Report on Multiclass Classification by Sparse Multinomial Logistic Regression

Zequan Xiong

School of Mathematical Sciences

Nankai University

Update: August 15, 2025

Background: High-Dimensional Multiclass Classification

- 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)
- **Recall**: Previous work in 2019[1] only designed Slope for binary classification by Logistic Regression.
- Feature selection for **multiclass classification** has not yet been rigorously well-studied before this paper published and the goal of this paper is to fill the gap.

Background: High-Dimensional Multiclass Classification

- L-class classification, Features $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^d$; Outcome class label $Y \in \{1, \ldots, L\}$. We can model it as

$$Y \mid (\boldsymbol{X} = \boldsymbol{x}) \sim \text{Mult}\left(p_1(\boldsymbol{x}), \ldots, p_L(\boldsymbol{x})\right)$$

  where

$$p_l(\boldsymbol{x}) = P(Y = l \mid \boldsymbol{X} = \boldsymbol{x}), l = 1, \ldots, L$$

  .

- Classifier: $\eta : \mathcal{X} \to \{1, \ldots, L\}$.
- Misclassification Error: $R(\eta) = P(Y \neq \eta(\boldsymbol{x}))$.
- **Optimal Classifier**: $\eta^*(\boldsymbol{x}) = \arg \max_{1 \leqslant l \leqslant L} p_l(\boldsymbol{x})$ with $R\left(\eta^*\right) = 1 - \mathbb{E}_{\boldsymbol{X}} \max_{1 \leqslant l \leqslant L} p_l(\boldsymbol{x})$.

Strategies in Multiclass Classification

A first strategy for multiclass classification is to reduce it to a series of binary classifications.

- OvA: one-vs-all, for L-class classification we need to train L models, where each class is compared against all others.
- OvO: one-vs-one, for L-class classification we need to train $\binom{L}{2} = \frac{L(L-1)}{2}$ models, where all pairs of classes are compared to each other.

Remark For feature selection, all models in OvA and OvO may select **different features**. And that's why we don't use such strategy.

Sometimes OvA and OvO will have better performance in classification, but at the same time computational cost will increase rapidly since we have much more models to train.

Extend Binary Classification to Multiclass

Multiclass Classification ERM

A common approach to design a multiclass classifier $\widehat{\eta}$ is based on **empirical risk minimization** (ERM):

$$\widehat{R}_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} I\left\{Y_i \neq \eta\left(\boldsymbol{x}_i\right)\right\} \tag{1}$$

A crucial drawback of ERM is in minimization of 0-1 loss that makes it computationally infeasible. It is common to replace 0-1 loss by *related convex surrogate*.

For logistic regression, related convex surrogate will be $L(y, f(x)) = \log(1 + \exp(-y \cdot f(x)))$.

Multinomial Logistic Regression

Multinomial logistic regression model:

$$\ln \frac{p_l(\boldsymbol{x})}{p_L(\boldsymbol{x})} = \boldsymbol{\beta}_l^T \boldsymbol{x}, \quad l = 1, \ldots, L-1 \tag{2}$$

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

$$p_l(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\beta}_l^T \boldsymbol{x}\right)}{\sum_{k=1}^{L} \exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right)}, \quad l = 1, \ldots, L, \boldsymbol{\beta}_L = 0 \tag{3}$$

The Bayes classifier is then a linear classifier $\eta^*(\boldsymbol{x}) = \arg\max_{1 \le l \le L} p_l(\boldsymbol{x}) = \arg\max_{1 \le l \le L} \boldsymbol{\beta}_l^T \boldsymbol{x}$.

Penalized Maximum Likelihood Estimation

Log-likelihood function

- Let $B \in \mathbb{R}^{d\times L}$ be the matrix of the regression coefficients in (2) with the columns $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L$ (recall that $\boldsymbol{\beta}_L = \mathbf{0}$) and let $f_B(\boldsymbol{x}, y)$ be the corresponding joint distribution of $(\boldsymbol{X}, Y)$, i.e. $\mathrm{d}f_B(\boldsymbol{x}, y) = \prod_{l=1}^{L} p_l(\boldsymbol{x})^{\xi_l} \mathrm{d}P_X(\boldsymbol{x})$.
- The conditional log-likelihood function is

$$\ell(B) = \sum_{i=1}^{n} \left\{ \boldsymbol{X}_i^T B \boldsymbol{\xi}_i - \ln \sum_{l=1}^{L} \exp\left(\boldsymbol{\beta}_l^T \mathbf{X}_i\right) \right\} \tag{4}$$

- By maximizing $\ell(B)$ we can get B. The MLE $\widehat{\boldsymbol{\beta}}$'s though not available in the closed form, can be nevertheless obtained numerically by the fast iteratively reweighted least squares algorithm.

Sparsity of B

- For binary classification, the sparsity can be naturally measured by ℓ_0 : $\|\beta\|_0$.
- For multiclass case we may think of several ways to measure the sparsity of B:
  - The number of non-zero rows of B (row-sparse).
  - Element-wise sparsity.
  - Reduced-rank + Row-sparse. [3]
- This paper chooses row-sparse corresponds to the assumption that part of the features **do not have any impact on classification** at all.

Misclassification Excess Risk Bounds

Assumption

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

Remark This assumption prevents the conditional variances $\mathrm{Var}\left(\xi_l \mid \boldsymbol{X} = \boldsymbol{x}\right) = p_l(\boldsymbol{x})\left(1 - p_l(\boldsymbol{x})\right)$ to be arbitrarily close to zero.

Penalized Maximum Likelihood Model Selection Criterion

- $\mathfrak{M}$: The set of all 2^d possible models $M \subseteq \{1, \cdots, d\}$.
- $\mathcal{B}_M : \{B \in \mathbb{R}^{d \times L} : B_L = \mathbf{0}$ and $B_{j\cdot} = \mathbf{0}$ iff $j \notin M\}$.
- Under the model M, the MLE $\widehat{B}_M$ of B is then

$$\widehat{B}_M = \arg \max_{\widetilde{B} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \boldsymbol{X}_i^T \widetilde{B} \boldsymbol{\xi}_i - \ln \sum_{l=1}^{L} \exp\left(\widetilde{\boldsymbol{\beta}}_l^T \boldsymbol{X}_i \right) \right\} \tag{5}$$

  where $\widetilde{\boldsymbol{\beta}}_l = \widetilde{B}_{\cdot l}, l = 1, \ldots L$ are the columns of $\widetilde{B}$.

- Selection Criterion:

$$\widehat{M} = \arg \min_{M \in \mathfrak{M}} \left\{ -\ell(\widehat{B}) + \mathrm{Pen}(|M|) \right\} \tag{6}$$

From Infeasible to Feasible

In order to make the original problem computational feasible (as 0-1 loss is impossible to optimize), the paper provide the penalty below:

$$\mathrm{Pen}(|M|) = c_1|M|(L-1) + c_2|M| \ln\left(\frac{d\mathrm{e}}{|M|}\right) \tag{7}$$

1. First part is an AIC penalty, this part measures **the number of parameters**.
2. Second part measures the **combinatorial complexity**.

The penalty term transforms the discrete combinatorial problem of ”feature subset selection” into a continuous optimization problem of ”minimizing empirical risk + penalty term,” thereby avoiding exponential enumeration searches.

Upper Bound and Minimax Lower Bound of the Excess Risk

- Nonzero rows of matrix B : r_B.
- The set of all d_0-sparse linear L-class classifiers:

$$\mathcal{C}_L\left(d_0\right)=\left\{\eta(\mathbf{x})=\arg\max_{1\leqslant l\leqslant L}\boldsymbol{\beta}_l^T\mathbf{x}:B\in\mathbb{R}^{d\times L},B_{\cdot L}=\mathbf{0}\text{ and }r_B\leq d_0\right\}$$

- The complexity penalty:

$$\mathrm{Pen}(|M|)=c_1|M|(L-1)+c_2|M|\ln\left(\frac{d\mathrm{e}}{|M|}\right) \tag{8}$$

The absolute constants $c_1, c_2 > 0$ are given in the proof of Theorem 3.

Remark Recall the penalty in [1] below:

$$\mathrm{Pen}(|M|)=c|M|\ln\frac{d\mathrm{e}}{|M|}$$

The Upper Bound of the Excess Risk

Theorem 1: Upper Bound of the Excess Risk

Under Assumption (A) and penalty in (8),

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}\left(\widehat{\eta}_{\widehat{M}}, \eta^*\right) \quad \leqslant \quad C_1(\delta)\sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$$

for some $C_1(\delta)$ depending on δ, for all $1 \leq d_0 \leq \min(d, n)$.

Remark This theorem build the upper bound of the excess risk for multinomial logistic regression for the first time.

Minimax Lower Bound of the Excess Risk

Theorem 2: Minimax Lower Bound of the Excess Risk

Consider a d_0-sparse agnostic multinomial logistic regression model (1)-(2), where $2 \leqslant d_0 \ln\left(\frac{de}{d_0}\right) \leqslant n$ and $d_0(L-1) \leqslant n$. Then for some $C_2 > 0$,

$$\inf_{\widetilde{\eta}} \sup_{\eta^* \in \mathcal{C}_L(d_0), P_X} \mathcal{E}\left(\widetilde{\eta}, \eta^*\right) \geqslant C_2 \sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$$

Remark Combining theorem 1, we now have

$$\text{Excess Risk} \sim \sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$$

Different Cases of L

Let two parts in $\mathrm{Pen}(|M|) = c_1|M|(L-1) + c_2|M| \ln\left(\frac{d\mathrm{e}}{|M|}\right)$ equal to each other. One may obtain $L = 2 + \ln(d/d_0)$.

1. **Small Number of Classes**: $L \leqslant 2 + \ln\left(\frac{d}{d_0}\right)$. Multiclass classification for such a small number of classes is essentially not harder than binary.
2. **Large Number of Classes**: $2 + \ln\left(\frac{d}{d_0}\right) < L \leq \frac{n}{d_0}$. $\mathrm{Pen}(|M|) \sim c|M|(L-1)$ is an AIC type penalty.
3. **Impossible**: $L > \frac{n}{d_0}$. At this time coefficients matrix B will be larger than sample size.

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

Improved Bounds Under Low-Noise Condition

Assumption

Assumption (B) Consider the multinomial logistic regression model (3) and assume that there exist $C > 0, \alpha \geqslant 0$ and $h^* > 0$ such that for all $0 < h \leqslant h^*$,

$$P\left(p_{(1)}(\boldsymbol{X}) - p_{(2)}(\boldsymbol{X}) \leqslant h\right) \leqslant Ch^{\alpha} \tag{9}$$

Where $p_{(i)}(\boldsymbol{X})$ stands for the i-th largest probability.

Remark Assumption (B) implies that with high probability (depending on the parameter α) the most likely class is sufficiently distinguished from others.

Improved Bounds for Misclassification Excess Risks

Theorem 3: Improved Bounds for Misclassification Excess Risks

Consider a d_0-sparse multinomial logistic regression model and let $\widehat{M}$ be a model selected with the complexity penalty $\mathrm{Pen}(|M|)$. Then, under Assumptions (A) and (B), there exists $C(\delta)$ such that

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}\left(\widehat{\eta}_{\widehat{M}}, \eta^*\right) \leqslant C(\delta) \left(\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n} \right)^{\frac{\alpha+1}{\alpha+2}} \tag{10}$$

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

Remark Theorem 1 is a particular case of Theorem 3 with $\alpha = 0$. Theorem 3 tells us the minimax bound can be improved in low-noise condition.

Multinomial Logistic Group Lasso and Slope

Multinomial Logistic Group Lasso

To capture row-sparsity we consider a multinomial logistic group Lasso classifier defined as follows. For a given tuning parameter $\lambda > 0$,

$$\widehat{B}_{gL} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp\left(\widetilde{\boldsymbol{\beta}}_l^T \boldsymbol{X}_i \right) \right) - \boldsymbol{X}_i^T \widetilde{B} \xi_i \right) + \lambda \sum_{j=1}^{d} |\widetilde{B}|_j \right\}$$

Where $|\widetilde{B}|_j = |\widetilde{B}_{j\cdot}|_2$ is the ℓ_2-norm of the j-th row of $\widetilde{B}$.

Multinomial Logistic Group Slope

Multinomial logistic group Slope is a more general variation of multinomial logistic group Lasso. Namely,

$$\widehat{B}_{gS} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp\left(\widetilde{\boldsymbol{\beta}}_l^T \boldsymbol{X}_i \right) \right) - \boldsymbol{X}_i^T \widetilde{B} \boldsymbol{\xi}_i \right) + \sum_{j=1}^{d} \lambda_j |\widetilde{B}|_{(j)} \right\} \tag{11}$$

where the rows' ℓ_2-norms $|\widetilde{B}|_{(1)} \geqslant \ldots \geqslant |\widetilde{B}|_{(d)}$ are the descendingly ordered and $\lambda_1 \geqslant \ldots \geqslant \lambda_d > 0$ are the tuning parameters.

Constraints for Convex Relaxation

Assumption (C) For the components X_j of the random feature vector $X \in \mathbb{R}^d$, the following conditions hold:

1. $\mathbb{E}X_j^2 = 1$ (features are scaled);
2. There exist constants $\kappa_1, \kappa_2, w > 1$ and $\gamma \geqslant 1/2$ such that $(\mathbb{E}|X_j|^p)^{1/p} \leqslant \kappa_1 p^\gamma$ for all $2 \leqslant p \leqslant \kappa_2 \ln(wd)$ (moments grow polynomially up to order $\ln d$).

This assumption ensures that for $n \geqslant C_1(\ln d)^{\max(2\gamma-1,1)}$,

$$\mathbb{E} \max_{1\leqslant j\leqslant d} \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 \leqslant C_2$$

for some constants $C_1 = C_1(\kappa_1, \kappa_2, w, \gamma)$ and $C_2 = C_2(\kappa_1, \kappa_2, w)$

Excess Risk Bounds for Group Slope Classifiers

Theorem 4

Consider a d_0-sparse multinomial logistic regression. Apply the multinomial logistic group Slope classifier (11) with λ_j 's satisfying

$$\max_{1 \leqslant j \leqslant d} \frac{\sqrt{L + \ln(d/j)}}{\lambda_j} \leqslant C_0 \sqrt{n} \tag{15}$$

with the constant C_0 derived in the proof. Assume Assumptions (A)-(C) and let $n \geqslant C_1 \ln d$. Then,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}\left(\widehat{\eta}_{gS}, \eta^*\right) \leqslant C(\delta) \left(\sum_{j=1}^{d_0} \frac{\lambda_j}{\sqrt{j}} \right)^{\frac{2(\alpha+1)}{\alpha+2}} \tag{12}$$

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

Two Specific Choices of Penalty Coefficients λ_j

Equal λ_j

This choice is equal to case of multinomial logistic group Lasso. Take

$$\lambda = C_0 \sqrt{\frac{L + \ln d}{n}} \tag{16}$$

to satisfy Theorem 4.

Corollary 1 With λ_j above and under Assumptions (A)-(C) and $n \geqslant C_1 \ln d$,

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

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

Variable λ_j

Consider

$$\lambda_j = C_0 \sqrt{\frac{L + \ln(d/j)}{n}} \tag{13}$$

Corollary 2 With λ_j in (13) and under Assumptions (A)-(C) and $n \geqslant C_1 \ln d$,

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

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

It is adaptively rate-optimal for both small and large number of classes, and, unlike the penalized likelihood classifier $\widehat{\eta}_{\widehat{M}}$, is **computationally feasible**.

Contribution of this Paper

The paper presents a unified, minimax-optimal theory for high-dimensional multiclass classification with sparse multinomial logistic regression:

1. **Penalized Maximum Likelihood Estimation**: The paper proposes the penalty to control the complexity in multinomial classification:

$$\mathrm{Pen}(|M|) = c_1|M|(L-1) + c_2|M| \ln\left(\frac{d\mathrm{e}}{|M|}\right) \tag{14}$$

2. **Minimax lower bounds**: The paper derived non-asymptotic upper bounds and minimax lower bounds for misclassification excess risk:

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) \leqslant C_1(\delta)\sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{d\mathrm{e}}{d_0}\right)}{n}} \tag{15}$$

And it will be proven that the upper bound is tight.

3. **Phase transition phenomenon**: The paper discovered the number of classes L will affect the dominant term of the risk. And it will be discussed in two circumstances: $L \leqslant 2 + \ln(d/d_0)$ and $L > 2 + \ln(d/d_0)$.
4. **Risk improvment under low-noise conditions**: A multiclass extension of the low-noise (Tsybakov) [2] condition is introduced. And under this condition, the risk bound is improved to

$$\mathcal{E}(\widehat{\eta}_{\widehat{M}}, \eta^*) \leqslant C(\delta) \left(\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n} \right) \tag{16}$$

5. **Computationally feasible methods**: Multinomial logistic group Lasso and Slope classifiers are designed to replace combinatorial search with convex optimization.

Limitations

1. The study primarily focuses on a **single form** of sparsity (row-wise sparsity in the parameter matrix), neglecting other sparsity patterns.
2. The theoretical results heavily depend on restrictive assumptions such as Assumptions (A)-(C). For instance, Assumption (A) requires class probabilities to be bounded away from 0 and 1.

Bibliography

[1] Felix Abramovich and Vadim Grinshtein. “High-dimensional classification by sparse logistic regression”. In: *IEEE Transactions on Information Theory* 65.5 (May 2019), pp. 3068–3079. DOI: 10.1109/tit.2018.2884963.

[2] Enno Mammen and Alexandre B. Tsybakov. *Smooth discrimination analysis*. Humboldt- Univ., Wirtschaftswiss. Fak, 1998.

[3] Canhong Wen et al. “Simultaneous dimension reduction and variable selection for multinomial logistic regression”. In: *INFORMS Journal on Computing* 35.5 (Sept. 2023), pp. 1044–1060. DOI: 10.1287/ijoc.2022.0132.