^2$$

and

$$\sup_{B\in\mathcal{M}(d_0,\mathbf{m})} \mathbb{E}||\widehat{B}_{sgS}-B||_{sgS} \leq \frac{C_2}{\nu_{gS}(d_0)}\left(\sum_{j=1}^{d_0}\frac{\lambda_j}{\sqrt{j}} + \sqrt{\sum_{j=1}^{d_0}\left(\sum_{l=1}^{m_j}\frac{\kappa_l}{\sqrt{l}}\right)^2}\right)^2.$$

The proof is given in the Appendix A.

In particular, take $\lambda_j = c_1\sqrt{\frac{\ln(de/j)}{n}}$, $j = 1,\ldots,d$ and $\kappa_l = c_2\sqrt{\frac{\ln(Le/l)}{n}}$, $l = 1,\ldots,L$ with $c_1 = \frac{1440C_0\sqrt{\pi}}{7}$ and $c_2 = \frac{2880\sqrt{\pi}}{7}$. By a straightforward calculus one can verify that these λ_j's and κ_l's satisfy the condition (13), and Theorem 3 then implies:

Corollary 4 *Apply Theorem 3 with*

$$\lambda_j = c_1\sqrt{\frac{\ln(de/j)}{n}} \quad \text{and} \quad \kappa_l = c_2\sqrt{\frac{\ln(Le/l)}{n}}, \tag{14}$$

where $c_1 = \frac{1440\sqrt{2\pi}C_0}{7}$, $c_2 = \frac{2880\sqrt{\pi}}{7}$ and C_0 is given in Lemma 13. Then, under Assumptions A-B$_1$,

$$\sup_{\eta^*\in\mathcal{C}_L(d_0,\mathbf{m})} \mathcal{E}(\widehat{\eta}_{sgS},\eta^*) \leq \left(\frac{C}{\nu_{gS}(d_0)}\,\frac{d_0\ln\left(\frac{de}{d_0}\right) + \sum_{j\in\mathcal{J}(B)} m_j\ln\left(\frac{Le}{m_j}\right)}{n}\right)^{\frac{\alpha+1}{\alpha+2}} \tag{15}$$

In addition,

$$\sup_{B\in\mathcal{M}(d_0,\boldsymbol{m})} \mathbb{E}\|V^{\frac{1}{2}}(\widehat{B}_{sgS}-B)\|_F^2 \leq \frac{C_1}{\nu_{gS}(d_0)}\frac{d_0\ln\left(\frac{de}{d_0}\right) + \sum_{j\in\mathcal{J}(B)} m_j\ln\left(\frac{Le}{m_j}\right)}{n}$$

and

$$\sup_{B\in\mathcal{M}(d_0,\mathbf{m})} \mathbb{E}||\widehat{B}_{sgS}-B||_{sgS} \le \frac{C_2}{\nu_{gS}(d_0)} \, \frac{d_0\ln\left(\frac{de}{d_0}\right)+\sum_{j\in\mathcal{J}(B)} m_j \ln\left(\frac{Le}{m_j}\right)}{n}$$

Corollary 4 shows that with a proper choice of tuning parameters, the bounds for misclassification excess risk for the global row-wise sparsity (8) are improved under the stronger double row-wise sparsity assumption. The multinomial logistic sparse group Slope classifier $\widehat{\eta}_{sgS}$ is adaptive to $d_0, \mathbf{m}$ and α.

Similar to global sparsity, there is a phase transition between small and large number of classes. The numerator in the upper bounds contains two terms. The first term $d_0 \ln(de/d_0)$ corresponds again to the error of selecting a subset of d_0 nonzero rows out of d, while the second term $\sum_{j\in\mathcal{J}(B)} m_j \ln(Le/m_j)$ appears due to simultaneous estimation of d_0 m_j-sparse vectors from $\mathbb{R}^L$. Since $d_0\ln(Le) \le \sum_{j\in\mathcal{J}(B)} m_j \ln(Le/m_j) \le d_0 L$, the first term is always dominating for small number of classes with $L \le \ln(de/d_0)$, while the second term is the main one for large number of classes with $L \ge d/d_0$.

It also follows from Theorem 3 that, similar to the group Lasso, the multinomial logistic sparse group Lasso classifier with constant $\lambda = c_1\sqrt{\frac{\ln d}{n}}$ and $\kappa = c_2\sqrt{\frac{\ln L}{n}}$ in (12) is sub-optimal up to the differences in the log-terms:

$$\sup_{\eta^*\in\mathcal{C}_L(d_0,\mathbf{m})} \mathcal{E}(\widehat{\eta}_{sgL},\eta^*) \le \left(\frac{C}{\nu_{gS}(d_0)}\,\frac{d_0\ln d+\ln L\cdot\sum_{j\in\mathcal{J}(B)} m_j}{n}\right)^{\frac{\alpha+1}{\alpha+2}}.$$

Note that unlike global row-wise sparsity, interpretation of local (and, therefore, the double) row-wise sparsity assumption depends on the chosen constraint on B. Thus, a non-zero row of B may be sparse (in terms of l_0-norm) under the symmetric constraint $\sum_{l=1}^L \boldsymbol{\beta}_l = \mathbf{0}$ but not necessarily sparse under another possible constraint, e.g., $\boldsymbol{\beta}_L = \mathbf{0}$ and vice versa.

3.3 Low-rank sparsity

So far we considered various types of row-wise sparsity of the regression coefficients matrix B. A more general approach is to assume the existence of some underlying hidden low-dimensional structure, where there is a smaller number of latent variables that define the outcome classes. The row-wise sparsity is a particular case of such a general case. The natural measure of such type of sparsity (sometimes called also *spectral* sparsity) is $rank(B)$.

Direct penalization of $rank(B)$ implies a non-convex optimization since $rank(B)$ is not a convex function although She (2013) proposed a computationally fast procedure for its solution for GLM. To convexify rank penalization note that $rank(B) = ||\boldsymbol{\gamma}||_0$, where

$\gamma_1, \ldots, \gamma_{\min(L-1,d)}$ are the singular values of B. Similar to Lasso, we replace $||\boldsymbol{\gamma}||_0$ by $||\boldsymbol{\gamma}||_1$ aka a nuclear norm $||B||_*$ or, Schatten S_1-norm. Nuclear penalties have been intensively studied in statistical and machine learning for multivariate regression and matrix completion (e.g., Bach, 2008; Candes and Plan, 2010; Bunea et al., 2011; Koltchinskii et al., 2011; Alquier et al., 2019). Powers et al. (2018) considered nuclear penalization in multinomial logistic classification. They developed numerical algorithms for its solution but did not investigate theoretical properties of the resulting classifier.

We start from establishing a minimax lower bound for misclassification excess risk over a set of L-class linear classifiers with law rank coefficients matrices. Let $\mathcal{M}^*(r_0) = \{B \in \mathcal{M} : rank(B) \leq r_0\}$ and $\mathcal{C}^*_L(r_0) = \{\eta(\boldsymbol{x}) = \arg\max_{1 \leq l \leq L} \boldsymbol{\beta}_l^T \boldsymbol{x} : B \in \mathcal{M}^*(r_0)\}$.

Theorem 5 *Consider an agnostic multinomial regression model (1)-(2) with $rank(B) \leq r_0$, where $1 \leq r_0 \leq \min(L-1, d)$ and $r_0(L+d) \leq n$. Then,*

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}^*_L(r_0),\ \mathbb{P}_X} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{r_0((L-1)+d)}{n}} \tag{16}$$

for some $C > 0$.

The proof is given in the Appendix C.

We now show that estimating B by penalized maximum likelihood estimator with a nuclear penalty of the form $\lambda||B||_*$ with a properly chosen tuning parameter λ leads to a linear classifier that achieves the lower bound (16) up to a multiplicative term depending on the marginal distribution $\mathbb{P}_X$ of $\boldsymbol{X}$.

Define

$$\widehat{B}_{nu} = \underset{\widetilde{B}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\ln\left(\sum_{l=1}^L \exp(\widetilde{\beta}_l^T \boldsymbol{x}_i) \right) - \boldsymbol{\beta}_{y_i}^T \boldsymbol{x}_i \right) + \lambda||B||_* \right\}, \tag{17}$$

with $\lambda > 0$, and the corresponding classifier $\widehat{\eta}_{nu}(\boldsymbol{x}) = \operatorname{argmax}_{1 \leq l \leq L} \widehat{\boldsymbol{\beta}}^T_{nu,l} \boldsymbol{x}$. Similar to group Slope and group Lasso classifiers from Section 3.1, there is no need to impose an additional symmetric constraint $\widetilde{B} \in \mathcal{M}$ in (17) since centering rows to zero means can only decrease the nuclear norm of a matrix (Powers et al., 2018).

Let $\tau_1(V) \geq \ldots \geq \tau_d(V)$ be the ordered eigenvalues of the second moment matrix $V = \mathbb{E}_X(\boldsymbol{X}\boldsymbol{X}^T)$.

Assumption $\mathbf{B_2}$ *Assume that $\tau_d(V) > 0$.*

Theorem 6 *Consider a multinomial regression model (1)-(2) and the nuclear penalized classifier $\widehat{\eta}_{nu}(\boldsymbol{x})$ with*

$$\lambda = C\sqrt{\tau_1(V)}\ \frac{\sqrt{L-1}+\sqrt{d}}{\sqrt{n}}, \tag{18}$$

where $C > 0$ is specified in the proof.

Then, under Assumption B_2

$$\sup_{\eta^* \in \mathcal{C}^*_L(r_0)} \mathcal{E}(\widehat{\eta}_{nu}, \eta^*) \leq \sqrt{C\ \frac{\tau_1(V)}{\tau_d(V)}\ \frac{r_0((L-1)+d)}{n}}\ . \tag{19}$$

Furthermore, under the additional low-noise Assumption A,

$$\sup_{\eta^* \in \mathcal{C}^*_L(r_0)} \mathcal{E}(\widehat{\eta}_{nu}, \eta^*) \leq \left(C\ \frac{\tau_1(V)}{\tau_d(V)}\ \frac{r_0((L-1)+d)}{n}\right)^{\frac{\alpha+1}{\alpha+2}}\ . \tag{20}$$

In addition,

$$\sup_{B \in \mathcal{M}^*(r_0)} \mathbb{E}\|V^{\frac{1}{2}}(\widehat{B}_{nu} - B)\|_F^2 \leq C_1\ \frac{\tau_1(V)}{\tau_d(V)}\ \frac{r_0((L-1)+d)}{n}$$

and

$$\sup_{B \in \mathcal{M}^*(r_0)} \mathbb{E}||\widehat{B}_{nu} - B||_* \leq C_2\ \frac{\tau_1(V)}{\tau_d(V)}\ \frac{r_0((L-1)+d)}{n}$$

The proof is given in the Appendix A.

Similar upper bounds for the misclassification excess risk with the extra $\ln^{3/2}(n^{3/2}L)$-term can be derived from Corollary 10 of Lei et al. (2019) using (23) from Appendix A.

Summarizing, up to a multiplicative constant depending on the eigenvalues of the second moment matrix of $\boldsymbol{X}$, $\widehat{\eta}_{nu}(\boldsymbol{x})$ attains the minimax misclassification excess risk and is adaptive to the unknown low-rank sparsity of the regression coefficients matrix.

4. Example

To illustrate the performance of the derived sparse multinomial logistic regression classifiers we applied them to the data set *Cancer sites* considered in Vincent and Hansen (2014). It consists of bead-based expression data for $n = 162$ microRNAs with $d = 372$ features from $L = 18$ classes of normal and cancer issue samples. The number of samples in each class ranges from 5 to 26. Vincent and Hansen (2014) used sparse group Lasso classifier for this data.

We compared the performance of sparse group Slope with λ_j's and κ_ℓ's of the form given in (14), sparse group Lasso (replicating Vincent and Hansen, 2014), random forest

Classifier	Average misclass. error	# features	# non-zero coefficients
sparse group Slope	0.159 (0.019)	60-67	186-271
sparse group Lasso	0.165 (0.018)	51-79	382-592
random forest	0.209 (0.009)	-	-
XGBoost	0.250 (0.026)	-	-

Table 1: Average misclassification errors with their standard errors (in brackets) and feature selection for various classifiers.

and the well-known gradient boosting trees XGBoost classifiers on the above data set, where we developed the proximal gradient algorithm for solving sparse group Slope in (12) – see Appendix D.

To remove various technical variations, following Vincent and Hansen (2014), the data was first normalized by centering and scaling the rows of the design matrix, and then standardized by centering and scaling the columns. We split the data into training (75%) and test (25%) sets. The tuning parameters of all classification procedures were chosen by 10-fold cross-validation on the training set, and the misclassification errors of the resulting classifiers were measured on the test set. We repeated the process 10 times, randomly partitioning the data into train and test sets.

Table 1 presents the average (over 10 random splits) misclassification errors for the test sets, the numbers of selected features (non-zero rows of the regression coefficients matrix B) and the overall numbers of non-zero coefficients in B. It shows that both sparse multinomial logistic regression classifiers outperform their nonparametric counterparts for this data. Sparse group Slope yielded smaller misclassification errors than sparse group Lasso and, in addition, resulted in much sparser models.

5. Concluding remarks

In this paper we discussed high-dimensional multiclass classification by sparse multinomial logistic regression. Multiclass setup allows one to consider various types of sparsity associated with different assumptions on a matrix of regression coefficients. We proposed penalized MLE feature selection procedures with convex penalties capturing a specific type of sparsity at hand and showed that the resulting classifiers are optimal in the minimax sense. We presented the results for global row-wise, double row-wise and low-rank sparsity scenarios but one can consider also other related types of sparsity, e.g., group-sparsity, when features may have a group structure, or class-dependent sparsity, where each class has its own sparse subset of predictive features that implies column-wise sparsity, combinations of row-wise and low-rank sparsities, etc. The developed approach is general (see Appendix A.1 and Theorem 7 there) although a specific type of a penalty should be properly chosen w.r.t. a particular type of sparsity at hand.

In this paper we assume that $\mathbb{P}_X$ has a bounded support. Using a slightly different techniques, the main results remain valid also for Gaussian design (see Bellec et al., 2018; Alquier et al., 2019, for binary classification).

Even when the considered multinomial logistic regression model is misspecified and the Bayes classifier η^* is not linear, the misclassification excess risk can still be decomposed as

$$R(\widehat{\eta}_{\widehat{B}}) - R(\eta^*) = \left(R(\widehat{\eta}_{\widehat{B}}) - R(\eta^*_L)\right) + \left(R(\eta^*_L) - R(\eta^*)\right), \tag{21}$$

where $\eta^*_L = \arg\min_{\eta \in \mathcal{C}_L} R(\eta)$ is the best possible (oracle) linear classifier. The results of the paper can then be applied to the first term in the RHS of (21) representing the estimation error, whereas the approximation error in the second term measures the ability of linear classifiers to perform as good as η^*. Enriching the class of linear classifiers may improve the approximation error but will increase the resulting estimation error in (21). In a way, it is similar to the variance/bias tradeoff in regression.

Acknowledgments

The work was supported by the Israel Science Foundation (ISF), Grants ISF-589/18 and ISF-1095/22. The authors would like to thank Amir Beck, Guillaume Lecué and the anonymous referees for helpful comments.

Appendix A. Proofs of the upper bounds (Theorems 1, 3 and 6)

Throughout the proofs we use various generic positive constants, not necessarily the same each time they are used even within a single equation.

Throughout the proofs let $|\mathbf{a}|_2$ be the Euclidean norm of a vector $\mathbf{a}$, $||A||_2$ and $||A||_F$ respectively the operator/spectral and Frobenius norms of a matrix A. The Frobenius inner product of two matrices A_1 and A_2 is $\langle A_1, A_2 \rangle = tr(A_1^T A_2)$. Denote $||g(\boldsymbol{x})||_{L_2}$ for the L_2-norm of a function g and $||g(\boldsymbol{x})||_{L_2(\mathbb{P}_X)} = (\int_{\mathcal{X}} g(\boldsymbol{x})^2 d\mathbb{P}_X(\boldsymbol{x}))^{1/2}$ for the L_2-norm of g w.r.t. the measure $\mathbb{P}_X$. Recall that $V = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$.

A.1 Upper bounds for misclassification excess risk for a general penalized MLE plug-in linear classifier

Consider first a generic setup. Let $\mathcal{M} = \{B \in \mathbb{R}^{d\times L} : B\mathbf{1} = \mathbf{0}\}$ be the set of regression matrices satisfying the symmetric constraint and $\mathcal{M}_0 \subseteq \mathcal{M}$ be its subset of sparse matrices, where the notion of sparsity depends on the particular problem at hand. Let $B \in \mathcal{M}_0$ and consider a penalized MLE estimator $\widehat{B}$ of the form

$$\widehat{B} = \arg\min_{\widetilde{B}\in\mathcal{M}} \left\{-l(\widetilde{B}) + ||\widetilde{B}||\right\}, \tag{22}$$

where the regularized matrix norm $||\cdot||$ induces the given type of sparsity, and the corresponding plug-in linear classifier

$$\widehat{\eta}_{\widehat{B}}(\boldsymbol{x}) = \arg\max_{1\le l\le L} \widehat{\boldsymbol{\beta}}_l^T \boldsymbol{x}.$$

The Kullback-Leibler divergence between two multinomial distributions with probabilities vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ is $KL(\mathbf{p}_1, \mathbf{p}_2) = \sum_{l=1}^L p_{1l} \ln\left(\frac{p_{1l}}{p_{2l}}\right)$. Let $f_B(\boldsymbol{x}, y)$ be the joint distribution of $(\boldsymbol{X}, Y)$, i.e., $df_B(\boldsymbol{x}, y) = \prod_{l=1}^L p_l(\boldsymbol{x})^{\xi_l} d\mathbb{P}_X(\boldsymbol{x})$, where $p_l(\boldsymbol{x})$ are given in (2). For two given regression coefficients matrices B_1 and B_2 the Kullback-Leibler divergence between the distributions f_{B_1} and f_{B_2} is then $d_{KL}(f_{B_1}, f_{B_2}) = \int KL(\mathbf{p}_1(\boldsymbol{x}), \mathbf{p}_2(\boldsymbol{x}))d\mathbb{P}_X(\boldsymbol{x})$. We exploit the well-known result (e.g., Pires and Szepesvári, 2016; Abramovich et al., 2021) that relates the misclassification excess risk $\mathcal{E}(\hat{\eta}, \eta^*)$ and the Kullback-Leibler risk $\mathbb{E}d_{KL}(f_B, f_{\widehat{B}})$ under the low-noise Assumption A:

$$\mathcal{E}(\widehat{\eta}_{\widehat{M}}, \eta^*) \le C\left(\mathbb{E}d_{KL}(f_B, f_{\widehat{B}})\right)^{\frac{\alpha+1}{\alpha+2}}. \tag{23}$$

We now extend the results of Alquier et al. (2019) for univariate response to multivariate (multinomial) $\mathbf{Y}$ to bound the Kullback-Leibler risk $\mathbb{E}d^2_{KL}(f_B, f_{\widehat{B}})$. Define $\theta_l(\boldsymbol{x}) = \boldsymbol{\beta}_l^T\boldsymbol{x},\ l = 1, \ldots, L$, where due to the symmetric constraint, $\sum_{l=1}^L \theta_l(\boldsymbol{x}) = 0$. It is easy to verify that in terms of θ_l's, the multinomial log-likelihood is Lipschitz w.r.t. the l_2-norm. Furthermore, for $\mathbb{P}_X$ with a bounded support, $|\boldsymbol{\beta}_l^T\boldsymbol{x}| \le C$ and $d_{KL}(\cdot,\cdot)$ is strongly convex (Abramovich et al., 2021): for any two matrices B_1 and B_2 satisfying the symmetric constraint, $d_{KL}(f_{B_1}, f_{B_2}) \ge C\sum_{l=1}^L ||\theta_{1l}(\boldsymbol{x}) - \theta_{2l}(\boldsymbol{x})||^2_{L_2(\mathbb{P}_X)}$ (the multivariate analogue of Bernstein condition in terminology of (Alquier et al., 2019). These two conditions allow us to adopt the general approach of Alquier et al. (2019) to bound $\mathbb{E}d^2_{KL}(f_B, f_{\widehat{B}})$.

Define the following quantities along the lines of Alquier et al. (2019). Let $\mathcal{B}_{||\cdot||} = \{B \in \mathcal{M} : ||B|| \le 1\}$ be the unit ball of matrices satisfying the symmetric constraint w.r.t. $||\cdot||$-norm in (22). Let $\widehat{Rad}(\mathcal{B}_{||\cdot||})$ be the empirical (multivariate) Rademacher complexity of

$\mathcal{B}_{||\cdot||}$, namely,

$$\widehat{Rad}(\mathcal{B}_{||\cdot||}) = \mathbb{E}_\Sigma \left\{ \frac{1}{\sqrt{n}} \sup_{B \in \mathcal{B}} \sum_{i=1}^{n} \sum_{l=1}^{L} \sigma_{il} \boldsymbol{\beta}_l^T \boldsymbol{X}_i \middle| \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n \right\}$$
$$= \mathbb{E}_\Sigma \left\{ \frac{1}{\sqrt{n}} \sup_{B \in \mathcal{B}} tr(\Sigma B^T X^T) \right\},$$

where the elements σ_{il}'s of $\Sigma \in \mathbb{R}^{n \times L}$ are i.i.d. Rademacher random variables with $P(\sigma_{il} = 1) = P(\sigma_{il} = -1) = 1/2$, and

$$Rad(\mathcal{B}_{||\cdot||}) = \mathbb{E}_X \left\{ \widehat{Rad}(\mathcal{B}_{||\cdot||}) \right\}$$

be the Rademacher complexity of $\mathcal{B}$.

Define a *complexity function*

$$r(\rho) = \sqrt{\frac{C_0 Rad(\mathcal{B}_{||\cdot||})\rho}{2R^2\sqrt{n}}}, \quad \rho > 0,$$

where the exact value of $C_0 > 0$ is specified in Alquier et al. (2019).

Let $\mathcal{T}(\rho) = \{B' \in \mathcal{M} : ||B'|| = \rho,\ ||V^{\frac{1}{2}} B'||_F^2 \leq r^2(2\rho)\}$. For a given matrix $B \in \mathcal{M}_0$ define $\Gamma_B(\rho) = \bigcup_{B' : ||B' - B|| < \frac{\rho}{20}} \partial || \cdot ||(B')$, where the subdifferential $\partial || \cdot ||(B') = \{G \in \mathcal{M} : ||B' + B''|| - ||B'|| \geq \langle B'', G \rangle,\ \forall B'' \in \mathcal{M}\}$. The *sparsity parameter* is

$$\Delta(\rho) = \inf_{B' \in \mathcal{T}(\rho)} \sup_{G \in \Gamma_B(\rho)} \langle B', G \rangle.$$

Finally, let ρ^* be any solution of the sparsity inequality

$$\Delta(\rho^*) \geq \frac{4}{5}\rho^* \tag{24}$$

The quantity ρ^* depends on a particular norm in (22) and the second moment matrix V, and plays a key role in establishing the upper bound for $Ed_{KL}(f_{B_1}, f_{B_2})$.

We have the following generic theorem:

Theorem 7 *Let $\widehat{B}$ be the solution of (22). Assume that there exists ρ^* such that $\Delta(\rho^*) \geq \frac{4}{5}\rho^*$ and $Rad(\mathcal{B}_{||\cdot||}) \leq \frac{7}{720}\sqrt{n}$. Then,*

$$\mathbb{E} d_{KL}(f_B, f_{\widehat{B}}) \leq C\rho^*, \tag{25}$$

for some $C > 0$.

In addition,

$$\mathbb{E}||V^{\frac{1}{2}}(\widehat{B} - B)||_F^2 \leq C_1 \rho^*$$

and

$$\mathbb{E}\|\widehat{B} - B\| \leq C_2\rho^*$$

for some $C_1, C_2 > 0$.

Theorem 7 is an extension of Theorem 2.2 (or more general Theorem 9.2) of Alquier et al. (2019) for multivariate response and anisotropic design. Its proof repeats the proof of Lemma 1 in Abramovich et al. (2021) with the particular group Slope norm considered there replaced by a general norm $||\cdot||$.

Remark 8 In fact, from the definition of the sparsity parameter $\Delta(\rho)$ and ρ^* it follows that Theorem 7 holds even if the true regression matrix B is only "approximately sparse" in the sense that there exists a sparse matrix $B' \in \mathcal{M}_0$ such that $||B - B'|| \leq \rho^*/20$ (see also Alquier et al., 2019).

We will now apply the general upper bound (25) for the group Slope, sparse group Slope and nuclear norms to complete the proofs of Theorems 1, 3 and 6 by finding the corresponding $Rad(\mathcal{B}_{||\cdot||})$ and ρ^*.

A.2 Proof of Theorem 1

The proof of Theorem 1 is somewhat different from that of Theorem 4 of Abramovich et al. (2021) for isotropic $\boldsymbol{X}$.

For given $\lambda_1 \geq \ldots \geq \lambda_d$ consider the group Slope norm $||B||_\lambda = \sum_{j=1}^d \lambda_j |B|_{(j)}$. Let $B \in \mathcal{M}(d_0)$ with a set of zero rows $\mathcal{J}(B)$ and $B' \in \mathbb{R}^{d\times L}$ such that $\|B' - B\|_\lambda = \rho^*$ and $\|V^{\frac{1}{2}}(B' - B)\|_F^2 \leq \frac{C_0 Rad(\mathcal{B}_\lambda)\rho^*}{R^2\sqrt{n}}$, where ρ^* will be defined later .

Let $\mathcal{G}$ be a set all of matrices of the form $\sum_{j\in\mathcal{J}(B)} \lambda_{\pi(j)} \mathbf{e}_j \frac{B_{j\cdot}}{|B_{j\cdot}|_2} + \sum_{j\in\mathcal{J}^c(B)} \lambda_{\pi(j)} \mathbf{e}_j \mathbf{v}_j^T$, where $\pi = (\pi(1), \ldots, \pi(d))$ is a permutation of $\{1, \ldots, d\}$ and $\mathbf{v}_j$'s are unit vectors in $\mathbb{R}^L$, and note that $\|B\|_\lambda = \max_{G\in\mathcal{G}}\langle B, G\rangle$.

In particular, $\partial\|\cdot\|_\lambda(B) \supseteq \text{argmax}_{G\in\mathcal{G}}\langle B, G\rangle$. Hence, we can find a permutation of $\{\lambda_j\}_{j=d_0+1}^d$ such that the corresponding $G \in \text{argmax}_{G\in\mathcal{G}}\langle B, G\rangle \subseteq \partial\|\cdot\|_\lambda(B)$ and $\sum_{j\in\mathcal{J}^c(B)} G_{\cdot j}^T(B'-$

$B)_{j\cdot} \geq \sum_{j=d_0+1}^{d} \lambda_j |B' - B|_{(j)}$. Then,

$$\begin{aligned} \langle G, B' - B \rangle &= \sum_{j \in \mathcal{J}(B)} G_{\cdot j}^T (B' - B)_{j\cdot} + \sum_{j \in \mathcal{J}^c(B)} G_{\cdot j}^T (B' - B)_{j\cdot} \\ &\geq \sum_{j \in \mathcal{J}^c(B)} G_{\cdot j}^T (B' - B)_{j\cdot} - \sum_{j=1}^{d_0} \lambda_j |B' - B|_{(j)} \geq \\ &\geq \sum_{j=1}^{d} \lambda_j |B' - B|_{(j)} - 2\sum_{j=1}^{d_0} \lambda_j |B' - B|_{(j)} \\ &= \rho^* - 2\sum_{j=1}^{d_0} \lambda_j |B' - B|_{(j)}. \end{aligned} \tag{26}$$

By Assumption B_1,

$$\frac{1}{\nu_{gS}(d_0)} \|V^{\frac{1}{2}}(B' - B)\|_F^2 \geq \|\Pi_{d_0}(B' - B)\|_F^2 = \sum_{j=1}^{d_0} |B' - B|_{(j)}^2 \ . \tag{27}$$

For any $1 \leq j \leq d_0$ we also have

$$\sum_{j'=1}^{d_0} |B' - B|_{(j')}^2 \geq \sum_{j'=1}^{j} |B' - B|_{(j')}^2 \geq j |B' - B|_{(j)}^2$$

and, therefore, $|B' - B|_{(j)} \leq \sqrt{\sum_{j'=1}^{d_0} |B' - B|_{(j')}^2} / \sqrt{j}$.

Taking

$$\rho^* = \frac{100 C_0 C}{\nu_{gS}(d_0)} \frac{Rad(\mathcal{B}_\lambda) \left(\sum_{j=1}^{d_0} \lambda_j / \sqrt{j} \right)^2}{\sqrt{n}}$$

(27) implies

$$\begin{aligned} \sum_{j=1}^{d_0} \lambda_j |B' - B|_{(j)} &\leq \frac{1}{\sqrt{\nu_{gS}(d_0)}} \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right) \|V^{\frac{1}{2}}(B' - B)\|_F \\ &\leq \frac{1}{\sqrt{\nu_{gS}(d_0)}} \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right) \sqrt{\frac{C_0 Rad(\mathcal{B}_\lambda) \rho^*}{R^2 \sqrt{n}}} \\ &\leq \frac{1}{10} \rho^*. \end{aligned}$$

Thus, combining with (26)

$$\langle G, B' - B \rangle \geq \frac{4}{5} \rho^*$$

for every $B' - B \in \mathcal{T}(\rho^*)$ and, therefore,

$$\Delta(\rho^*) \geq \frac{4}{5} \rho^*.$$

Furthermore, by Lemma 2 of Abramovich et al. (2021),

$$Rad(\mathcal{B}_\lambda) \leq C \max_{1\leq j\leq d} \frac{\sqrt{L+\ln(d/j)}}{\lambda_j}$$

for some $C>0$. Hence, for λ_j satisfying (10), $Rad(\mathcal{B}_\lambda) \leq \frac{7}{720}\sqrt{n}$ and we can apply Theorem 7 to complete the proof.

A.3 Proof of Theorem 3

For given $\lambda_1 \geq \ldots \geq \lambda_d > 0$ and $\kappa_1 \geq \ldots \geq \kappa_L > 0$, consider the sparse group Slope norm $||B||_{\kappa,\lambda} = \sum_{j=1}^d \lambda_j |B|_{(j)} + \sum_{j=1}^d \sum_{l=1}^L \kappa_l |B|_{j(l)}$, where $|B|_{(1)} \geq \ldots \geq |B|_{(d)}$ are the descendingly ordered l_2-norms of the rows of B and $|B|_{j(1)} \geq \ldots \geq B_{j(L)}$ are descendingly ordered absolute values of entries of its rows. Let $\mathcal{B}_{\kappa,\lambda}$ be the unit ball of matrices w.r.t. this norm.

Lemma 9 *Let $B \in \mathcal{M}(d_0, \mathbf{m})$. Under Assumption B_1, define*

$$\rho^* = \frac{100 C_0 C}{\nu_{gS}(d_0)} \frac{Rad(\mathcal{B}_{\kappa,\lambda}) \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} + \sqrt{\sum_{j=1}^{d_0} \left(\sum_{l=1}^{m_j} \frac{\kappa_l}{\sqrt{l}} \right)^2} \right)^2}{\sqrt{n}}. \tag{28}$$

Then, ρ^* satisfies the sparsity inequality (24), i.e., $\Delta(\rho^*) \geq \frac{4}{5}\rho^*$.

To apply Theorem 7 to complete the proof, we need also to show that $Rad(\mathcal{B}_{\kappa,\lambda}) \leq \frac{7}{720}\sqrt{n}$:

Lemma 10 *Let $\kappa_L \geq \sqrt{\frac{\pi}{2}} \frac{2880}{7} \frac{1}{\sqrt{n}}$. Then,*

$$Rad(\mathcal{B}_{\kappa,\lambda}) \leq \frac{7}{1440}\sqrt{n} + C_0 \sqrt{\frac{\pi}{2}} \max_{1\leq j\leq d} \frac{\sqrt{2\sum_{j=1}^L \frac{1}{l} \left(\frac{Le}{l}\right)^l e^{-C^2 n l \kappa_l^2} + 2\log\left(\frac{de}{j}\right)}}{\lambda_j},$$

where $C = \sqrt{\frac{2}{\pi}} \frac{7}{2880}$ and $C_0 > 0$ is given in the proof.

In particular, for λ_j's and κ_l's satisfying (13), $Rad(\mathcal{B}_{\kappa,\lambda}) \leq \frac{7}{720}\sqrt{n}$.

A.4 Proof of Theorem 6

Let $||B||_\lambda = \lambda ||B||_*$ and $\mathcal{B}_\lambda$ the corresponding unit ball. Define

$$\rho^* = 100\lambda \frac{C_0 r_0 Rad(\mathcal{B}_\lambda)}{2R^2 \tau_d(V)\sqrt{n}}.$$

Extending Lemma 4.4 of Lecué and Mendelson (2018) for the anisotropic case by using $\|B\|_* < \frac{1}{\sqrt{\tau_d(V)}} \|V^{\frac{1}{2}} B\|_*$, we have $\Delta(\rho^*) \geq \frac{4}{5}\rho^*$.

To apply Theorem 7 we need to show that for λ in (18), $Rad(\mathcal{B}_\lambda) \leq \frac{7}{720}\sqrt{n}$:

Lemma 11

$$Rad(\mathcal{B}_\lambda) \le C_0 \sqrt{\tau_1(V)}\ \frac{\sqrt{L-1}+\sqrt{d}}{\lambda}$$

for some $C_0 > 0$.

Thus, taking $C = \frac{720C_0}{7}$, the choice of $\lambda = \sqrt{\tau_1(V)}\ \frac{(\sqrt{L-1}+\sqrt{d})}{\sqrt{n}}$ implies $Rad(\mathcal{B}_\lambda) \le \frac{7}{720}\sqrt{n}$.

Appendix B. Proofs of lemmas

B.1 Proof of Lemma 9

We use the arguments similar to those in the proof of Theorem 1.

Let $\mathcal{J}$ be the set of indices of non-zero rows of B and $\mathcal{L}_j$ be the set of indices of non-zero entries of the j-th row for $j \in \mathcal{J}$. Obviously, $|\mathcal{J}| = d_0$ and $|\mathcal{L}_j| = m_j$. Consider a matrix B' such that $\|B'-B\|_{\kappa,\lambda} = \rho^*$ and $\|V^{\frac{1}{2}}(B'-B)\|_F^2 \le r^2(2\rho^*) = \frac{C_0 Rad(\mathcal{B})C}{\sqrt{n}}\rho^*$.

We can decompose $\|B\|_{\kappa,\lambda}$ into two additive components: $\|B\|_{\kappa,0} = \sum_{j=1}^d \sum_{l=1}^L \kappa_l |B|_{j(l)}$ and $\|B\|_{0,\lambda} = \sum_{j=1}^d \lambda_j |B|_{(j)}$.

Define two matrices $G, H \in \mathbb{R}^{d\times L}$ as follows. For every $j \in \mathcal{J}$ let $\pi_j(1), \ldots, \pi_j(m_j)$ be the indices of descendingly ordered nonzero entries $|B|_{j(l)}$'s and set $G_{j\pi_j(l)} = \kappa_{\pi_j(l)}\,\mathrm{sign}(B_{j\pi_j(l)})$. Similarly, let $\tilde{\pi}(1), \ldots, \tilde{\pi}(d_0)$ be the indices of descendingly ordered Euclidean norms $|B|_{(j)}$ of d_0 nonzero rows of B and set $H_{jl} = \lambda_{\tilde{\pi}(j)} \frac{B_{\tilde{\pi}(j)l}}{|B_{\tilde{\pi}(j)\cdot}|_2}$. The entries of G and H corresponding to zero entries of B will be defined later.

By construction, $tr(G^T B) = \|B\|_{\kappa,0}$ and $tr(H^T B) = \|B\|_{0,\lambda}$, while for any B', $tr(G^T B') \le \|B'\|_{\kappa,0}$ and $tr(H^T B') \le \|B'\|_{0,\lambda}$. Thus, G and H are in $\partial\|\cdot\|_{\kappa,0}(B)$ and $\partial\|\cdot\|_{0,\lambda}(B)$ respectively.

We have

$$\sum_{j=1}^d \sum_{l\in\mathcal{L}_j} G_{jl}\left|B'_{jl}-B_{jl}\right| \le \sum_{j=1}^d \sum_{l=1}^{m_j} \kappa_l \left|B'-B\right|_{j(l)}$$

and

$$\sum_{j\in\mathcal{J}} \left|\sum_{l=1}^L H_{jl}\left(B'_{jl}-B_{jl}\right)\right| \le \sum_{j=1}^{d_0} \lambda_j \left|B'-B\right|_{(j)}.$$

Hence,

$$\begin{aligned} tr\left(G^T\left(B'-B\right)\right) &= \sum_{j=1}^d \sum_{l\in\mathcal{L}_j} G_{jl}\left(B'_{jl}-B_{jl}\right) + \sum_{j=1}^d \sum_{l\in\mathcal{L}_j^C} G_{jl}\left(B'_{jl}-B_{jl}\right) \\ &\ge \sum_{j=1}^d \sum_{l\in\mathcal{L}_j^C} G_{jl}\left(B'_{jl}-B_{jl}\right) - \sum_{j=1}^d \sum_{l=1}^{m_j} \kappa_l \left|B'-B\right|_{j(l)}, \end{aligned} \tag{29}$$

and

$$\begin{aligned} tr\left(H^T\left(B'-B\right)\right) &= \sum_{j\in\mathcal{J}}\sum_{l=1}^{L} H_{jl}\left(B'_{jl}-B_{jl}\right)+\sum_{j\in\mathcal{J}^C}\sum_{l=1}^{L} H_{jl}\left(B'_{jl}-B_{jl}\right) \\ &\geq \sum_{j\in\mathcal{J}^C}\sum_{l=1}^{L} H_{jl}\left(B'_{jl}-B_{jl}\right)-\sum_{j=1}^{d_0}\lambda_j\left|B'-B\right|_{(j)} \end{aligned} \tag{30}$$

To bound the first terms of the RHSs in (29) and (30) from below for a given B' complete the entries of G and H corresponding to zero entries of B in such a way that

$$\sum_{j=1}^{d}\sum_{l\in\mathcal{L}_j^C} G_{lj}\left(B'_{lj}-B_{lj}\right)\geq\sum_{j=1}^{d}\sum_{l=m_j+1}^{L}\kappa_l\left|B'-B\right|_{j(l)},$$

and

$$\sum_{j\in\mathcal{J}^C}\sum_{l=1}^{L} H_{lj}\left(B'_{lj}-B_{lj}\right)\geq\sum_{j=d_0+1}^{d}\lambda_j\left|B'-B\right|_{(j)}.$$

Thus,

$$\begin{aligned} tr\left(G^T\left(B'-B\right)\right) &\geq \sum_{j=1}^{d}\sum_{l=1}^{L}\kappa_l\left|B'-B\right|_{j(l)}-2\sum_{j=1}^{d}\sum_{l=1}^{m_j}\kappa_l\left|B'-B\right|_{j(l)} \\ &= \|B'-B\|_{\kappa,0}-2\sum_{j=1}^{d}\sum_{l=1}^{m_j}\kappa_l\left|B'-B\right|_{j(l)}, \end{aligned}$$

and

$$\begin{aligned} tr\left(H^T\left(B'-B\right)\right) &\geq \sum_{j=1}^{d}\lambda_j\left|B'-B\right|_{(j)}-2\sum_{j=1}^{d_0}\lambda_j\left|B'-B\right|_{(j)} \\ &= \|B'-B\|_{0,\lambda}-2\sum_{j=1}^{d_0}\lambda_j\left|B'-B\right|_{(j)}. \end{aligned}$$

Consider $Z=G+H$. Evidently, $Z\in\partial\|\cdot\|_{\kappa,\lambda}(B)$ and

$$\begin{aligned} tr\left(Z^T\left(B'-B\right)\right) &\geq \|B'-B\|_{\kappa,\lambda}-2\sum_{j=1}^{d_0}\lambda_j\left|B'-B\right|_{(j)}-2\sum_{j=1}^{d}\sum_{l=1}^{m_j}\kappa_l\left|B'-B\right|_{j(l)} \\ &= \rho^*-2\sum_{j=1}^{d_0}\lambda_j\left|B'-B\right|_{(j)}-2\sum_{j=1}^{d}\sum_{l=1}^{m_j}\kappa_l\left|B'-B\right|_{j(l)}. \end{aligned}$$

By Assumption B_1,

$$\frac{1}{\nu_{gS}(d_0)}\|V^{\frac{1}{2}}(B'-B)\|_F^2\geq\|\Pi_{d_0}(B'-B)\|_F^2=\sum_{j=1}^{d_0}|B'-B|_{(j)}^2\geq\sum_{j\in\mathcal{J}}\sum_{l=1}^{m_j}|B'-B|_{j(l)}^2\ .$$

Since $|B'-B|_{(j)} \leq \sqrt{\sum_{j'=1}^{d_0} |B'-B|^2_{(j')}}/\sqrt{j}$,

$$\sum_{j=1}^{d_0} \lambda_j \left|B'-B\right|_{(j)} \leq \frac{1}{\nu_{gS}(d_0)} \sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \|V^{1/2}(B'-B)\|_F.$$

On the other hand, by the Cauchy-Schwartz inequality,

$$\begin{aligned}
\sum_{j=1}^{d} \sum_{l=1}^{m_j} \kappa_l \left|B'-B\right|_{j(l)} &\leq \sum_{j\in\mathcal{J}} \left|(B'-B)\right|_j \sum_{l=1}^{m_j} \frac{\kappa_l}{\sqrt{l}} \\
&\leq \sqrt{\sum_{j=1}^{d}\left(\sum_{l=1}^{m_j} \frac{\kappa_l}{\sqrt{l}}\right)^2} \sqrt{\sum_{j'=1}^{d_0} |B'-B|^2_{(j')}} \\
&\leq \frac{1}{\nu_{gS}(d_0)} \sqrt{\sum_{j=1}^{d}\left(\sum_{l=1}^{m_j} \frac{\kappa_l}{\sqrt{l}}\right)^2} \|V^{1/2}(B'-B)\|_F
\end{aligned}$$

Thus, for any $B-B' \in \mathcal{T}(\rho^*)$, we found $Z \in \partial\|\cdot\|_{\kappa,\lambda}(B)$ such that

$$tr\left(Z^T\left(B'-B\right)\right) \geq \rho^* - 2\frac{1}{\nu_{gS}(d_0)} \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} + \sqrt{\sum_{j=1}^{d}\left(\sum_{l=1}^{m_j} \frac{\kappa_l}{\sqrt{l}}\right)^2}\right) \|V^{1/2}(B'-B)\|_F.$$

Hence,

$$\begin{aligned}
\Delta(\rho^*) &= \inf_{B''\in\mathcal{T}(\rho^*)} \sup_{Z\in\partial\|\cdot\|_{\kappa,\lambda}(B)} tr\left(Z^T B''\right) \\
&\geq \rho^* - 2\left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} + \sqrt{\sum_{j=1}^{d}\left(\sum_{l=1}^{m_j} \frac{\kappa_l}{\sqrt{l}}\right)^2}\right) \sqrt{\frac{C_0 Rad(\mathcal{B}_{\lambda,\kappa})C}{\nu_{gS}(d_0)\sqrt{n}}\rho^*}.
\end{aligned}$$

and, therefore, for ρ^* from (28), $\Delta(\rho^*) \geq \frac{4}{5}\rho^*$.

B.2 Proof of Lemma 10

To prove Lemma 10 we first bound the empirical Rademacher complexity $\widehat{Rad}(\mathcal{B}_{\kappa,\lambda})$. As a first step, we bound the empirical Rademacher complexity by the empirical Gaussian complexity

$$\widehat{G}(\mathcal{B}_{\kappa,\lambda}) = \mathbb{E}_G\left\{\frac{1}{\sqrt{n}} \sup_{B\in\mathcal{B}_{\kappa,\lambda}} \sum_{i=1}^{n}\sum_{l=1}^{L} G_{il}\boldsymbol{\beta}_l^T \boldsymbol{X}_i \middle| \boldsymbol{X}_1 = \boldsymbol{x}_1,\ldots,\boldsymbol{X}_n = \boldsymbol{x}_n\right\} = \mathbb{E}_G\left\{\frac{1}{\sqrt{n}} \sup_{B\in\mathcal{B}_{\kappa,\lambda}} tr(B^T Z)\right\},$$

where G_{il} are i.i.d. $N(0,1)$ and $Z = X^T G$. We have $\widehat{Rad}(\mathcal{B}_{\kappa,\lambda}) \leq \sqrt{\frac{\pi}{2}}\widehat{G}(\mathcal{B}_{\kappa,\lambda})$ (see, e.g., Wainwright, 2019, Section 5.2).

Define

$$\delta_j = \sqrt{\sum_{l=1}^{L} \left(\left|Z_{j(l)}\right| - \sqrt{\frac{2}{\pi}} \frac{7}{1440} |X|_{2j} \ \kappa_l \right)_+^2}, \quad j = 1, \ldots, d.$$

To bound $\widehat{G}(\mathcal{B}_{\kappa,\lambda})$ we need the following two lemmas:

Lemma 12

$$\widehat{G}(\mathcal{B}_{\kappa,\lambda}) \le \sqrt{\frac{2}{\pi}} \frac{7}{1440} \max_{1 \le j \le d} |X_{\cdot j}|_2 + \mathbb{E}_G \max_{1 \le j \le d} \frac{\delta_{(j)}}{\lambda_j}.$$

Lemma 13 *Let* $\kappa_L \ge \sqrt{\frac{\pi}{2}} \frac{2880}{7\sqrt{n}}$. *Then, conditionally on* X,

$$\mathbb{E}_G \max_{1 \le j \le d} \frac{\delta_{(j)}}{\lambda_j} \le C_0 \max_{1 \le j \le d} \left\{ \frac{1}{\sqrt{n}} |X|_{2j} \ \frac{\sqrt{2 \sum_{j=1}^{L} \frac{1}{l} \left(\frac{Le}{l}\right)^l e^{-C^2 n l \kappa_l^2} + 2 \log\left(\frac{de}{j}\right)}}{\lambda_j} \right\},$$

where $C = \sqrt{\frac{2}{\pi}} \frac{7}{2880}$ *and* $C_0 = 2\,(1 + \sqrt{\pi})$.

Lemmas 12 and 13 together imply

$$\widehat{Rad}(\mathcal{B}_{\kappa,\lambda}) \le \left(\frac{7}{1440} \sqrt{n} + C_0 \sqrt{\frac{\pi}{2}} \max_{1 \le j \le d} \frac{\sqrt{2 \sum_{j=1}^{L} \frac{1}{l} \left(\frac{Le}{l}\right)^l e^{-C^2 n l \kappa_l^2} + 2 \log\left(\frac{de}{j}\right)}}{\lambda_j} \right) \max_{1 \le j \le d} \frac{1}{\sqrt{n}} |X_{\cdot j}|_2.$$

Hence,

$$\begin{aligned} Rad(\mathcal{B}_{\kappa,\lambda}) &= \mathbb{E}_X \left\{ \widehat{Rad}(\mathcal{B}_{\kappa,\lambda}) \right\} \\ &\le \frac{7}{1440} \sqrt{n} + C_0 \sqrt{\frac{\pi}{2}} \max_{1 \le j \le d} \frac{\sqrt{8 \sum_{j=1}^{L} \frac{1}{l} \left(\frac{Le}{l}\right)^l e^{-C^2 n l \kappa_l^2} + 2 \log\left(\frac{de}{j}\right)}}{\lambda_j}. \end{aligned}$$

Proof of Lemma 12

Define two unit balls w.r.t. $||\cdot||_{\kappa,0}$ and $||\cdot||_{0,\lambda}$: $\mathcal{B}_\kappa = \left\{ B : \sum_{j=1}^{d} \sum_{l=1}^{L} \kappa_l \left|B_{j(l)}\right| \le 1 \right\}$ and $\mathcal{B}_\lambda = \left\{ B : \sum_{j=1}^{d} \lambda_j |B|_{(j)} \le 1 \right\}$ and note that $\mathcal{B}_{\kappa,\lambda} \subseteq \mathcal{B}_\kappa \cap \mathcal{B}_\lambda$.

For any matrix $A \in \mathbb{R}^{d \times L}$ we have

$$\begin{aligned} E_G \sup_{B \in \mathcal{B}_{\kappa,\lambda}} \langle Z, B \rangle \le \mathbb{E}_G \sup_{B \in \mathcal{B}_\kappa \cap \mathcal{B}_\lambda} \langle Z, B \rangle \ &= \ \mathbb{E}_G \sup_{B \in \mathcal{B}_\kappa \cap \mathcal{B}_\lambda} \{ \langle A, B \rangle + \langle Z - A, B \rangle \} \\ &\le \mathbb{E}_G \left\{ \sup_{B \in \mathcal{B}_\kappa \cap \mathcal{B}_\lambda} \langle A, B \rangle + \sup_{B \in \mathcal{B}_\kappa \cap \mathcal{B}_\lambda} \langle Z - A, B \rangle \right\} \\ &\le \mathbb{E}_G \left\{ \sup_{B \in \mathcal{B}_\kappa} \langle A, B \rangle + \sup_{B \in \mathcal{B}_\lambda} \langle Z - A, B \rangle \right\}. \end{aligned}$$

Similar to the results for the group Slope of Abramovich et al. (2021), $\sup_{B\in\mathcal{B}_\kappa}\langle A,B\rangle \le \max_{jl}\frac{|A_{j(l)}|}{\kappa_l}$ and $\sup_{B\in\mathcal{B}_\lambda}\langle Z-A,B\rangle \le \max_j \frac{|Z-A|_{(j)}}{\lambda_j}$. Thus,

$$E\sup_{B\in\mathcal{B}_{\kappa,\lambda}}\langle Z,B\rangle \le \mathbb{E}\left\{\max_{j,l}\frac{|A_{j(l)}|}{\kappa_l}+\max_j\frac{|Z-A|_{(j)}}{\lambda_j}\right\}.$$

In particular, consider a matrix A such that $A_{j(l)} = \text{sign}\left(Z_{j(l)}\right)\min\left\{\left|Z_{j(l)}\right|, \sqrt{\frac{2}{\pi}}\frac{7}{1440}|X_{\cdot j}|_2\kappa_l\right\}$. We then have

$$E\sup_{B\in\mathcal{B}_{\kappa,\lambda}}\langle Z,B\rangle \le \sqrt{\frac{2}{\pi}}\frac{7}{1440}\max_{1\le j\le d}|X_{\cdot j}|_2+\mathbb{E}\left\{\max_{1\le j\le d}\frac{\left(\sqrt{\sum_{l=1}^L\left(\left|Z_{j(l)}\right|-\sqrt{\frac{2}{\pi}}\frac{7}{1440}|X_{\cdot j}|_2\ \kappa_l\right)_+^2}\right)_{(j)}}{\lambda_j}\right\}.$$

Proof of Lemma 13

Denoting $C=\frac{1}{\sqrt{2\pi}}\frac{7}{1440}$, we have

$$\begin{aligned}\mathbb{E}_G\left\{\left(\frac{|Z_{j(l)}|}{\frac{1}{\sqrt{n}}|X_{\cdot j}|_2}-2C\sqrt{n}\kappa_l\right)_+^2\right\} &= \int_0^\infty 2sP\left(\left(\frac{|Z_{j(l)}|}{\frac{1}{\sqrt{n}}|X_{\cdot j}|_2}-2C\sqrt{n}\kappa_l\right)_+^2 > s^2\right)ds\\ &\le \int_0^\infty 2sP\left(\frac{|Z_{j(l)}|}{\frac{1}{\sqrt{n}}|X_{\cdot j}|_2} > s+2C\sqrt{n}\kappa_l\right)ds \qquad (31)\\ &\le \int_0^\infty 2s\binom{L}{l}P\left(\frac{|Z_{jl}|}{\frac{1}{\sqrt{n}}|X_{\cdot j}|_2} > s+2C\sqrt{n}\kappa_l\right)^l ds.\end{aligned}$$

Note that conditionally on X, $\frac{Z_{jl}}{\frac{1}{\sqrt{n}}|X_{\cdot j}|_2}$ is an $\mathcal{N}(0,1)$ Gaussian random variable and, therefore, (31) yields

$$\begin{aligned}\mathbb{E}_G\left\{\left(\frac{|Z_{j(l)}|}{\frac{1}{\sqrt{n}}|X_{\cdot j}|_2}-2C\sqrt{n}\kappa_l\right)_+^2\right\} &\le \int_0^\infty 2^{l+1}s\binom{L}{l}e^{-\frac{l\left(s+2C\sqrt{n}\kappa_l\right)^2}{2}}ds\\ &\le \int_0^\infty 2^{l+1}\frac{1}{l}\binom{L}{l}l\left(s+2C\sqrt{n}\kappa_l\right)e^{-\frac{l\left(s+2C\sqrt{n}\kappa_l\right)^2}{2}}ds \qquad (32)\\ &= 2^{l+1}\frac{1}{l}\binom{L}{l}e^{-2C^2nl\kappa_l^2} \le 2^{l+1}\frac{1}{l}\left(\frac{Le}{l}\right)^l e^{-2C^2nl\kappa_l^2}.\end{aligned}$$

For $\kappa_l \ge \frac{1}{C\sqrt{n}}$, (32) implies

$$\mathbb{E}\left\{\left(\left|Z_{j(l)}\right|-2C|X_{\cdot j}|_2\kappa_l\right)_+^2\right\} \le 2\frac{1}{l}\left(\frac{Le}{l}\right)^l e^{-C^2nl\kappa_l^2}\ \frac{1}{n}|X_{\cdot j}|_2^2$$

Hence, by Jensen inequality,

$$\mathbb{E}\sqrt{\sum_{l=1}^{L}\left(\left|Z_{j(l)}\right|-2C|X_{\cdot j}|_2\kappa_l\right)_+^2} \leq \sqrt{\mathbb{E}\left[\sum_{l=1}^{L}\left(\left|Z_{j(l)}\right|-2C|X_{\cdot j}|_2\kappa_l\right)_+^2\right]}$$
$$\leq \sqrt{2\sum_{j=1}^{L}\frac{1}{l}\left(\frac{Le}{l}\right)^l e^{-C^2 nl\kappa_l^2}}\ \frac{1}{\sqrt{n}}|X_{\cdot j}|_2.$$

Let

$$M_j = \sqrt{2\sum_{j=1}^{L}\frac{1}{l}\left(\frac{Le}{l}\right)^l e^{-C^2 nl\kappa_l^2}}.$$

One can verify that the function $f_j(\mathbf{z}) = \sqrt{\sum_{l=1}^{L}\left(|z_{j(l)}|-2C|X_{\cdot j}|_2\kappa_l\right)_+^2} : \mathbb{R}^L \to \mathbb{R}$ is a 1-Lipschitz function. Recall that $Z_{jl} \sim \mathcal{N}(0, \frac{1}{\sqrt{n}}|X_{\cdot j}|_2)$ and, therefore, by the Tsirelson-Ibragimov-Sudakov inequality (Boucheron et al., 2013, Theorem 5.6), for any $s, u \geq 1$,

$$P\left(f_j(Z) > s\frac{1}{\sqrt{n}}|X_{\cdot j}|_2\sqrt{2M_j^2+2u}\right) \leq P\left(f_j(Z) > \frac{1}{\sqrt{n}}|X_{\cdot j}|_2 M_j + \frac{1}{\sqrt{n}}|X_{\cdot j}|_2 s\sqrt{u}\right)$$
$$\leq P\left(f_j(Z) > Ef_j(Z) + \frac{1}{\sqrt{n}}|X_{\cdot j}|_2 s\sqrt{u}\right) \leq e^{-\frac{s^2}{2}u}.$$

Thus, for $s \geq 2$, we have,

$$P\left(\frac{f_{(j)}(Z)}{\lambda_j} > s\frac{1}{\sqrt{n}}|X_{\cdot j}|_2\frac{\sqrt{2M_j^2+2\log(de/j)}}{\lambda_j}\right)$$
$$\leq \binom{d}{j}P\left(\frac{f_j(Z)}{\lambda_j} > s\frac{1}{\sqrt{n}}|X_{\cdot j}|_2\frac{\sqrt{2M_j^2+2\log(de/j)}}{\lambda_j}\right)^j$$
$$\leq \binom{d}{j}e^{-j\frac{s^2}{2}\log(de/j)} \leq \left(\frac{de}{j}\right)^{-j\left(\frac{s^2}{2}-1\right)} \leq \left(\frac{de}{j}\right)^{-j\frac{s^2}{4}},$$

and applying the union bound,

$$P\left(\max_j\frac{f_{(j)}(Z)}{\lambda_j} > s\max_j\frac{1}{\sqrt{n}}|X_{\cdot j}|_2\frac{\sqrt{2M_j^2+2\log(de/j)}}{\lambda_j}\right) \leq \sum_{j=1}^{d}\left(\frac{de}{j}\right)^{-j\frac{s^2}{4}} \leq \sum_{j=1}^{d}e^{-j\frac{s^2}{4}}$$
$$\leq \frac{e^{-\frac{s^2}{4}}}{1-e^{-\frac{s^2}{4}}} \leq 2e^{-\frac{s^2}{4}}, \tag{33}$$

Finally, (33) implies

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{\max_{1\le j\le d}\frac{\delta_{(j)}}{\lambda_j}}{\max_{j=1}^{d}\frac{1}{\sqrt{n}}|X_{\cdot j}|_2\frac{\sqrt{2\sum_{j=1}^{L}\frac{1}{l}\left(\frac{Le}{l}\right)^l e^{-C^2nl\kappa_l^2}+2\log\left(\frac{de}{j}\right)}}{\lambda_j}}\right\}\\
&=\int_0^\infty P\left(\max_j\frac{f_{(j)}(Z)}{\lambda_j}>s\max_j\frac{1}{\sqrt{n}}|X_{\cdot j}|_2\frac{\sqrt{2M_j^2+2\log\left(de/j\right)}}{\lambda_j}\right)\\
&\le 2\left(1+\sqrt{\pi}\right).
\end{aligned}
$$

B.3 Proof of Lemma 11

Let $U\in\mathbb{R}^{L\times(L-1)}$ be a matrix with orthonormal columns such that $UU^T=I-\frac{1}{L}\mathbf{1}\mathbf{1}^T$. One can easily verify that $B=BUU^T$. Recall that

$$
Rad(\mathcal{B}_\lambda)=\mathbb{E}_X\mathbb{E}_\Sigma\left[\frac{1}{\sqrt{n}}\sup_{B\in\mathcal{B}_\lambda}tr(\Sigma UU^TB^TX^T)\right]=\mathbb{E}_X\mathbb{E}_\Sigma\left[\frac{1}{\sqrt{n}}\sup_{B\in\mathcal{B}_\lambda}tr(U^TB^TK)\right],
$$

where $K=X^T\Sigma U\in\mathbb{R}^{d\times(L-1)}$. By duality of Schatten norms,

$$
\frac{1}{\sqrt{n}}\sup_{B\in\mathcal{B}_\lambda}tr(U^TB^TK)=\frac{1}{\lambda}\frac{1}{\sqrt{n}}\sup_{\|B\|_*\le 1}tr(U^TB^TK)=\frac{1}{\lambda}\frac{1}{\sqrt{n}}\|X^T\Sigma U\|_2.
$$

Denote $v(X)=\|\frac{1}{\sqrt{n}}X\|_2$ and $\omega(X)=\|\frac{1}{\sqrt{n}}X\|_F\le v(X)\sqrt{d}$. By Theorem 3.2 of Rudelson and Vershynin (2013), conditionally on X, for any $s,t>1$

$$
P\left(\frac{1}{\sqrt{n}}\|X^T\Sigma U\|_2>C\left(s\omega(X)+t\sqrt{L-1}\,v(X)\right)\Big|X\right)\le 2\exp\left(-\frac{\omega^2(X)}{v^2(X)}s^2-(L-1)t^2\right),\tag{34}
$$

where $C>0$ is given in their theorem.

Assume first that $v(X)\sqrt{L-1}\ge\omega(X)$. Take $s=\frac{v(X)}{\omega(X)}\sqrt{L-1}t>1$ in (34) to get

$$
P\left(\frac{1}{\sqrt{n}}\|X^T\Sigma\|_2>2Ct\sqrt{L-1}\,v(X)\right)\le 2\exp\left(-2t^2(L-1)\right).
$$

Setting $u=2Ct\sqrt{L-1}\,v(X)$ yields

$$
P\left(\frac{1}{\sqrt{n}}\|X^T\Sigma\|_2>u\;\Big|X\right)\le 2\exp\left(-\frac{u^2}{2C^2v(X)^2}\right),
$$

for any $u>2C\sqrt{L-1}v(X)$ and, therefore, the empirical Rademacher complexity

$$
\widehat{Rad}(\mathcal{B}_\lambda)\le\frac{1}{\lambda}\left(2C\sqrt{L-1}v(X)+2\int_{2C\sqrt{L-1}v(X)}^\infty e^{-\frac{u^2}{2C^2v(X)^2}}du\right)\le C\frac{1}{\lambda}\sqrt{L-1}v(X).
$$

Similarly, for $v(X)\sqrt{L-1} < \omega(X)$, take $t = \frac{\omega(X)}{v(X)\sqrt{L-1}} s > 1$ in (34) and $u = 2Cs\omega(X)$ to get

$$P\left(\frac{1}{\sqrt{n}}\|X^T\Sigma\|_2 > u \,\Big|X\right) \leq 2\exp\left(-\frac{u^2}{2C^2 v(X)^2}\right),$$

for any $u > 2C\omega(X)$ and, therefore,

$$\widehat{Rad}(\mathcal{B}_\lambda) \leq C\frac{1}{\lambda}\omega(X) \leq C\frac{1}{\lambda}v(X)\sqrt{d}.$$

Combining both cases we have

$$\widehat{Rad}(\mathcal{B}_\lambda) \leq C\frac{1}{\lambda}v(X)(\sqrt{L-1}+\sqrt{d}). \tag{35}$$

To complete the proof of the lemma apply the results of Vershynin (2012, Section 5.4.1) for sub-Gaussian matrices with independent rows to get

$$E_X v(X) \leq \sqrt{E_X v^2(X)} \leq C\sqrt{\tau_1(V)}. \tag{36}$$

Appendix C. Proof of Theorem 5

Consider the class $\tilde{\mathcal{C}}_L(r_0)$ of r_0-globally sparse linear L-class classifiers from Section 3.1 but with the *known* subset of r_0 significant features. Evidently, $\tilde{\mathcal{C}}_L(r_0) \subset \mathcal{C}^*_L(r_0)$. Apply now the results of Abramovich et al. (2021, Theorem 2) on the lower bounds for global row-wise sparse classification to get

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}^*_L(r_0),\ \mathbb{P}_X} \mathcal{E}(\tilde{\eta}, \eta^*) \geq \inf_{\tilde{\eta}} \sup_{\eta^* \in \tilde{\mathcal{C}}_L(r_0),\ \mathbb{P}_X} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{r_0(L-1)}{n}}. \tag{37}$$

On the other hand, consider r_0-class classification, where all d features are significant ($d_0 = d$). It is obvious that $\mathcal{C}_{r_0}(d) \subset \mathcal{C}^*_{r_0}(r_0)$ and that r_0-class classification cannot be harder than the L-class one. Thus, exploiting again Theorem 2 of Abramovich et al. (2021) we have

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}^*_L(r_0),\ \mathbb{P}_X} \mathcal{E}(\tilde{\eta}, \eta^*) \geq \inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}^*_{r_0}(r_0),\ \mathbb{P}_X} \mathcal{E}(\tilde{\eta}, \eta^*) \geq \inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{r_0}(d),\ \mathbb{P}_X} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{r_0 d}{n}}. \tag{38}$$

Combining (37) and (38) completes the proof of the theorem.

Appendix D. Sparse group Slope algorithm

The penalized MLE minimization problem in (12) involves a sum of a convex smooth log-likelihood and a convex but non-smooth penalty consisting of two terms. A common approach to solve such optimization problems is by the proximal gradient method (e.g., Beck, 2017). A general proximal operator of a given convex function f is defined as

$$\operatorname{prox}_f(a) = \arg\min_b \left\{ \frac{1}{2}\|a-b\|^2 + f(b) \right\}.$$

For the setup at hand consider the proximal operator

$$\operatorname{prox}_{\|\cdot\|_{\kappa,\lambda}}(A) = \arg\min_B \left\{ \frac{1}{2}\|A-B\|_F^2 + \|B\|_{\kappa,\lambda} \right\}, \tag{39}$$

where recall that $\|B\|_{\kappa,\lambda} = \sum_{j=1}^d \lambda_j |B|_{(j)} + \sum_{j=1}^d \sum_{l=1}^L \kappa_l |B|_{j(l)} = \|B\|_\lambda + \sum_{j=1}^d \|B_{j\cdot}\|_\kappa$.

There exist the efficient proximal gradient descent algorithms for computing proximal operators $\operatorname{prox}_{\|\cdot\|_\kappa}$ and $\operatorname{prox}_{\|\cdot\|_\lambda}$ for $\|\cdot\|_\kappa$ and $\|\cdot\|_\lambda$ separately (see respectively Bogdan et al., 2015; Brzyski et al., 2019). We now show that applying $\operatorname{prox}_{\|\cdot\|_\kappa}$ and $\operatorname{prox}_{\|\cdot\|_\lambda}$ consecutively results in $\operatorname{prox}_{\|\cdot\|_{\kappa,\lambda}}$ as depicted by Algorithm 1:

Algorithm 1: $\operatorname{prox}_{\|\cdot\|_{\kappa,\lambda}}(A)$

```
for j → 1 ... d do
    U_j· = prox_{||·||_κ}(A_j·)
end
B ← prox_{||·||_λ}(U)
```

The proof relies on the second prox theorem (Beck, 2017, Theorem 6.39) and the following general lemma:

Lemma 14 *Assume that for all a, $\partial g(\operatorname{prox}_f(a)) \supseteq \partial g(a)$, then for all b, $\operatorname{prox}_{f+g}(b) = \operatorname{prox}_f(\operatorname{prox}_g(b))$.*

Proof For a given b, let $a = \operatorname{prox}_g(b)$ and $z = \operatorname{prox}_f(a)$. By the second prox theorem, $b - a \in \partial g(a)$ and $a - z \in \partial f(z)$. By the condition, $\partial g(z) \supseteq \partial g(a)$, and therefore,

$$b - z = b - a + a - z \in \partial f(z) + \partial g(z) = \partial(f+g)(z)$$

which implies by the second prox theorem that $z = prox_{f+g}(b)$. ■

Applying Lemma 14 for $g(A) = \sum_{j=1}^d \|A_{j,\cdot}\|_\kappa$ and $f(A) = \|A\|_\lambda$ relies on the following lemma:

Lemma 15 *For $Z, A \in \mathbb{R}^{d\times L}$ such that $Z \in \partial\|\cdot\|_\lambda(A)$ and for any $j \in \{1,\ldots,d\}$, there exists $c_j \geq 0$ such that $Z_{j\cdot} = c_j A_{j\cdot}$.*

Proof Let $Z \in \partial\|\cdot\|_\lambda(A)$. Thus,

$$Z \in \underset{\|Z\|_\lambda^* \le 1}{\operatorname{argmax}}\, tr(Z^T A) = \underset{\|Z\|_\lambda^* \le 1}{\operatorname{argmax}} \sum_{j=1}^{d} Z_{j\cdot}^T A_{j\cdot} \ ,$$

where $\|\cdot\|_\lambda^*$ is the dual norm. Since the norm $\|\cdot\|_\lambda$ is invariant to rotation of the rows, so does its dual norm $\|\cdot\|_\lambda^*$ because we can always rotate the rows of the norming matrix. Thus, the maximum above is when $Z_{j\cdot} = c_j A_{j\cdot}$ for some $c_j \ge 0$. ■

Let $Z = \text{prox}_{\|\cdot\|_\lambda}(A)$. By the second prox theorem we have $A - Z \in \partial\|\cdot\|_\lambda(Z)$, and by Lemma 15, $A_j - Z_j = c_j Z_j$ for some $c_j > 0$. Thus, $Z_j = \frac{1}{1+c_j} A_j$.

Let $V \in \partial(\sum_{j=1}^d \|e_j^T \cdot\|_\kappa)(A)$, that is, $V_{j\cdot} \in \partial\|\cdot\|_\kappa(A_{j\cdot})$. By the definition of the subgradient, for any $\mathbf{u} \in \mathbb{R}^L$,

$$\|A_{j\cdot}\|_\kappa + V_{j\cdot}^T(\mathbf{u} - A_{j\cdot}) \le \|\mathbf{u}\|_\kappa$$

Let $\mathbf{u'} \in \mathbb{R}^L$. Then,

$$\begin{aligned}\|Z_{j\cdot}\|_\kappa + V_{j\cdot}^T(\mathbf{u'} - Z_{j\cdot}) &= \frac{1}{1+c_j}\|A_{j\cdot}\|_\kappa + \frac{1}{1+c_j} V_{j\cdot}^T((1+c_j)\mathbf{u'} - A_{j\cdot}) \\ &\le \frac{1}{1+c_j}\|(1+c_j)\mathbf{u'}\|_\kappa = \|\mathbf{u'}\|_\kappa\end{aligned}$$

and, therefore, $V_{j\cdot}^T(\mathbf{u'} - Z_{j\cdot}) \le \|\mathbf{u'}\|_\kappa - \|Z_{j\cdot}\|_\kappa$ implying $V_{j\cdot} \in \partial\|\cdot\|_\kappa(Z_{j\cdot})$. Hence, $V \in \partial(\sum_{j=1}^d \|e_j^T \cdot\|_\kappa)(Z)$ and the condition for Lemma 14 holds, i.e.

$$\partial\|\cdot\|_\kappa(\text{prox}_{\|\cdot\|_\lambda}(A)) \supseteq \partial\|\cdot\|_\kappa(A).$$

References

Felix Abramovich and Vadim Grinshtein. High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079, 2019.

Felix Abramovich, Vadim Grinshtein, and Tomer Levy. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 67(7):4637–4646, 2021.

Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *The Annals of Statistics*, 47:2117–2144, 2019.

Francis R. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

Amir Beck. *First-order Methods in Optimization*. SIAM, 2017.

Pierre C. Bellec, Guillaume Lecué, and Alexander B. Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989 – 1010, 2004.

Peter J. Bickel, Ya'acov Ritov, and Alexander B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

Damian Brzyski, Alexej Gossmann, Weijie Su, and Małgorzata Bogdan. Group slope–adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525):419–433, 2019.

Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Di-Rong Chen and Tao Sun. Consistency of multiclass empirical risk minimization methods based on convex loss. *Journal of Machine Learning Research*, 7(86):2435–2447, 2006.

Le-Yu Chen and Sokbae Lee. Binary classification with covariate selection through ℓ_0-penalised empirical risk minimisation. *The Econometrics Journal*, 24(1):103–120, 2021.

Amit Daniely, Sivan Sabato, and Shai Shwartz. Multiclass learning approaches: a theoretical comparison with implications. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16(72):2377–2404, 2015.

Jianqing Fan and Yingying Fan. High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605 – 2637, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Vladimir Koltchinskii, Karim Lounici, and Alexander B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. *The Annals of Statistics*, 46:611–641, 2018.

Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65 (5):2995–3021, 2019.

Yurii Maximov and Dar'ya Reshetova. Tight risk bounds for multi-class margin classifiers. *Pattern Recognition and Image Analysis*, 26(4):673–680, 2016.

Peter McCullagh and John A Nelder. *Generalized Linear Models*. Routledge, 1989.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, USA, second edition, 2018.

Bernardo Ávila Pires and Csaba Szepesvári. Multiclass classification calibration functions. *arXiv:1609.06385*, 2016.

Scott Powers, Trevor Hastie, and Robert Tibshirani. Nuclear penalized multinomial regression with an application to predicting at bat outcomes in baseball. *Statistical Modelling*, 18(5-6):388–410, 2018.

Henry Reeve and Ata Kaban. Optimistic bounds for multi-output learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, 2014.

Yiyuan She. Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, 6(2):197–209, 2013.

Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2000.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Gitta Kutyniok and Yonina C. Eldar, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, Cambridge, 2012.

Martin Vincent and Niels Richard Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics* & *Data Analysis*, 71(C):771–786, 2014.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.