Multiclass Classification by Sparse Multinomial Logistic Regression

Felix Abramovich, Vadim Grinshtein, and Tomer Levy

Abstract—In this paper we consider high-dimensional multiclass classification by sparse multinomial logistic regression. We propose first a feature selection procedure based on penalized maximum likelihood with a complexity penalty on the model size and derive the nonasymptotic bounds for misclassification excess risk of the resulting classifier. We establish also their tightness by deriving the corresponding minimax lower bounds. In particular, we show that there is a phase transition between small and large number of classes. The bounds can be reduced under the additional low noise condition. To find a penalized maximum likelihood solution with a complexity penalty requires, however, a combinatorial search over all possible models. To design a feature selection procedure computationally feasible for high-dimensional data, we propose multinomial logistic group Lasso and Slope classifiers and show that they also achieve the minimax order.

Index Terms—Complexity penalty, convex relaxation, feature selection, high-dimensionality, minimaxity, misclassification excess risk, sparsity.

I. INTRODUCTION

CLASSIFICATION is one of the core problems in statistical learning and has been intensively studied in statistical and machine learning literature. Nevertheless, while the theory for binary classification is well developed (see [15], [20], [37], and references therein for a comprehensive review), its multiclass extensions are much less complete.

Consider a general $L$-class classification with a (high-dimensional) vector of features $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ and the outcome class label $Y \in \{1, \ldots, L\}$. We can model it as

$$Y|\mathbf{X} = \mathbf{x}) \sim \text{Mult}(p_1(\mathbf{x}), \ldots, p_L(\mathbf{x})),$$

where $p_l(\mathbf{x}) = P(Y = l|\mathbf{X} = \mathbf{x})$, $l = 1, \ldots, L$.

A classifier is a measurable function $\eta : \mathcal{X} \rightarrow \{1, \ldots, L\}$ and its accuracy is defined by a misclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$. The optimal classifier that minimizes this error is the Bayes classifier $\eta^*(\mathbf{x}) = \arg \max_{1 \leq l \leq L} p_l(\mathbf{x})$ with

$$R(\eta^*) = 1 - E_X \max_{1 \leq l \leq L} p_l(\mathbf{x}).$$

The probabilities $p_l(\mathbf{x})$’s are, however, unknown and one should derive a classifier $\hat{\eta}(\mathbf{x})$ from the available data $D$: a random sample of $n$ independent observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from the joint distribution of $(\mathbf{X}, Y)$. The corresponding (conditional) misclassification error of $\hat{\eta}$ is $R(\hat{\eta}) = P(Y \neq \hat{\eta}(\mathbf{x})|D)$ and the goodness of $\hat{\eta}$ w.r.t. $\eta^*$ is measured by the misclassification excess risk $\mathcal{E}(\hat{\eta}, \eta^*) = ER(\hat{\eta}) - R(\eta^*)$. The goal is then to find a classifier $\hat{\eta}$ within given family with minimal $\mathcal{E}(\hat{\eta}, \eta^*)$.

A first strategy in multiclass classification is to reduce it to a series of binary classifications. The probably 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 to extend binary classification approaches for multiclass case. Thus, a common approach to design a multiclass classifier $\hat{\eta}$ is based on empirical risk minimization (ERM), where minimization of a true misclassification error $R(\eta)$ is replaced by minimization of the corresponding empirical risk $\hat{R}_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} I\{Y_i \neq \eta(\mathbf{x}_i)\}$ over a given class of classifiers. For binary classification, tight risk bounds for ERM classifiers have been established in terms of VC-dimension, Rademacher complexity or covering numbers (see [15], [20], [37], and references therein). Their extensions to multiclass case, however, are not straightforward. See [27] for a comprehensive survey of the state-of-the-art results on the upper bounds for misclassification excess risk of multiclass ERM classifiers. A comparison of error bounds

0000-0000/00/$00.00 © 2015 IEEE
for ERM classifiers with those for OvA and OvO is given in [18].

A crucial drawback of ERM is in minimization of 0-1 loss that makes it computationally infeasible. A typical remedy is to replace 0-1 loss by a related convex surrogate. The resulting solution approximates then the minimizer of the corresponding surrogate risk. The goal is to find a surrogate loss such that minimization of its risk leads to a Bayes classifier \( \eta^* \) (aka Fisher consistent or calibrated loss). Various calibrated surrogate risk. The goal is to find a surrogate loss such that minimization of its risk leads to a Bayes classifier \( \eta^* \) (aka Fisher consistent or calibrated loss). Various calibrated losses for multiclass classification have been considered in the literature (e.g., [7], [8], [17], [34], [42]).

An alternative approach to ERM is to estimate \( p_l(x) \)'s from the data by some \( \widehat{p}_l(x) \)'s and to use a plug-in classifier of the form \( \widehat{\eta}(x) = \arg \max_{1 \leq l \leq L} \widehat{p}_l(x) \). A standard approach is to assume some (parametric or nonparametric) model for \( p_l(x) \). The most commonly used model is multinomial logistic regression, where it is assumed that \( p_l(x) = \frac{\exp(\beta_l^T x)}{\sum_{l'=1}^L \exp(\beta_{l'}^T x)} \) and \( \beta_l \in \mathbb{R}^d \), \( l = 1, \ldots, L \) are unknown vectors of regression coefficients. The corresponding Bayes classifier is, therefore, a linear classifier \( \eta^*_l(x) = \arg \max_{1 \leq l \leq L} p_l(x) = \arg \max_{1 \leq l \leq L} \beta_l^T x \). One then estimates \( \beta_l \)'s from the data by the maximum likelihood estimators (MLE) \( \hat{\beta}_l \)'s and derives the plug-in (linear) classifier \( \widehat{\eta}(x) = \arg \max_{1 \leq l \leq L} \beta_l^T x \). Unlike ERM, the MLE \( \hat{\beta}_l \)'s though not available in the closed form, can be nevertheless obtained numerically by the fast iteratively reweighted least squares algorithm [29, Section 2.5].

The general challenge modern statistics faces with is high-dimensionality of the data, where the number of features \( d \) is large and might be even larger than the sample size \( n \) (“large \( d \) small \( n \)” setups) that raises a severe “curse of dimensionality” problem. Reducing the dimensionality of a feature space by selecting a sparse subset of “significant” features becomes crucial.

For binary classification [20, Chapter 18] and [37, Chapter 4] considered model selection from classifiers within a sequence of classes by penalized ERM with the structural penalty depending on the VC-dimension of a class. See also [15, Section 8] for related penalized ERM approaches and references therein. [3] explored feature selection in high-dimensional logistic regression classification.

To the best of our knowledge, feature selection for multiclass classification has not yet been rigorously well-studied and the goal of this paper is to fill the gap. Thus, we propose a model/feature selection procedure based on penalized maximum likelihood with a certain complexity penalty on the model size. We establish the non-asymptotic upper bounds for misclassification excess risk of the resulting plug-in classifier which is also adaptive to the unknown sparsity and show their tightness by deriving the corresponding minimax lower bound over a set of sparse linear classifiers. It turns out that there appear interesting phenomena in the multiclass setup. In particular, we find that there is a phase transition. For \( L \leq 2 + \ln(d/d_0) \), where \( d_0 \) is the size of the true (unknown) model, the multiclass effect is not manifested and the minimax misclassification excess risk over the set of \( d_0 \)-sparse linear classifiers is of the order \( \sqrt{\frac{d_0 \ln \left( \frac{d_0}{n} \right)}{n}} \) regardless of \( L \). For larger \( L \), it increases as \( \sqrt{\frac{d_0 (L-1)}{n}} \) and does not depend on \( d \). We also show that these bounds can be improved under the additional low-noise assumption.

Any penalized maximum likelihood procedure that involves a complexity penalty requires, however, a combinatorial search over all possible models that makes its use computationally infeasible for large \( d \). A common remedy is then to use a convex surrogate, where the original combinatorial minimization is replaced by a related convex program. In this paper we consider Slope convex relaxation which can be viewed as generalization of the celebrated Lasso and show that for the properly chosen tuning parameters, the resulting multinomial logistic group Slope multiclass classifier is also minimax rate-optimal.

The rest of the paper is organized as follows. In Section II we present sparse multinomial logistic regression model and propose a feature selection procedure. The bounds for misclassification excess risk of the resulting plug-in classifier are derived in Section III. In Section IV we introduce the additional low-noise assumption that allows one to improve the bounds. In Section V we develop group Slope convex relaxation techniques for multiclass classification with Lasso as its particular case, and establish the misclassification excess risk bounds for the resulting classifier. All the proofs are given in the Appendix.

II. CONSTRUCTION OF A CLASSIFIER

A. Multinomial logistic regression model

Consider \( d \)-dimensional \( L \)-class classification setup that can be written in the following form:

\[
Y | (X = x) \sim \text{Mult}(p_1(x), \ldots, p_L(x)),
\] (1)
where \( \mathbf{X} \in \mathbb{R}^d \) is a vector of linearly independent features with a marginal probability distribution \( P_X \) with a support \( \mathcal{X} \subseteq \mathbb{R}^d \) and \( \sum_{j=1}^{L} p_j(x) = 1 \) for any \( x \in \mathcal{X} \).

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

\[
\ln \frac{p_l(x)}{p_{l'}(x)} = \beta_l^T \mathbf{x}, \quad l = 1, \ldots, L - 1, \tag{2}
\]

and \( \beta_l \in \mathbb{R}^d \) are the vectors of the (unknown) regression coefficients. Hence,

\[
p_l(x) = \frac{\exp(\beta_l^T \mathbf{x})}{1 + \sum_{k=1}^{L-1} \exp(\beta_k^T \mathbf{x})}, \quad l = 1, \ldots, L - 1
\]

and

\[
p_L(x) = \frac{1}{1 + \sum_{k=1}^{L-1} \exp(\beta_k^T \mathbf{x})}
\]

or, in a somewhat more compact form,

\[
p_l(x) = \frac{\exp(\beta_l^T \mathbf{x})}{\sum_{k=1}^{L} \exp(\beta_k^T \mathbf{x})}, \quad l = 1, \ldots, L
\]

with \( \beta_L = 0 \). The Bayes classifier is then a linear classifier

\[
\eta^*(\mathbf{x}) = \arg \max_{1 \leq l \leq L} p_l(\mathbf{x}) = \arg \max_{1 \leq l \leq L} \beta_l^T \mathbf{x}.
\]

The choice of the last class as a reference class is, in fact, quite arbitrary. One can consider an equivalent model with any other reference class \( h \) instead: \( \ln \frac{p_l(x)}{p_h(x)} = \gamma_l \mathbf{x}, \quad l \neq h \). Evidently, there is one-to-one transformation: \( \gamma_l = \beta_l - \beta_h \) and \( \beta_l = \gamma_l - \gamma_l \). Change of a reference class is, therefore, just a matter of reparametrization of the same model.

**B. Penalized maximum likelihood estimation**

To each possible value \( y \in \{1, \ldots, L\} \) of \( Y \) assign the indicator vector \( \mathbf{1}_y \in \{0,1\}^L \) with \( \mathbf{1}_y = I\{y = l\}, \quad l = 1, \ldots, L \). Let \( B \in \mathbb{R}^{d \times L} \) be the matrix of the regression coefficients in (2) with the columns \( \beta_1, \ldots, \beta_L \) (recall that \( \beta_L = 0 \)) and let \( f_B(\mathbf{x}, y) \) be the corresponding joint distribution of \((\mathbf{X}, Y)\), i.e. \( df_B(\mathbf{x}, y) = \prod_{l=1}^{L} p_l(\mathbf{x})^{\mathbf{1}_y} \, dP_X(\mathbf{x}) \), where

\[
p_l(x) = \frac{\exp(\beta_l^T \mathbf{x})}{\sum_{k=1}^{L} \exp(\beta_k^T \mathbf{x})}. 
\]

Given a random sample \((\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n) \sim f_B(\mathbf{X}, \mathbf{Y})\), the conditional log-likelihood function is

\[
\ell(B) = \sum_{i=1}^{n} \left\{ \mathbf{X}_i^T B \mathbf{1}_{\mathbf{Y}_i} - \ln \sum_{l=1}^{L} \exp(\beta_l^T \mathbf{X}_i) \right\}, \tag{3}
\]

and one can find the maximum likelihood estimator (MLE) for \( B \) by maximizing \( \ell(B) \).

The era of “Big Data” brought the challenge of dealing with problems, where the number of features \( d \) is very large and may be even larger than the sample size \( n \) (“large \( d \) small \( n \)” setups). Nevertheless, it is commonly assumed that the true underlying model is sparse and most of the features do not have a significant impact on classification. Reducing the dimensionality of a feature space by selecting a sparse subset of “significant” features is then crucial. Thus, \([11]\), \([21]\) showed that even binary high-dimensional classification without a proper feature selection might be as bad as just pure guessing.

For binary classification, where the regression matrix \( B \) reduces to a single vector \( \beta \in \mathbb{R}^d \), the sparsity is naturally measured by the \( l_0 \) (quasi)-norm \( |\beta|_0 \) – the number of non-zero entries of \( \beta \) (see, e.g., \([3]\)). For multiclass case one can think of several possible ways to extend the notion of sparsity. The most evident measure of sparsity is the number of non-zero rows of \( B \) that corresponds to the assumption that part of the features do not have any impact on classification at all and, therefore, have zero coefficients in (2) for all \( l \). It can be viewed as global row-wise sparsity. One can easily verify that such a measure is invariant under the choice of the reference class in (2).

In what follows we assume the following assumption:

**Assumption (A). Assume that there exists \( 0 < \delta < 1/2 \) such that \( \delta < p_l(\mathbf{x}) < 1 - \delta \) or, equivalently, \( |\beta_l^T \mathbf{x}| < C_0 \) with \( C_0 = \ln \frac{1-\delta}{\delta} \) for all \( x \in \mathcal{X} \) and all \( l = 1, \ldots, L \).**

Assumption (A) prevents the conditional variances \( \text{Var}(\xi_l | \mathbf{X} = \mathbf{x}) = p_l(\mathbf{x})/(1 - p_l(\mathbf{x})) \) to be arbitrarily close to zero, where any MLE-based procedure may fail.

Let \( \mathcal{M} \) be the set of all \( 2^d \) possible models \( M \subseteq \{1, \ldots, d\} \). For a given model \( M \) define a set of matrices \( \mathcal{B}_M = \{ B \in \mathbb{R}^{d \times L} : B_L = 0 \text{ and } B_j = 0 \text{ iff } j \notin M \} \). Obviously, all matrices in \( \mathcal{B}_M \) have the same number of non-zero rows which can be naturally defined as a model size \( |M| \).

Under the model \( M \), the MLE \( \hat{B}_M \) of \( B \) is then

\[
\hat{B}_M = \arg \max_{B \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \mathbf{X}_i^T B \xi_i - \ln \sum_{l=1}^{L} \exp(\beta_l^T \mathbf{X}_i) \right\}, \tag{4}
\]

where \( \beta_l = \tilde{B}_l, \quad l = 1, \ldots, L \) are the columns of \( \tilde{B} \).

Select the model \( \hat{M} \) by the penalized maximum likelihood model selection criterion of the form

\[
\hat{M} = \arg \min_{M \in \mathcal{M}} \left\{ \sum_{i=1}^{n} \left( \ln \left( \sum_{l=1}^{L} \exp(\beta_{Ml}^T \mathbf{X}_i) \right) - \mathbf{X}_i^T \hat{B}_M \xi_i \right) + \text{Pen}(|M|) \right\}, \tag{5}
\]

with the complexity penalty \( \text{Pen}(\cdot) \) on the model size \( |M| \).
Finally, for the selected model \( \hat{M} \) the resulting plug-in classifier
\[
\hat{\eta}_{\hat{M}}(x) = \mathop{\arg\max}_{1 \leq i \leq L} \hat{\beta}_{\hat{M}}^T x \tag{6}
\]

The proper choice of the complexity penalty \( \text{Pen}(\cdot) \) in (5) is obviously the core of the proposed approach.

III. MISCLASSIFICATION EXCESS RISK BOUNDS

We now derive the (non-asymptotic) upper bound for misclassification excess risk of the penalized maximum likelihood classifier (6) derived in Section II for a particular type of the classification excess risk of the penalized maximum likelihood classifier of Theorem 3.

Consider a \( \ell \)-class classifier \( \hat{c} \) derived in Section II for a particular type of the classification excess risk of the penalized maximum likelihood classifier of Theorem 3.

The proper choice of the complexity penalty \( \text{Pen}(\cdot) \) in (5) is obviously the core of the proposed approach.

Theorem 1. Consider a \( d_0 \)-sparse multinomial logistic regression model (1)-(2).

Let \( \hat{M} \) be a model selected in (4)-(5) with the complexity penalty
\[
\text{Pen}(|\hat{M}|) = c_1 |\hat{M}| (L - 1) + c_2 |\hat{M}| \ln \left( \frac{d \varepsilon}{|\hat{M}|} \right), \tag{7}
\]
where the absolute constants \( c_1, c_2 > 0 \) are given in the proof of Theorem 3.

Then, under Assumption (A),
\[
\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{\hat{M}}; \eta^*)
\leq C_1(\delta) \left( \frac{d_0(L - 1) + d_0 \ln \left( \frac{d \varepsilon}{d_0} \right)}{n} \right) \tag{8}
\]
for some \( C_1(\delta) \) depending on \( \delta \), simultaneously for all \( 1 \leq d_0 \leq \min(d, n) \).

Theorem 1 is a particular case of a more general Theorem 3 given in the next Section IV.

The complexity penalty \( \text{Pen}(|\hat{M}|) \) in (7) contains two terms. The first one is proportional to \( |\hat{M}| (L - 1) \) – the overall number of estimated parameters in the model \( M \) and is an AIC-type penalty. The second one is proportional to \( |\hat{M}| \ln \left( \frac{d}{|M|} \right) \sim \ln \left( \frac{d}{|M|} \right) \) – the log of the number of all possible models of size \( |\hat{M}| \) and typically appears in model selection in various regression and classification setups (see, e.g. [1]-[3], [13], [16]).

Theorem 2 below shows that for an agnostic model, where the Bayes risk \( R(\eta^*) > 0 \), the upper bound (8) for the misclassification excess risk established in Theorem 1 is essentially tight and up to a possibly different constant coincides with the corresponding minimax lower bound over \( C_L(d_0) \):

**Theorem 2.** Consider a \( d_0 \)-sparse agnostic multinomial logistic regression model (1)-(2), where \( 2 \leq d_0 \ln \left( \frac{d \varepsilon}{d_0} \right) \leq n \) and \( d_0(L - 1) \leq n \). Then,
\[
\inf_{\hat{\eta}} \sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}, \eta^*) \geq C_2 \sqrt{\frac{d_0(L - 1) + d_0 \ln \left( \frac{d \varepsilon}{d_0} \right)}{n}}
\]
for some \( C_2 > 0 \), where the infimum is taken over all classifiers \( \hat{\eta} \) based on the data \( (X_i, Y_i), i = 1, \ldots, n \).

The above bounds imply, in particular, that as \( d \) and \( L \) grow with \( n \) and assuming that \( \delta \) is bounded away from zero, there are two different regimes:

1. **Small number of classes:** \( L \leq 2 + \ln \left( \frac{d}{d_0} \right) \).

In this case, the complexity penalty \( \text{Pen}(|\hat{M}|) \sim c|M| \ln \left( \frac{d}{|M|} \right) \) does not depend on \( L \). The resulting (rate-optimal) misclassification excess risk is of the order \( \sqrt{\frac{d_0(L - 1)}{n}} \) regardless of \( L \) and the error in feature selection dominates in the overall excess risk. Multiclass classification for such a small number of classes is essentially not harder than binary (see the results of [3] for \( L = 2 \)).

2. **Large number of classes:** \( 2 + \ln \left( \frac{d}{d_0} \right) < L \leq \frac{n}{d_0} \).

In this regime, \( \text{Pen}(|\hat{M}|) \sim c|M| \ln \left( \frac{d}{|M|} \right) \) is an AIC type penalty (see above), the misclassification excess risk increases with \( L \) as \( \sqrt{\frac{d_0(L - 1)}{n}} \) regardless of \( d \) and the main contribution to the overall error comes from estimating the large number \( (d_0(L - 1)) \) of parameters in the model.

For \( L > \frac{n}{d_0} \) the number of parameters in the model becomes larger than the sample size and consistent classification is impossible.

In particular, without sparsity assumption, i.e. in the case \( d_0 = \frac{d}{n} \), the misclassification excess risk is of the order \( \sqrt{\frac{d(L - 1)}{n}} \) for all \( 1 \leq L \leq \frac{n}{d} \).

Note that even if the considered multinomial logistic regression model is misspecified and the Bayes classifier \( \eta^* \) is not linear, we still have the following excess risk decomposition
\[
R(\hat{\eta}_{\hat{M}}) - R(\eta^*) = (R(\hat{\eta}_{\hat{M}}) - R(\eta^*_L)) + (R(\eta^*_L) - R(\eta^*)). \tag{9}
\]
where $\eta_L^* = \arg\min_{\eta \in C_L(d)} R(\eta)$ is the oracle (ideal) linear classifier. The above results can then be still applied to the first term in the RHS of (9) representing the estimation error, whereas the second term is an approximation error that measures the ability of linear classifiers to perform as good as $\eta^*$. Enriching the class of classifiers may improve the approximation error but will necessarily increase the estimation error in (9). In a way, it is similar to the variance/bias tradeoff in regression.

IV. IMPROVED BOUNDS UNDER LOW-NOISE CONDITION

Intuitively, it is clear that misclassification error is particularly large where it is difficult to separate the class with the highest probability from others, i.e. at those $x \in X$, where the largest probability $p_1(x)$ is close to the second largest $p_2(x)$ (see also [23]).

Define the following multiclass extension of the low-noise (aka Tsybakov) condition as in [25], [35]:

**Assumption (B).** Consider the multinomial logistic regression model (7)-(2) and assume that there exist $C > 0$, $\alpha \geq 0$ and $h^* > 0$ such that for all $0 < h \leq h^*$,

$$P \left( p_1(X) - p_2(X) \leq h \right) \leq C h^\alpha$$

(10)

(see also [4], [17]). Assumption (B) implies that with high probability (depending on the parameter $\alpha$) the most likely class is sufficiently distinguished from others. The two extreme cases $\alpha = 0$ and $\alpha = \infty$ correspond respectively to no assumption on the noise considered in the previous sections and to existence of a hard margin of size $h^*$ separating $p_1(x)$ and $p_2(x)$. A straightforward multiclass extension of [9] Lemma 5] implies that (10) is equivalent to the condition that there exists $C_1(\alpha)$ such that for any classifier $\eta$,

$$P \left( \eta(X) \neq \eta^*(X) \right) \leq C_1(\alpha) \mathcal{E}(\eta, \eta^*)^{\frac{\alpha + 1}{\alpha + 2}}$$

(11)

We now show that under the additional low-noise condition (10), the bounds for the misclassification excess risks established in the previous Section III can be improved:

**Theorem 3.** Consider a $d_0$-sparse multinomial logistic regression model (1)-(2) and let $\tilde{M}$ be a model selected in (5) with the complexity penalty (7).

Then, under Assumptions (A) and (B), there exists $C(\delta)$ such that

$$\sup_{\eta^* \in C_L(d_0)} \mathcal{E}(\tilde{\eta}_{\tilde{M}}, \eta^*) \leq C(\delta) \left( \frac{d_0(L - 1) + d_0 \ln \left( \frac{d}{d_0} \right)}{n} \right)^{\frac{\alpha + 1}{\alpha + 2}}$$

for all $1 \leq d_0 \leq \min(d, n)$ and all $\alpha \geq 0$.

Note that $\tilde{\eta}_{\tilde{M}}$ is inherently adaptive to both $d_0$ and $\alpha$. As we have mentioned, Theorem 1 is a particular case of Theorem 3 with $\alpha = 0$.

To conclude this section we note that the error bounds can be also improved under other types of additional constraints on the marginal distribution $P_X$, e.g., a so-called strong density assumption ([6] for binary classification) or a cluster assumption ([28], [31]).

V. MULTINOMIAL LOGISTIC GROUP LASSO AND SLOPE

Despite strong theoretical results on penalized maximum likelihood classifiers with complexity penalties established in the previous sections, selecting the model $\tilde{M}$ in (5) requires a combinatorial search over all possible models in $\mathcal{M}$ that makes it computationally infeasible when the number of features is large. A common approach to handle this problem is convex relaxation, where the original combinatorial minimization is replaced by a related convex surrogate. The most well-known examples include the celebrated Lasso, where the $l_0$-norm in the penalty is replaced by $l_1$-norm and, its recently developed more general variation Slope that uses a sorted $l_1$-type norm ([14]). Lasso and Slope estimators have been intensively studied in Gaussian regression (see, e.g., [10], [12], [33] among others), and their logistic modifications in logistic regression ([3], [5], [36]). [3] investigated logistic Lasso and Slope classifiers for the binary case. In this section we consider multinomial logistic group Lasso and Slope classifiers and extend the corresponding results of [3] for multiclass classification.

Recall that we consider a global row-wise sparsity, where the coefficient regression matrix $B$ has a subset of zero rows. To capture such type of sparsity we consider a multinomial logistic group Lasso classifier defined as follows. For a given tuning parameter $\lambda > 0$, find

$$\tilde{B}_{gL} = \arg\min_B \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \ln \left( \sum_{l=1}^{L} \exp(\tilde{\beta}_{i,l}^T x_i) \right) - X_i^T \tilde{B}_l \xi_i \right) + \lambda \sum_{j=1}^{d} |B_l|_2 \right\},$$

(12)

where $|B_l|_2 = |\tilde{B}_j|_2$ is the $l_2$-norm of the $j$-th row of $\tilde{B}$ and define the corresponding classifier $\tilde{\eta}_{gL}(x) = \arg\max_{1 \leq l \leq L} \tilde{\beta}_{gL,l}^T x$. An efficient algorithm for computing multinomial logistic group Lasso is given in [38].
Multinomial logistic group Slope is a more general variation of [12]. Namely,
\[
\hat{B}_{gS} = \arg \min_{\hat{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \ln \left( \sum_{l=1}^{L} \exp(\beta^T_l \mathbf{x}_i) \right) - \mathbf{x}_i^T \hat{B} \xi_i \right) \right\} + \sum_{j=1}^{d} \lambda_j |\hat{B}|(j),
\]
where the rows’ l2-norms $|\hat{B}|(j) \geq \ldots \geq |\hat{B}|(d)$ are the descendingly ordered and $\lambda_1 \geq \ldots \geq \lambda_d > 0$ are the tuning parameters, and set $\hat{g}_{gS}(\mathbf{x}) = \arg \max_{1 \leq l \leq L} \hat{B}_{gS,l}^T \mathbf{x}$. Multinomial logistic group Lasso [12] can be evidently viewed as a particular case of [13] with equal $\lambda_j$’s.

Note that unlike complexity penalties, the solution of [13] is identifiable without the additional constraint $\hat{\beta}_L = 0$. Moreover, since the unconstrained log-likelihood [3] satisfies $\ell(\hat{\beta}_1, \ldots, \hat{\beta}_L) = \ell(\hat{\beta}_1 - \mathbf{c}, \ldots, \hat{\beta}_L - \mathbf{c})$ for any vector $\mathbf{c} \in \mathbb{R}^d$, the solution can be improved by taking $\tilde{c}_j = \arg \min_{c_j} \sum_{i=1}^{L} (\hat{B}_{jl} - c_j)^2$, i.e. $\tilde{c}_j = \hat{B}_{j.}$. Hence, $\hat{B}_{gS}$ necessarily satisfies the symmetric constraint $\sum_{l=1}^{L} \hat{B}_{gS,l} = 0$ or, equivalently, $\hat{B}_{gS} \mathbf{1} = 0$ (zero mean rows).

As usual for convex relaxation methods, one needs some (mild) constraints on the design. In particular, we assume the following assumption on the marginal distribution $P_X$:

**Assumption (C).** Assume that for all (generally dependent) components $X_j$’s of a random features vector $\mathbf{X} \in \mathbb{R}^d$,

1. $EX_j^2 = 1$ ($X_j$’s are scaled)
2. there exist constants $\kappa_1, \kappa_2, w > 1$ and $\gamma \geq 1/2$ such that $E(|X_j|^p)^{1/p} \leq \kappa_1 \mathbb{P}^{-\gamma}$ for all $2 \leq p \leq \kappa_2 \ln(\text{wd})$ ($X_j$’s have polynomially growing moments up to the order $\ln d$)

In particular, Assumption (C) evidently holds for (scaled) Gaussian and sub-Gaussian $X_j$’s with $\gamma = 1/2$ for all moments. Assumption (C) ensures that for $n \geq C_1 (\ln d)^{\max(2\gamma-1,1)}$,
\[
E \max_{1 \leq j \leq d} \frac{1}{n} \sum_{i=1}^{n} X_j^2 \leq C_2
\]
for some constants $C_1 = C_1(\kappa_1, \kappa_2, w, \gamma)$ and $C_2 = C_2(\kappa_1, \kappa_2, w)$ ([23] proof of Theorem A1). Moreover, [14] might be violated if the moments condition in Assumption (C) holds only up to the order of $\ln(\text{wd}) / \ln \ln(\text{wd})$. We will need [14] in the proof of the upper bound for misclassification excess risk of a general multinomial logistic group Slope classifier [13].

For simplicity of exposition, in what follows we consider $\gamma = 1/2$ corresponding to $n \geq C_1 \ln d$, where $C_1$ is given in [23].

**Theorem 4.** Consider a $d_0$-sparse multinomial logistic regression [1]-[2].

Apply the multinomial logistic group Slope classifier [13] with $\lambda_j$’s satisfying
\[
\max_{1 \leq j \leq d} \sqrt{L + \ln(d/j)} \leq C_0 \sqrt{n}
\]
with the constant $C_0$ derived in the proof.

Assume Assumptions (A)-(C) and let $n \geq C_1 \ln d$.

Then,
\[
\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{gS}, \eta^*) \leq C(\delta) \left( \sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right)^{\frac{2(\alpha+1)}{\alpha+2}}
\]
for some constant $C(\delta)$ depending on $\delta$.

We now consider two specific choices of $\lambda_j$’s:

1. **Equal $\lambda_j$ (multinomial logistic group Lasso).**
   Take
   \[
   \lambda = C_0 \sqrt{\frac{L + \ln d}{n}}
   \]
   to satisfy (15). Note that $\sum_{j=1}^{d_0} \frac{1}{\sqrt{j}} \leq 2 \sqrt{d_0}$ that yields the following corollary of Theorem 4.

**Corollary 1.** Consider a $d_0$-sparse multinomial logistic regression [1]-[2]. Apply the multinomial logistic group Lasso classifier [12] with $\lambda$ from (16).

Then, under Assumptions (A)-(C) and $n \geq C_1 \ln d$,
\[
\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{gL}, \eta^*) \leq C(\delta) \left( \frac{d_0(L-1) + d_0 \ln(d\epsilon)}{n} \right)^{\frac{\alpha+1}{\alpha+2}}
\]
for all $1 \leq d_0 \leq \min(d, n)$ and all $\alpha \geq 0$.

Thus, unless $d$ grows faster than exponentially with $n$, the multinomial logistic group Lasso classifier $\hat{\eta}_{gL}$ achieves a minimax order for large number of classes (see Section III), whereas for small $L$ it rate-optimal for sparse cases, where $d_0 \ll d$, but only sub-optimal (up to an extra logarithmic loss) for dense cases, where $d_0 \sim d$. We conjecture that similar to the results of [10] for Gaussian regression, $\hat{\eta}_{gL}$ with adaptively chosen $\lambda$ can achieve the minimax rate in the latter case as well but the proof of this conjecture is beyond the scope of the paper.

2. **Variable $\lambda_j$’s.** Consider
\[
\lambda_j = C_0 \sqrt{\frac{L + \ln(d/j)}{n}}
\]
that evidently satisfies [13]. One can also verify that
\[\sum_{j=1}^{d_0} \sqrt{L + \ln(d/j)} \leq 2L \sqrt{d_0(L + \ln(d_0))} \]
\[\leq 4 \sqrt{d_0 \left( L - 1 + \ln \left( \frac{d}{d_0} \right) \right)} \]

Theorem 4 implies then:

**Corollary 2.** Consider a \(d_0\)-sparse multinomial logistic regression ([2]). Apply the multinomial logistic group Slope classifier ([13]) with \(\lambda_j\)'s from [17].

Then, under Assumptions (A)-(C) and \(n \geq C_1 \ln d\),
\[\mathop{\sup}_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\tilde{\eta}, \eta^*) \leq C(\delta) \left( \frac{d_0(L - 1) + d_0 \ln \left( \frac{d}{d_0} \right)}{n} \right)^{\frac{3}{2}} \]
for all \(1 \leq d_0 \leq \min(d, n)\) and all \(\alpha \geq 0\).

Hence, if the number of features grows at most exponentially with \(n\), the multinomial logistic group Slope classifier with variable \(\lambda_j\)'s from [17] is adaptively rate-optimal for both small and large number of classes, and, unlike the penalized likelihood classifier \(\tilde{\eta}_M\), is computationally feasible.

**APPENDIX**

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

We first introduce several notations that will be used throughout the proofs. Let \(|a|_2\) be the Euclidean norm of a vector \(a\), \(|A|_2\) the operator norm of a matrix \(A\) and \(|A|_F\) its Frobenius norm. Denote \(|g|_{L_2} = (\int_X g^2(x)dx)^{1/2}\) for a standard \(L_2\)-norm of a function \(g\) and \(|g|_{L_2}(P_X) = (\int_X g^2(x)dP_X(x))^{1/2}\) for the \(L_2\)-norm of \(g\) weighted w.r.t. the marginal distribution \(P_X\) of \(X\). In addition, the \(L_\infty\)-norm \(|g|_{L_\infty} = \sup_{x \in X} |g(x)|\).

**Proof of Theorem 2** On the one hand, it is obvious that feature selection and classification in multiclass case cannot be simpler than in binary, for which the results of [3] Section 6) imply that the minimax lower bound for misclassification risk over a set of \(d_0\)-sparse linear classifiers is of the order \(\sqrt{d_0 \ln(\frac{d}{d_0})}\).

On the other hand, for a given model \(M\) of size \(d_0\), consider the corresponding set of \(d_0\)-dimensional linear \(L\)-class classifiers \(\mathcal{C}_M = \{\eta(x) \in \mathcal{C}_L(d_0) : B \in \mathcal{B}_M\}\).

Obviously, 
\[\inf_{\tilde{\eta}} \mathop{\sup}_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\tilde{\eta}, \eta^*) \geq \inf_{\tilde{\eta}} \mathop{\sup}_{\eta^* \in \mathcal{C}_M} \mathcal{E}(\tilde{\eta}, \eta^*)\]
From the general results of [19] Theorem 5, it follows that
\[\inf_{\tilde{\eta}} \mathop{\sup}_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C \left( \frac{d_N(C_M^\alpha)}{n} \right)^{\frac{\alpha + 1}{\alpha + 2}}\]
for some \(C > 0\), where \(d_N(C_M^\alpha)\) is Natarajan dimension of \(C_M^\alpha\). Natarajan dimension is one of common multiclass extensions of VC-dimension ([30]) and ([13]) generalizes the corresponding well-known results for binary classification derived in terms of VC (e.g., [20] Chapter 14).

To complete the proof we use the bounds for Natarajan dimension of the set of \(d_0\)-dimensional linear \(L\)-classifiers established in [18] Theorem 3.1], namely, \(d_0(L - 1) \leq d_N(C_M^\alpha) \leq O(d_0L \ln(d_0L))\).

**Proof of Theorem 3** Let \(KL(p_1, p_2) = \sum_{l=1}^L p_{1l} \ln \left( \frac{p_{1l}}{p_{2l}} \right)\) and \(H^2(p_1, p_2) = \frac{1}{2} \sum_{l=1}^L (\sqrt{p_{1l}} - \sqrt{p_{2l}})^2\) be respectively the Kullback-Leibler divergence and the square Hellinger distance between two multinomial distributions with success probabilities vectors \(p_1\) and \(p_2\). Let also \(d_{KL}(f_B, f_{B\tilde{\eta}}) = \int KL(p_1(x), p_2(x))dP_X(x)\) and \(d_{H}^2(f_B, f_{B\tilde{\eta}}) = \int H^2(p_1(x), p_2(x))dP_X(x)\) be the corresponding Kullback-Leibler divergence and square Hellinger distance between \(f_B\) and \(f_{B\tilde{\eta}}\).

One can verify that for \(p_1\) and \(p_2\) satisfying Assumption (A), \(KL(p_1, p_2) \leq \frac{4(1-\delta)^2}{\delta^2} H^2(p_1, p_2)\) and, therefore, \(d_{KL}(f_B, f_{B\tilde{\eta}}) \leq \frac{4(1-\delta)^2}{\delta^2} d_{H}^2(f_B, f_{B\tilde{\eta}})\).

A common approach to derive the upper bounds for misclassification risk is to convert them to the bounds of some related surrogate risk (see Section 1) which can be established by various existing methods. See, e.g., [8], [9], [41], [42] among many others.

Thus, utilizing the results of [8] Section 5.2) for multiclass logistic regression corresponding to the logistic surrogate loss and applying then their Theorem 3.11 with the calibration function \(\delta'(\epsilon) = 0.5 ((1 - \epsilon) \ln(1 - \epsilon) + (1 + \epsilon) \ln(1 + \epsilon)) \geq 0.5\epsilon^2\) and \(\alpha' = \frac{\alpha}{\alpha + 1}\) implies that under the low-noise condition ([10]-[11]),
\[\mathcal{E}(\tilde{\eta}_{\bar{M}}, \eta^*) \leq C \left( Ed_{KL}(f_B, f_{B\tilde{\eta}}) \right)^{\frac{\alpha + 1}{\alpha + 2}} \leq C \left( \frac{1}{\delta^2} Ed_{H}^2(f_B, f_{B\tilde{\eta}}) \right)^{\frac{\alpha + 1}{\alpha + 2}}\]
and it is, therefore, sufficient to bound the square Hellinger risk \(Ed_{H}^2(f_B, f_{B\tilde{\eta}})\).

We will show now that the penalty ([7]) falls within a general class of penalties considered in Theorem 3.1 from the supplementary material which extends Theorem 1 of [40].
under weaker conditions. Applying Theorem S1, we find then an upper bound for $E d^2_H(f_B, f_{B_0})$.

It is easy to verify that

$$H^2(p_1, p_2) \geq \frac{1}{8} |p_1 - p_2|^2 \quad (20)$$

Furthermore, using the standard inequality $\ln(1+t) \leq t$, under Assumption (A) we have

$$|\ln f_{B_2}(x, y) - \ln f_{B_1}(x, y)| = \left| \sum_{l=1}^{L} \xi_l |p_{2l}(x) - p_{1l}(x)| \right| \leq \frac{1}{\delta} \max_{1 \leq l \leq L} |p_{2l}(x) - p_{1l}(x)| \quad (21)$$

where recall that $\xi \in \{0, 1\}^L$ is the indicator vector assigned to $y$.

For a given model $M$, consider the set of coefficient matrices $\mathcal{B}_M$ defined in Section II-B. One can easily verify that under Assumption (A), for any $B_1, B_2 \in \mathcal{B}_M$ with columns $\beta_1$'s and $\beta_2$'s respectively and the corresponding probability vectors $p_1(x), p_2(x)$

$$\delta(1-\delta) \|\beta_2 - \beta_1\|^T x \leq |p_{2l}(x) - p_{1l}(x)| \leq \frac{1}{4} \|\beta_2 - \beta_1\|^T x \quad (22)$$

for all $l = 1, \ldots, L - 1$ and any $x \in \mathcal{X}$.

Since $X_j$ are linearly independent, the matrix $G = E_X(XX^T)$ is positive definite. Consider the weighted Frobenius matrix norm $\|B\|_G = \sqrt{\text{tr}(B^T G B)}$. In particular, (22) implies

$$\sum_{l=1}^{L} |p_{2l} - p_{1l}| \leq \frac{1}{4} \|\beta_2 - \beta_1\|^T x \quad (23)$$

(recall that $\beta_{1l} = \beta_{2l} = 0$).

For each matrix $B_0 \in B_M$, consider the corresponding Hellinger ball $\mathcal{H}_{f_{B_0}, r} = \{f_B : d_H(f_B, f_{B_0}) \leq r, B \in B_M\}$. From (20) and (23) it then follows that if $f_B \in \mathcal{H}_{f_{B_0}, r}$, the corresponding $B \in B_M$ lies in the matrix ball $B_{f_{B_0}, r} = \{B \in \mathbb{R}^{[M] \times L} : \|B - B_0\|_G \leq r\}$ with $r = \frac{\sqrt{2}\sqrt{2r}}{2(\sqrt{1+\frac{1}{2}})}$. Furthermore, for any $x$ and any $1 \leq l \leq L - 1$, Cauchy–Schwarz inequality imply that

$$\sum_{l=1}^{L} |p_{2l}(x) - p_{1l}(x)|^2 \leq \frac{1}{4} \sum_{l=1}^{L-1} (\beta_{2l} - \beta_{1l})^T G(\beta_{2l} - \beta_{1l}) \cdot |G^{-1/2}x|^2 = \frac{1}{4} \|B_1 - B_2\|^2 \cdot |G^{-1/2}x|^2 \quad (24)$$

Let $N(B_{f_{B_0}, r}, \cdot, |G|, \epsilon)$ be the $\epsilon$-covering number of $B_{f_{B_0}, r}$ w.r.t. the $|G|$ norm. Note that since $\beta_L = 0$, the dimension of the vector space containing $B_{f_{B_0}, r}$ is $(L - 1)|M|$. We can use then the well-known results for the covering number of a ball to have

$$N(B_{f_{B_0}, r}, \cdot, |G|, \epsilon) = N(B_{f_{B_0}, r}, \cdot, |G|, \epsilon/r') \leq \left(1 + \frac{2r'}{\epsilon}\right)^{(L-1)|M|} \leq \left(\frac{3r'}{\epsilon}\right)^{(L-1)|M|}$$

(see, e.g., [39], Example 5.8).

Consider now the bracketing number $N([\mathcal{F}_{f_{B_0}, r}], \|\cdot\|_{L_2}, \epsilon)$, where $\mathcal{F}_{f_{B_0}, r} = \{f_B : f_B \in \mathcal{H}_{f_{B_0}, r}\}$. By (21) and (24), we have

$$|\ln f_{B_2}(x, y) - \ln f_{B_1}(x, y)| \leq \frac{1}{2 \delta} \|B_1 - B_2\|_G \cdot |G^{-1/2}x|^2 \quad (25)$$

Let $\mathcal{B}_k, k = 1, \ldots, N(B_{f_{B_0}, r}, \cdot, |G|, \epsilon \delta)$ be the cover set of $B_{f_{B_0}, r}$ w.r.t. the $|G|$ norm. Define $g_k^u(x, y) = \log f_{B_k}(x, y) - \frac{\epsilon}{2} |G^{-1/2}x|^2 |G^{-1/2}x|^2$ and $g_k^l(x, y) = \log f_{B_k}(x, y) + \frac{\epsilon}{2} |G^{-1/2}x|^2$. For each pair we have

$$\|g_k^u - g_k^l\|_{L_2} = \epsilon \|G^{-1/2}x\|_{L_2} = \epsilon \sqrt{E_X(X^T G^{-1} X)} = \epsilon$$

Finally, for any $\log f_B \in \mathcal{F}_{f_{B_0}, r}$, take $\{g_k^u, g_k^l\}$ such that $\|B - B_k\|_G < \epsilon \delta$. Therefore,

$$g_k^u(x, y) - \log f_B(x, y) \geq \frac{\epsilon}{2} |G^{-1/2}x|^2 \geq \frac{1}{2 \delta} \|B_k - B_0\|_G \cdot |G^{-1/2}x|^2 \geq 0,$$

$$g_k^l(x, y) - \log f_B(x, y) \leq \frac{1}{2 \delta} \|B_k - B_0\|_G \cdot |G^{-1/2}x|^2 - \frac{\epsilon}{2} |G^{-1/2}x|^2 \leq 0.$$
which imply that \( g_k^L(x, y) \leq \log f_B(x, y) \leq g_k^L(x, y) \). Hence, \( \{g_k^L, g_k^L\} \) are \( \epsilon \)-brackets that cover \( F_{\mathcal{B}, r} \) under \( \| \cdot \|_L \), so
\[
N_\epsilon(\mathcal{F}_{\mathcal{B}_0, r}, \| \cdot \|_L, \epsilon) \leq N(B_{\mathcal{B}_0, r}, \| \cdot \|_L, \epsilon) \leq \left( \frac{3r}{\epsilon} \right)^{(L-1)|M|} = \left( \frac{6\sqrt{2}}{\delta(1-\delta)} \right)^{(L-1)|M|}.
\]

The considered family of sparse multinomial logistic regression models satisfies then Assumption (D) in the supplementary material with \( A_m = \frac{18\sqrt{2}}{\epsilon(1-\delta)} \) and \( m_M = (L-1)|M| \). Note also that by Assumption (A), \( |\sum_{l=1}^L \xi_l \ln p_l(x)| \leq \max_{1 \leq l \leq L} \ln p_l(x) \leq \log(1/\delta) \) for all \( x \in \mathcal{X} \). Apply now Theorem [14] from supplementary material for a penalized maximum likelihood model selection procedure [5] with a complexity penalty \( \text{Pen}(|M|) = C_1 m_M \ln A_m + C_2 C_M \leq \hat{C}_1 (L-1)|M| + C_2 |M| \ln \left( \frac{de}{|M|} \right) \), where \( C_M = |M| \ln \left( \frac{de}{|M|} \right) \).

Thus,
\[
\text{Ed}_H^2(f_{\hat{B}_{SS}}, f_B) \leq C(\delta) \frac{\text{Pen}(d_0)}{n} \leq C(\delta) \frac{(L-1)d_0 + d_0 \ln \left( \frac{de}{d_0} \right)}{n},
\]
that together with (19) complete the proof.

**Proof of Theorem 2** First, recall that from (19) it follows that
\[
E(\tilde{\eta}_{GS}, \eta^*) \leq C \left( \text{Ed}_{KL}(f_B, f_{\hat{B}_{SS}}) \right)^{\frac{2}{1+2}},
\]
and, thus, it is sufficient to bound the Kullback-Leibler risk \( \text{Ed}_{KL}(f_B, f_{\hat{B}_{SS}}) \). For this purpose, we extend the corresponding results of [5] for logistic Slope to its group analogue in multinomial logistic regression model.

As we have mentioned, the solution of (13) satisfies the symmetric constraint \( \sum_{l=1}^L \beta_{gS, l} = 0 \). Let \( \theta_l(x) = \beta_l^T x, l = 1, \ldots, L \) with the constraint \( \sum_{l=1}^L \theta_l(x) = 0 \). Thus, \( p_l(x) = e^{\theta_l(x)} / \sum_{l=1}^L e^{\theta_l(x)} \) and in terms of \( \theta(x) \), the likelihood (2) is \( \ell(\theta(x)) = \sum_{l=1}^L \eta_l \theta_l(x) - \ln \left( \sum_{l=1}^L e^{\theta_l(x)} \right) \) which is Lipschitz with constant 2, i.e. \( |\ell(\theta_1(x)) - \ell(\theta_2(x))| \leq 2|\theta_1(x) - \theta_2(x)|_2 \). Furthermore, similar to Lemma 1 of [2] for binary logistic regression, re-writing the Kullback-Leibler divergence \( KL(p_1(x), p_2(x)) \) in terms of \( \theta(x) \) and expanding it in (multivariate) Taylor series, one can verify that under Assumption (A), \( KL(\theta_1(x), \theta_2(x)) \geq \frac{1}{2\pi} |\theta_1(x) - \theta_2(x)|^2 \) and, therefore, \( dk_1L(f_B, f_{\hat{B}_S}) \geq \frac{1}{2\pi} |\theta_1(x) - \theta_2(x)|_2^2 \) (a multivariate analogue of Bernstein condition in terminology of [5]). Lipschitz and Bernstein conditions allow us to apply the general approach of [5] and to extend their results to multinomial logistic group Lasso and group Slope. In particular, Assumption (A) corresponds to the bounded case considered there.

Let \( \mathcal{B} \) be a set of matrices \( B \) with zero mean rows, i.e. \( \mathcal{B} = \{ B \in \mathbb{R}^{d \times L} : B1 = 0 \} \). For a given regression coefficients matrix \( B \in \mathcal{B} \) with (zero mean) rows \( B_j \), define its group Slope norm \( |B|_\lambda = \sum_{j=1}^d \lambda_j |B(j)| \), where recall that \( |B(j)| \geq \ldots \geq |B(d)| \) are the descendingly ordered \( l_2 \)-norms of \( B_j \)’s, and consider the corresponding unit ball \( \mathcal{B}_\lambda \).

To derive an upper bound on \( \text{Ed}_{KL}(f_B, f_{\hat{B}_{SS}}) \) we define the following quantities along the lines of [5].

\[
\tilde{\text{Rad}}(\mathcal{B}_\lambda) = \frac{1}{n} \sup_{B \in \mathcal{B}_\lambda} \sum_{i=1}^n \sum_{l=1}^L \sigma_{il} |\hat{B}_l^T X_i|, \quad X_1 = x_1, \ldots, X_n = x_n = \mathcal{X} \)
\[
= \frac{1}{n} \sup_{B \in \mathcal{B}_\lambda} \text{tr}(S^T X^T),
\]
where the elements \( \sigma_{il} \)’s of \( S \in \mathbb{R}^{n \times L} \) are i.i.d. Rademacher random variables with \( P(\sigma_{il} = 1) = P(\sigma_{il} = -1) = 1/2 \), and
\[
\text{Rad}(\mathcal{B}_\lambda) = E_{\mathcal{X}} \left\{ \tilde{\text{Rad}}(\mathcal{B}_\lambda) \right\}
\]
be the Rademacher complexity of \( \mathcal{B}_\lambda \).

Define a complexity function
\[
r(\rho) = \sqrt{\frac{C_0 \text{Rad}(\mathcal{B}_\lambda) \rho}{2\delta^2 \sqrt{n}}}, \quad \rho > 0,
\]
where the exact value of \( C_0 > 0 \) is specified in [5].

Let \( \mathcal{M}(\rho) = \{ B \in \mathcal{B} : |B|_\lambda = \rho, \sum_{j=1}^d |B(j)|_2^2 \leq r^2(2\rho) \} \). For a given matrix \( B \in \mathcal{B} \) define \( \Gamma_B(\rho) = \{ B' \in \mathcal{B} : |B' - B|_\lambda < \frac{\rho}{m} \} \), where the subdifferential \( \partial |\lambda|(B') = \{ B'' \in \mathcal{B} : |B' + B''|_\lambda = |B'|_\lambda \geq \text{tr}((B')^T B'') \} \).

The sparsity parameter is
\[
\Delta(\rho) = \inf_{B' \in \mathcal{M}(\rho)} \sup_{H \in \Gamma_B(\rho)} \langle H, B' \rangle = \inf_{B' \in \mathcal{M}(\rho)} \sup_{H \in \Gamma_B(\rho)} \text{tr}(H^T B')
\]

Let \( B \in \mathcal{B} \) be \( d_0 \)-sparse and define
\[
\rho^* = \frac{C_0}{8000^2} \frac{\text{Rad}(\mathcal{B}_\lambda) \left( \sum_{j=1}^d \lambda_j / \sqrt{j} \right)^2}{\sqrt{n}}
\]
A straightforward extension of Lemma 4.3 of [24] for matrices implies that \( \Delta(\rho^*) > \frac{1}{2} \rho^* \) and, therefore, we can apply the following Lemma [1] which can be viewed as an extension of Theorem 2.2 (or more general Theorem 9.2) of [5] for our case. 

---

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3075137, IEEE Transactions on Information Theory
Lemma 1. Let $B \in \mathcal{B}$ be $d_0$-sparse and let $\lambda_j$’s be such that $\text{Rad}(B_{\lambda}) \leq \frac{7}{20} \sqrt{n}$. If $\rho^*$ defined in (25) satisfies $\Delta(\rho^*) \geq \frac{4}{5} \rho^*$, then

$$Ed_{KL}(f_B, f_{B_{\delta}}) \leq C(\delta) \left( \sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right)^2$$

(26)

for some constant $C(\delta)$ depending on $\delta$.

To satisfy the conditions of Lemma 1 and to complete the proof using [15], we need to find an upper bound for the Rademacher complexity $\text{Rad}(B_{\lambda})$ :

Lemma 2.

$$\text{Rad}(B_{\lambda}) \leq C \max_{1 \leq j \leq d} \frac{\sqrt{L + \ln \left( \frac{d}{j} \right)}}{\lambda_j},$$

(27)

where the exact constant $C$ is given in the proof.

Proof of Lemma 7 The proof is an extension of the proof of Theorem 9.2 in the supplementary material of [5] for the multiclass framework. In a slightly more general version of Proposition 9.1 of [5], we define the following event $\Omega_0$ for $t \geq 1$:

$$\Omega_0 = \left\{ \forall B_i \in \mathcal{B}, \left| \frac{1}{n}(\ell(B_i') - \ell(B)) - \mathbb{E}[\ell(B_i') - \ell(B)] \right| \leq \frac{7}{20} \delta^2 \max \left( r(2 \max \{|B_i' - B|_2, t \rho^*|^2, \sum_{i=1}^{L} ||(B_i - B_i')^2||_{L_2} \right) \right\}.$$ (28)

(Proposition 9.1 of [5] considers only $\Omega_0$).

As stated above, Assumption (A) implies the required Bernstein condition. The condition $\text{Rad}(B_{\lambda}) \leq \frac{7}{20} \sqrt{n}$ is needed for adjusting the scale of the norm w.r.t. to the loss required in Theorem 9.2 of [5]. Under these two conditions, we can follow the proof of Proposition 9.1 of [5] to get

$$d_{KL}(f_B, f_{B_{\delta}}) \leq 2\delta^2 r^2 (2\rho^*)^2 \leq C \frac{\rho^* \text{Rad}(B_{\lambda})}{\sqrt{n}}.$$ on the event $\Omega_0$ provided $\Delta(\rho^*) \geq \frac{4}{5} \rho^*$.

To extend the proof for $t > 1$, note that $t \rho^* \geq \rho^*$. Since $\Delta(\rho^*) \geq \frac{4}{5} \rho^*$, when $|B_i' - B|_{2} \geq t \rho^* \geq \rho^*$ we still have $\Delta(|B_i' - B|_{2}) \geq \frac{4}{5}|B_i' - B|_{2} \geq \rho^*$ (see [5], Lemma A.1]). Thus, following the arguments of Proposition 9.1, on the event $\Omega_0$ we have

$$d_{KL}(f_B, f_{B_{\delta}}) \leq 2\delta^2 r^2 (2\rho^*)^2 \leq C \frac{\rho^* \text{Rad}(B_{\lambda})}{\sqrt{n}} t.$$

To bound the probability of $\Omega_0$, we follow Proposition 9.3 of [5]. We consider the subsets

$$F_{j,i} = \left\{ B : \rho_{j,i} \leq |B_i' - B|_{\lambda} \leq \rho_j, t_{i-1}(\rho_{j,i}) \leq \sum_{i=1}^{L} ||(B_i - B_i')^2||_{L_2} \leq t_{i}(\rho_{j,i}) \right\},$$

where $\rho_j = 2^i \rho^*$ and $t_{j}(\rho) = 2^j r(\rho), i, j = 0, 1, \ldots$. Replace $\rho_j$ with $t \rho_j$ and go along the lines of the proof of Proposition 9.3 of [5] with the extended contraction inequality for Rademacher complexities for vector-valued Lipschitz functions of [26] to get

$$P\left( d_{KL}(f_B, f_{B_{\delta}}) \geq 2\delta^2 r^2 (2\rho^*)^2 t \right) \leq 2 \sum_{j=0}^{\infty} \sum_{i \in I_j} \exp \left( - \frac{1}{48} \tilde{C}(\delta) \frac{7}{20} \delta^2 n (2^j r(t^2 \rho^*))^2 \right).$$

Thus,

$$P \left( d_{KL}(f_B, f_{B_{\delta}}) \geq 2C_0 \frac{\rho^* \text{Rad}(B_{\lambda})}{\sqrt{n}} t \right) \leq 2 \sum_{j=0}^{\infty} \sum_{i \in I_j} \exp \left( - \frac{1}{48} \tilde{C}(\delta) \frac{7}{20} \delta^2 \sqrt{n} 2^{4j} \frac{1920}{7} \delta^2 t \text{Rad}(B_{\lambda}) \rho^* \right) \leq 2 \sum_{j=0}^{\infty} 2^{-4j} \exp \left( - \tilde{C}(\delta) \sqrt{n} 2^{4j} t \text{Rad}(B_{\lambda}) \rho^* \right) \leq 4 \sum_{j=0}^{\infty} 2^{-4j} \exp \left( - \tilde{C}(\delta) \sqrt{n} t \text{Rad}(B_{\lambda}) \rho^* \right) = 4 \exp \left( - \tilde{C}(\delta) \sqrt{n} t \text{Rad}(B_{\lambda}) \rho^* \right).$$
Hence,
\[
Ed_{KL}(f_B, f_{B^*}) \leq 8 \left( \frac{1}{C(\delta)} + \int_{c^* \in \Theta} \exp(-\tilde{C}(\delta)\sqrt{n} \text{Rad}(B_{\lambda})t \rho^*)dt \right) \cdot C \frac{\text{Rad}(B_{\lambda})\rho^*}{\sqrt{n}}
\]
\[
\leq C(\delta) \frac{\text{Rad}(B_{\lambda})\rho^*}{\sqrt{n}}
\] (29)
Substituting \( \rho^* \) from (25) into (29) under the conditions of the lemma completes the proof.

**Proof of Lemma 2** Recall that \( B1 = 0 \) for \( B \in B \). Define the matrix \( U \in \mathbb{R}^{L \times (L-1)} \) which (orthonormal) columns are the \( L - 1 \) eigenvectors of the matrix \( I_L - \frac{1}{2} I \), corresponding to the eigenvalue 1. One can easily verify that \( B = BUU^T \).

Then,
\[
\sup_{B \in \mathcal{B}_\lambda} \text{tr}(\Sigma B^T X^T) = \sup_{B \in \mathcal{B}_\lambda} \text{tr}(X^T \Sigma U U^T B^T) = \sup_{B \in \mathcal{B}_\lambda} \text{tr}(K^T U U^T B^T) = \sup_{B \in \mathcal{B}_\lambda} \sum_{j=1}^{d} K_j^T U B_j,
\]
where \( K = U^T X \). Let \( |K|_j = |K_j|_2 \). By Cauchy-Schwartz inequality and the definition of the group slope norm \( |B|_\lambda \), we have
\[
\sup_{B \in \mathcal{B}_\lambda} \sum_{j=1}^{d} K_j^T U B_j \leq \sup_{B \in \mathcal{B}_\lambda} \sum_{j=1}^{d} (UB)_j, 2 \cdot |K_j|_2
\]
\[
= \sup_{B \in \mathcal{B}_\lambda} \sum_{j=1}^{d} |B_j| \cdot |K|_j = \sup_{B \in \mathcal{B}_\lambda} \sum_{j=1}^{d} \lambda_j |B_j| \cdot |K|_j \frac{\lambda_j}{\lambda_j}
\]
\[
\leq \sup_{B \in \mathcal{B}_\lambda} \sum_{j=1}^{d} \lambda_j |B_j| \cdot |K|_j \frac{\lambda_j}{\lambda_j} \leq \max_{1 \leq j \leq d} \frac{|K|_j}{\lambda_j}
\]
Thus, \( \text{Rad}(B_{\lambda}) \leq E_{\Sigma} \left\{ \max_{1 \leq j \leq d} \frac{1}{\sqrt{n}} \frac{|K|_j}{\lambda_j} \left| X \right| \right\} \).

Let \( x_j \) be the columns of \( X \). By its definition, \( |U|_2^2 = (L - 1) \) and \( |U|_2 = 1 \). We can apply then the results of p. 8] to get conditionally on \( X \)
\[
P \left( \frac{|K|_j}{|x_j|} \geq t \sqrt{L + \ln(d/j)} \right) = P \left( \frac{|K|_j}{|x_j|} \geq t \sqrt{L - 1 + \frac{t}{\sqrt{2}} \ln \left( \frac{d}{j} \right)} \right)
\]
\[
\leq P \left( \frac{|K|_j}{|x_j|} \geq \frac{t}{2} \sqrt{L - 1 + \frac{t}{\sqrt{2}} \ln \left( \frac{d}{j} \right)} \right)
\]
\[
\leq P \left( \frac{|K|_j}{|x_j|} \geq \sqrt{L - 1 + \frac{t}{\sqrt{2}} \ln \left( \frac{d}{j} \right)} \right)
\]
\[
\leq 2 e^{-c t^2 \ln \left( \frac{d}{j} \right)} \leq \left( \frac{d}{j} \right)^{-c t^2}
\]
for all \( t \geq \sqrt{2} \) and a certain constant \( c > 0 \).

Hence, by standard probabilistic arguments, for all \( t \geq \max(\sqrt{2}, \frac{2}{\sqrt{e}}) \)
\[
P \left( \frac{|K|_j}{\lambda_j} > t \max_{1 \leq j \leq d} \frac{|x_j|}{\sqrt{L + \ln(d/j)}} \right)
\]
\[
\leq 2 \left( \frac{d}{j} \right) \left( \frac{d}{j} \right) \frac{de}{j} \leq 2 \left( \frac{d}{j} \right)^{-j (c t^2 - 1)}
\]
\[
\leq 2 \left( \frac{d}{j} \right)^{-j c t^2}
\]
and applying the union bound,
\[
P \left( \max_{1 \leq j \leq d} \frac{|K|_j}{\lambda_j} > t \max_{1 \leq j \leq d} \frac{|x_j|}{\sqrt{L + \ln(d/j)}} \right)
\]
\[
\leq 2 \sum_{j=1}^{d} \left( \frac{d}{j} \right)^{-j c t^2} \leq 2 \sum_{j=1}^{d} e^{-j c t^2} \leq 2 \frac{e^{-c t^2}}{1 - e^{-c t^2}}
\]
\[
\leq 4 e^{-c t^2}
\]
Therefore,
\[
E_{\Sigma} \left( \max_{1 \leq j \leq d} \frac{|K|_j}{\lambda_j} \left| \max_{1 \leq j \leq d} \frac{|x_j|}{\sqrt{L + \ln(d/j)}} \right| \right)
\]
\[
= \int_0^\infty P \left( \max_{1 \leq j \leq d} \frac{|K|_j}{\lambda_j} > t \max_{1 \leq j \leq d} \frac{|x_j|}{\sqrt{L + \ln(d/j)}} \right) dt
\]
\[
\leq \max \left( \sqrt{2}, \frac{2}{\sqrt{e}} \right) + 4 \int_0^\infty e^{-c t^2} dt
\]
Thus,
\[
\text{Rad}(B_{\lambda}) \leq C \frac{1}{\sqrt{n}} \max_{1 \leq j \leq d} \frac{|x_j|}{\sqrt{L + \ln(d/j)}} \lambda_j
\]
and by [14], \( \text{Rad}(B_{\lambda}) \leq C \max_{1 \leq j \leq d} \sqrt{L + \ln(d/j)} \lambda_j \).

**ACKNOWLEDGMENT**

The work was supported by the Israel Science Foundation (ISF), grant ISF-58918. The authors would like to thank Aryeh Kontorovich, Steve Hanneke and Boaz Nadler for valuable remarks. Helpful comments of the anonymous referees are gratefully acknowledged.

**REFERENCES**


Vadim Grinshtein received M.Sc. (1986) in Mathematics from Simferopol State University, USSR and Ph.D. (2001) in Mathematics from Tel Aviv University. He works at the Department of Mathematics and Computer Science, The Open University of Israel. His research interests include spectral theory of operators and mathematical statistics.

Tomer Levy received M.Sc. (2019) in Applied Mathematics and Computer Science from Weizmann Institute of Science, Israel. He is a Ph.D. candidate in the Department of Statistics and Operations Research, Tel Aviv University, under the supervision of Prof. Felix Abramovich. His research interests include high-dimensional and multiclass classification.