Minimax Lower Bound for Low-Rank Matrix-Variate Logistic Regression

Dec 14, 2021

Batoul Taki, Mohsen Ghassemi, Anand D. Sarwate, Waheed U. Bajwa

Department of Electrical and Computer Engineering
Rutgers University–New Brunswick

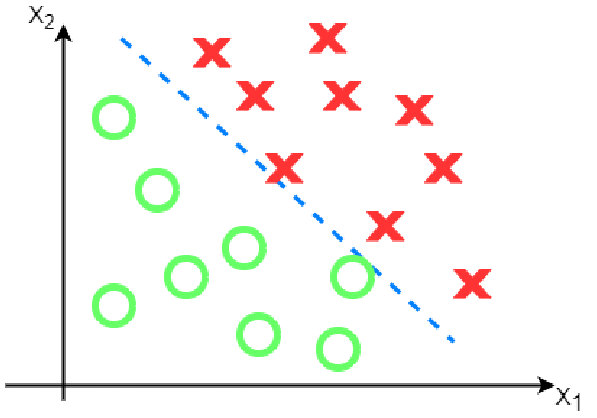
Outline

- Motivation and Model Overview
- Theoretical Result
- Construction of Our Theory

Part 1 (Motivation and Model Overview; Theoretical Result)

Part 2 (Construction of Our Theory)

Matrix-Variate Logistic Regression: Probability Model



Vector Logistic Regression:

$$\mathbb{P}_{y|\mathbf{x}}(y_i = 1|\mathbf{x}_i) = \frac{1}{1+\exp\left(-(\mathbf{b}^T\mathbf{x}_i + z)\right)}$$

$y_i \in \{0,1\}$: binary response (output class)

$\mathbf{b} \in \mathbb{R}^m$: unknown coefficient vector

z : zero-mean intercept (bias)

$\mathbf{x}_i \in \mathbb{R}^m$: covariate (data sample)

Matrix Logistic Regression:

$$\mathbb{P}_{y|\mathbf{X}}(y_i = 1|\mathbf{X}_i) = \frac{1}{1+\exp\left(-(\langle \mathbf{B}, \mathbf{X}_i\rangle + z)\right)}$$

$\mathbf{B} \in \mathbb{R}^{m_1 \times m_2}$

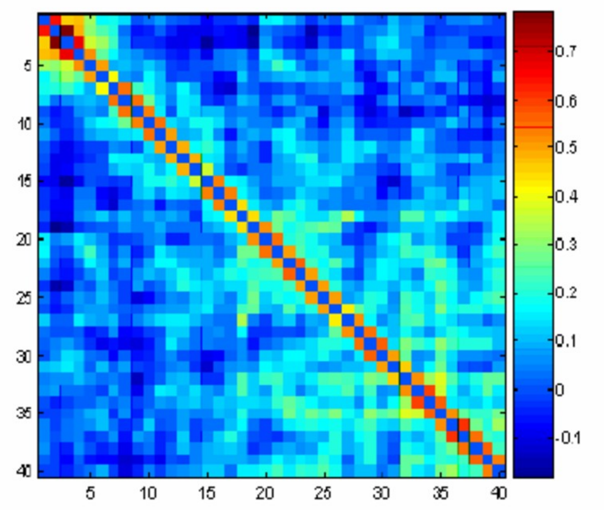
$\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$

Due to the inner product, both models are mathematically equivalent.

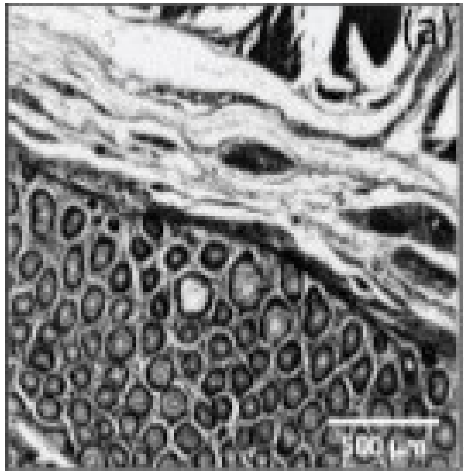
Why Matrix-Variate Logistic Regression?

$$\mathbb{P}_{y|\mathbf{X}}(y_i = 1|\mathbf{X}_i) = \frac{1}{1 + \exp\left(-(\langle \mathbf{B}, \mathbf{X}_i \rangle + z)\right)}$$

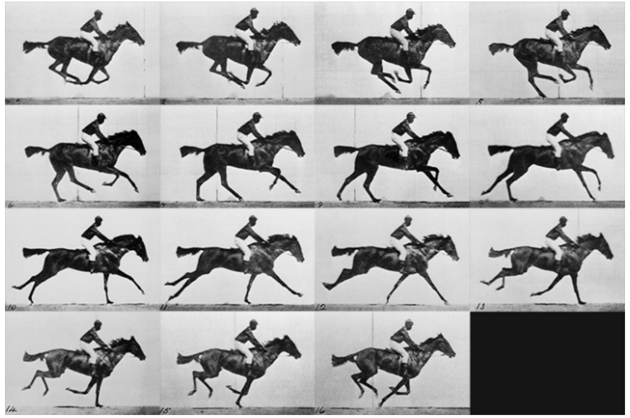
- In many practical applications covariates naturally take the form of two-dimensional arrays, such as:



Electroencephalography (EEG) data



Fiber-bundle Imaging



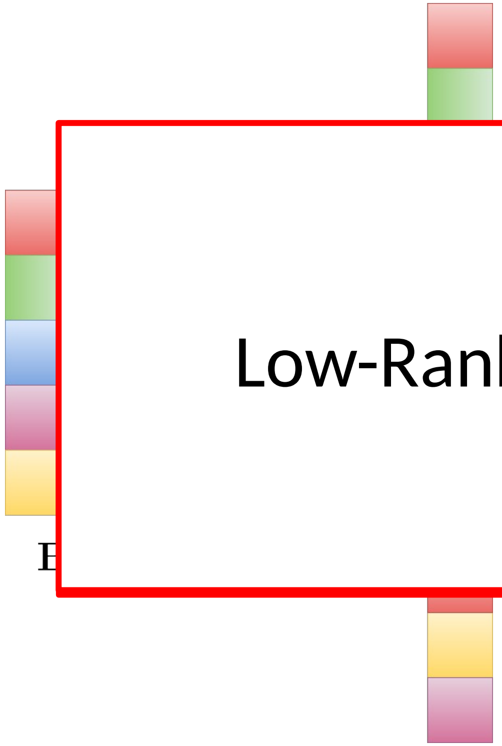
Spatial-temporal data

- The coefficients are also matrices, and contain rich information in their spatial structure.

Why Matrix-Variate Logistic Regression?

$$\mathbb{P}_{y|\mathbf{x}}(y_i = 1|\mathbf{X}_i) = \frac{1}{1+\exp\left(-(\langle \mathbf{B}, \mathbf{X}_i\rangle + z)\right)}$$

- For estimating B, classical machine learning techniques vectorize the data and estimate a coefficient vector.



Low-Rank Matrix-Variate Logistic Regression

$\mathbf{b} \in \mathbb{R}^{m_1 m_2}$

RUTGERS

Why Low-Rank Matrix-Variate Logistic Regression?

$$\mathbb{P}_{y|\mathbf{X}}(y_i = 1|\mathbf{X}_i) = \frac{1}{1 + \exp\left(-(\langle \mathbf{B}, \mathbf{X}_i \rangle + z)\right)}$$

- Low-rank structures may arise from the presence of redundant variables.
- The model's intrinsic degrees of freedom are smaller than its extrinsic dimensionality.

We can represent the data in a lower dimensional space

We can reduce the sample complexity of estimating the parameters

Prior Work:

- Vector based logistic regression
  - High-dimensional logistic regression [e.g F. Abramovich and V. Grinshtein 2018]
- Regularized matrix-variate logistic regression
  - Regularization for rank-optimized or sparse coefficient estimation [e.g J. Zhang and J. Jiang 2018, J. V. Shi et al 2014]
  - Regularization for inference on image data [e.g B. An and B. Zhang 2020]

Minimax Lower Bounds Provide Error Thresholds

Why Minimax Lower Bounds?

- They provide insights to:
  - The fundamental error thresholds of the estimation problem and the performance of corresponding algorithms.
- Indicate the parameters on which the minimax risk depends.

Prior Work

- Minimax lower bounds for graph-based logistic regression [e.g Q. Berthet and N. Baldin 2020].

RUTGERS

Outline of This Work

- Derive a minimax lower bound that is proportional to the rank and dimensions of the coefficient matrix.
- Reduce the sample complexity from the vector setting.
- Show that the methods used are easily extendible to the tensor case.

Model and Problem Formulation

$$\mathbb{P}_{y|\mathbf{x}}(y_i = 1|\mathbf{X}_i) = \frac{1}{1+\exp\left(-(\langle \mathbf{B}, \mathbf{X}_i\rangle + z)\right)}$$

- Consider the matrix LR problem:
  - Goal: Find estimate $\widehat{\mathrm{B}}$ of B using training data $\{\mathbf{X}_i, y_i\}_{i=1}^n$.
- Consider the case where B is a rank-r matrix. Specifically, the rank-r singular value decomposition of B is

$$\mathbf{B} = \mathbf{B}_1 \mathbf{G} \mathbf{B}_2^T \qquad \begin{bmatrix} | & & | \\ \mathbf{b}_1^1 & \cdots & \mathbf{b}_1^r \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \begin{bmatrix} — & \mathbf{b}_2^1 & — \\ & \vdots & \\ — & \mathbf{b}_2^r & — \end{bmatrix}$$

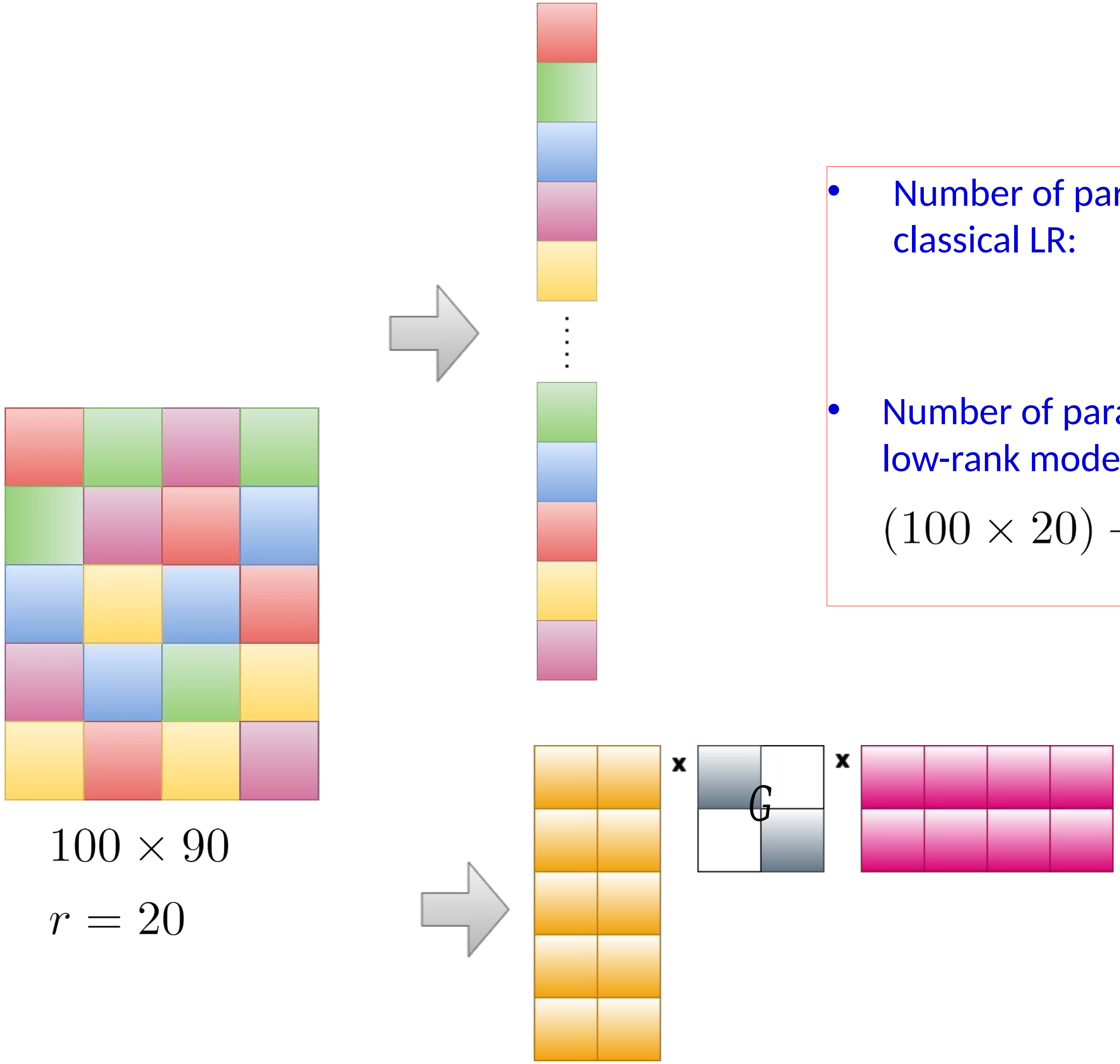
$\mathbf{B}_1 \in \mathbb{R}^{m_1 \times r}$, $\mathbf{B}_2 \in \mathbb{R}^{m_2 \times r}$ } Matrix of left/right singular vectors (with orthonormal columns)

$\mathbf{G} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r \times r}, \ \lambda_1 > 0 \ \forall i \in [r]$ } Matrix of singular values

$$\mathbb{P}_{y|\mathbf{x}}(y_i = 1|\mathbf{x}_i) = \frac{1}{1+\exp\left(-(\langle \mathbf{B}_1 \mathbf{G} \mathbf{B}_2^T, \mathbf{X}_i\rangle + z)\right)}$$

RUTGERS

Benefit of Low-Rank Models (Toy Example)



100×90

$r = 20$

- Number of parameters to learn in the case of classical LR:

$$9000$$

- Number of parameters to learn in the case of low-rank model

$$(100 \times 20) + (90 \times 20) + 20 = 3820$$

Model and Problem Formulation, Continued

Consider the parameter space, $\mathcal{P}_r$, of all rank-r matrices in $\mathbb{R}^{m_1 \times m_2}$, and a subset $\mathcal{B}_d \subset \mathcal{P}_r$ of rank-r matrices with finite energy. More formally.

$$\mathcal{B}_d(\mathbf{0}) \triangleq \{\mathbf{B}' \in \mathcal{P}_r : \|\mathbf{B}' - \mathbf{0}\|_F < d\}$$

The minimax risk is thus defined as the worst-case mean squared error (MSE) for the best estimator, i.e.,

$$\varepsilon^* = \inf_{\widehat{\mathbf{B}}} \sup_{\mathbf{B} \in \mathcal{B}_d(\mathbf{0})} \mathbb{E}_{\mathbf{y}, \underline{X}^c} \left\{ \|\widehat{\mathbf{B}} - \mathbf{B}\|_F^2 \right\}$$

RUTGERS

Main Results of Prior Works

Vector-based Logistic Regression: $\mathcal{O} = \frac{m_1 m_2}{n}$

Matrix Logistic Regression: ??

Main Result

Theorem 1 [Taki et al. 2021]

Consider the rank-r matrix LR problem with n i.i.d observations, $\{\mathbf{X}_i, y_i\}_{i=1}^{n}$ where the true coefficient matrix $\|\mathbf{B}\|_F^2 < d^2$.

Then, for covariate $\text{vec}(\mathbf{X}_i) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_1 m_2})$ the minimax risk is lower bounded by

$$\varepsilon^* \geq \frac{\left(\left[c_2\left(c_1 \mathbf{r}(\mathbf{m_1} + \mathbf{m_2} - \mathbf{2}) + c_1(\mathbf{r} - \mathbf{1})\right) - c_3\right] - 1\right)}{8\mathbf{n}\sigma\sqrt{\frac{2}{\pi}}}$$

where

$$c_1 = \left(1 - \frac{1}{10}\right)^2, \quad c_2 = \frac{\log_2(e)(\sqrt{2} - 1)}{4\sqrt{2}}, \quad c_3 = \left(\frac{3(\sqrt{2} - 1)}{\sqrt{8}}\right) \log_2\left(\frac{3}{2}\right)$$

Main Result and Discussion

$$\varepsilon^* \geq \frac{\left(\left[c_2\left(c_1\mathbf{r}(\mathbf{m_1}+\mathbf{m_2}-\mathbf{2})+c_1(\mathbf{r}-\mathbf{1})\right)-c_3\right]-1\right)}{8\mathbf{n}\sigma\sqrt{\frac{2}{\pi}}}$$

- Compared to the vector case, result shows a decrease in the lower bound.

Minimax risk for vector based LR: $\mathcal{O}\left(\frac{m_1 m_2}{n}\right)$

Minimax risk for rank-r matrix LR: $\mathcal{O}\left(\frac{r(m_1+m_2+1)}{n}\right)$

- Lower bound on the minimax risk is proportional to the intrinsic degrees of freedom in the coefficient matrix LR.

RUTGERS

The Exciting Part! Proof of Main Results

Proof of Theorem 1 uses an argument based on Fano's inequality, more specifically:

Suppose there exists an estimator, $\widehat{B}$, with worst-case MSE that matches ε^*.

→

This estimator can be used to solve a multiple hypothesis testing problem. The hypothesis test is an exponentially large family of distinct matrix coefficients:
$\mathcal{B}_L = \{\mathbf{B} : l \in [L]\} \subset \mathcal{B}_d(\mathbf{0})$

→

Minimax risk can be lower bounded by the probability of error in the hypothesis test.

Our goal: Further lower bound the probability of error.

Action items:

- Construct $\mathcal{B}_L$
- Find upper and lower bounds on the conditional mutual information $\mathbb{I}(\mathbf{y}; l|\underline{\mathbf{X}}^c)$

The Exciting Part! Proof of Main Results

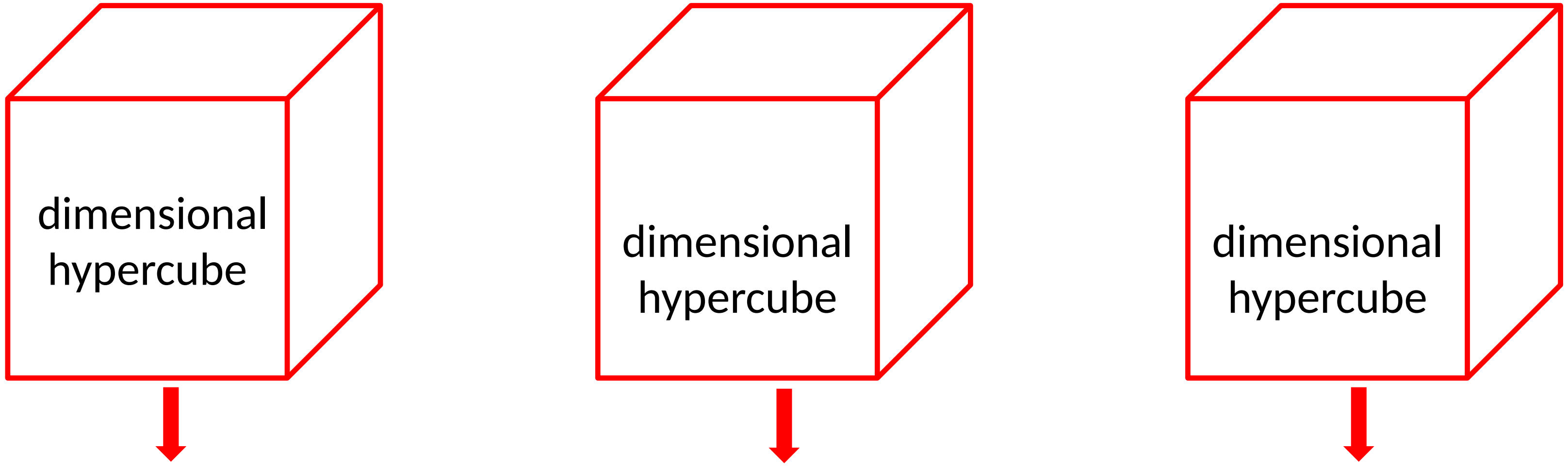
1. Constructing $\mathcal{B}_L$

a) We must **construct** $\mathcal{B}_L$ such that a **minimum distance condition holds**, namely:

$$\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2 \geq 8\delta$$

b) Since $\mathbf{B}_l = \mathbf{B}_1 \mathbf{G} \mathbf{B}_2^T$, we must **construct three separate sets** and **derive conditions** under which they exists simultaneously

Hypercube method: Construct a set of binary vectors/matrices with a minimum distance between any two distinct elements



$\mathbf{G}_f | f \in [F]$

- Square diagonal matrix

$\mathbf{B}_{1,p_1} | p_1 \in [P_1]$

- Orthonormal Columns
- Bounded energy

$\mathbf{B}_{1,p_2} | p_2 \in [P_2]$

- Orthonormal Columns
- Bounded energy

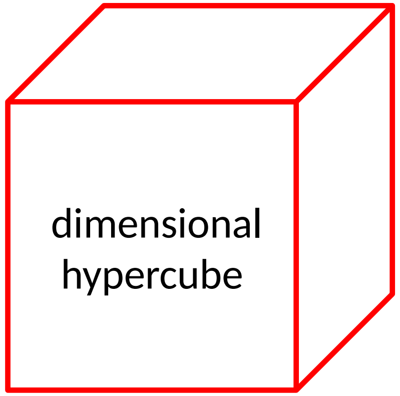
The Exciting Part! Proof of Main Results

Lemma 1: "Each hypercube exists with probability"

Lemma 1

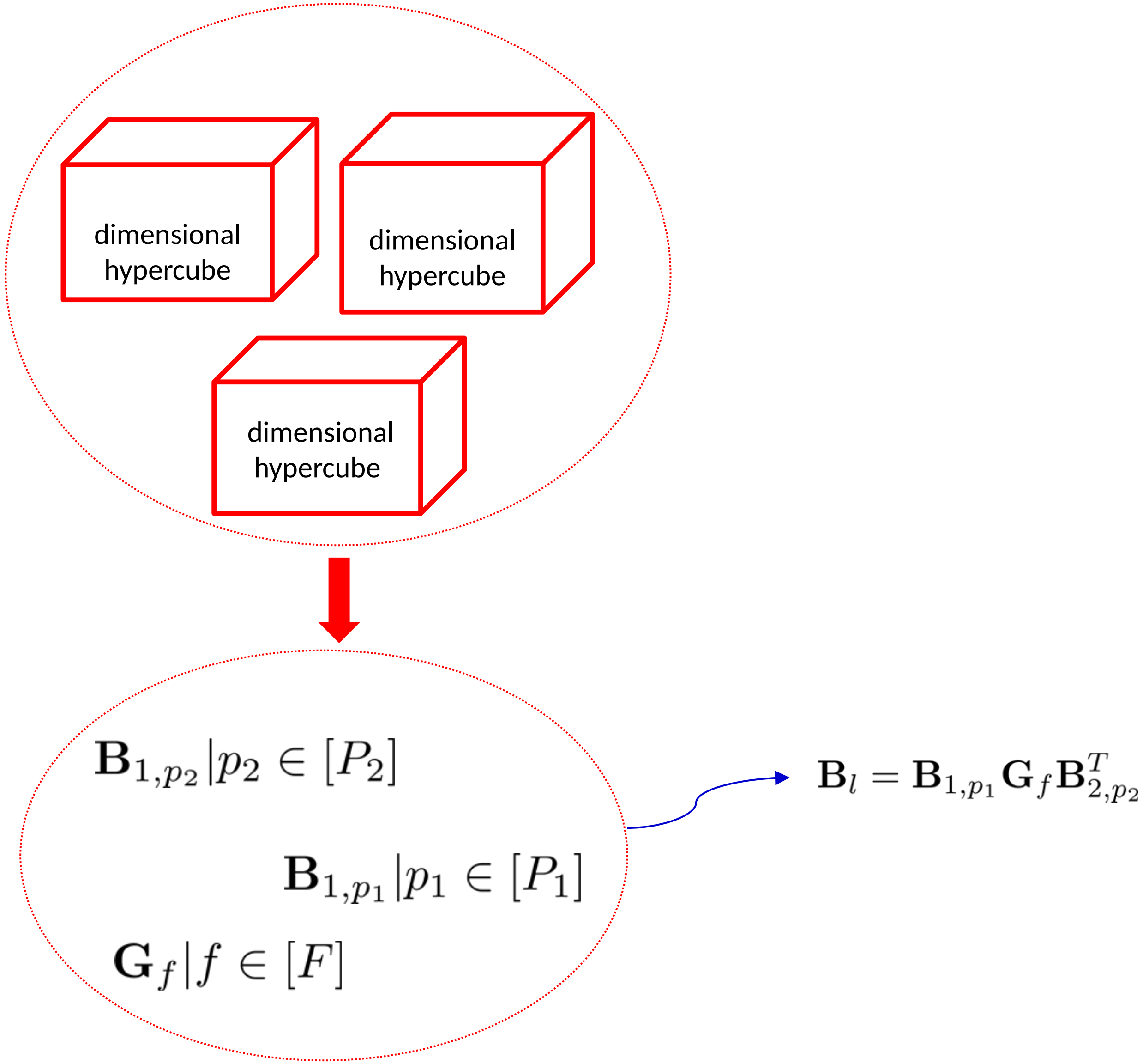
Let $r > 0$ and $F \geq 2$. Consider the set of F vectors $\{\mathbf{s}_f \in \mathbb{R}^{r-1} : f \in [F]\}$, where each entry in vector $\mathbf{s}_f$ is an independent and identically distributed random variable taking values $\left\{-\frac{1}{\sqrt{r-1}}, +\frac{1}{\sqrt{r-1}}\right\}$ uniformly. The probability that there exists a distinct pair (f, f') such that $\|\mathbf{s}_f - \mathbf{s}_{f'}\|_0 < \frac{r-1}{20}$ is upper bounded as follows:

$$\mathbb{P}(\exists (f, f') \in [F] \times [F], f \neq f' : \|\mathbf{s}_f - \mathbf{s}_{f'}\|_0 < \frac{r-1}{20}) \leq \exp\left[2\log(F) - \log(2) - \frac{1}{2}\left(1 - \frac{1}{10}\right)^2 (r-1)\right]. \quad (1)$$



RUTGERS

The Exciting Part! Proof of Main Results



RUTGERS

The Exciting Part! Proof of Main Results

Lemma 2: "For all sets to exists simultaneously, we can construct set $\mathcal{B}_L$ with L elements, where the distance between any two elements is bounded"

Lemma 2

There exists a collection of L matrices $B_L \triangleq \{\mathbf{B}_l : l \in [L]\} \subset \mathcal{B}_d(\mathbf{0})$ for some $d > 0$ of cardinality

$$L = 2^{\lfloor \frac{\log_2(e)}{4}\left(\left(1-\frac{1}{10}\right)^2(r(m_1+m_2-1)+\left(1-\frac{1}{10}\right)^2(r-1)\right)-\frac{3}{2}\log_2\left(\frac{3}{2}\right)\rfloor} \tag{1}$$

such that for any

$$\sqrt{\frac{8(r-1)}{r}} < \varepsilon \leq d\sqrt{\frac{r-1}{r}}, \tag{2}$$

we have

$$\frac{r\varepsilon^2}{r-1} < \|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2 \leq 4\frac{r\varepsilon^2}{r-1}. \tag{3}$$

Our packing:

RUTGERS

The Exciting Part! Proof of Main Results

1. Bounding $\mathbb{I}(\mathbf{y}; l|\underline{\mathbf{X}}^c)$

   a) Lower bound using Fano's inequality

   – We require the existence of an estimator producing estimate $\widehat{\mathbf{B}}$ and achieving minimax lower bound $\varepsilon^* = \sqrt{\delta}$

   – Consider the minimum distance decoder: $\widehat{l}(\mathbf{y}) \triangleq \underset{\mathbf{B}_{l'} \in \mathcal{B}_d(\mathbf{0})}{\arg\min} \left\| \widehat{\mathbf{B}} - \mathbf{B}_{l'} \right\|_F^2$

$\|\widehat{\mathbf{B}} - \mathbf{B}_l\|_F^2 < \sqrt{2\delta}$: **detect Bl and** $\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l) = 0$

$\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2 \geq \sqrt{2\delta}$:**detection error might occur**

$$\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l) \leq \mathbb{P}\left(\left\|\hat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2 \geq \sqrt{2\delta} \right)$$

Fano's inequality states that: $\mathbb{I}(\mathbf{y}; l) \geq \left(1 - \mathbb{P}(\widehat{l}(\mathbf{y}) \neq l)\right) \log_2(L) - 1 \triangleq u_1$

RUTGERS

The Exciting Part! Proof of Main Results

1. Bounding $\mathbb{I}(\mathbf{y}; l|\underline{\mathbf{X}}^c)$

   b) Upper bound using

$$\mathbb{I}(\mathbf{y}; l|\underline{\mathbf{X}}^c) \leq \frac{1}{L^2}\sum_{l,l'} \mathbb{E}_{\underline{\mathbf{X}}^c} D_{KL}(f_l(\mathbf{y}|\mathbf{X})||f_{l'}(\mathbf{y}|\mathbf{X})) \triangleq u_2$$

Lemmas 3 and 4 provide upper and lower bounds:

$$\frac{\sqrt{2}-1}{\sqrt{2}}\log_2 L - 1 \leq \mathbb{I}(\mathbf{y}; l|\mathbf{X}) \leq n\sigma\frac{2}{r}\sqrt{\frac{2}{\pi}}\varepsilon.$$

RUTGERS

Some Closing Remarks

The result is interesting because:

- The analysis is non-trivial because the model uses a logistic function. Moreover, the result explicitly leverages the low-rank structure thus the hypothesis set is constructed from three factor sets. We derive conditions under which all sets can exists, and can be generalized to the tensor case.
- Two **hypotheses may be far apart** but produce the **same model** (or same observation). Our result gives insight into the parameters in which an achievable minimax risk might depend.

Current Investigations and Future Work

Study the benefits of imposing similar low-rank structures in the multi-dimensional LR setting:

Minimax risk lower bounds on the coefficient estimation in tensor-variate logistic regression.

Develop algorithms that meet the minimax lower bounds.

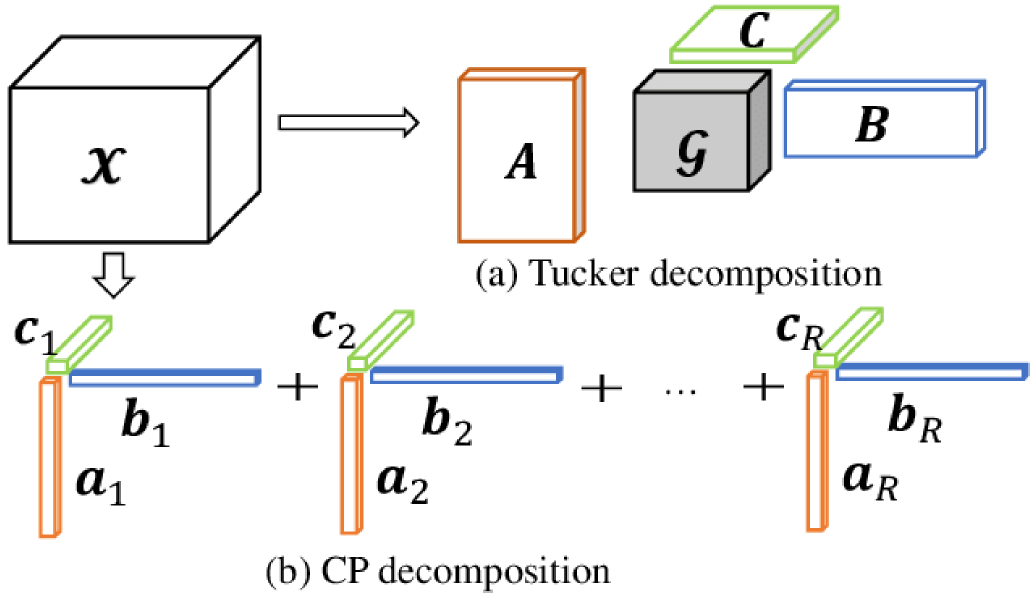
Test the performance of these algorithms on practical data.

Current Investigations and Future Work

Study the benefits of imposing similar low-rank structures in the multi-dimensional LR setting:

Minimax risk lower bounds on the coefficient estimation in tensor-variate logistic regression.

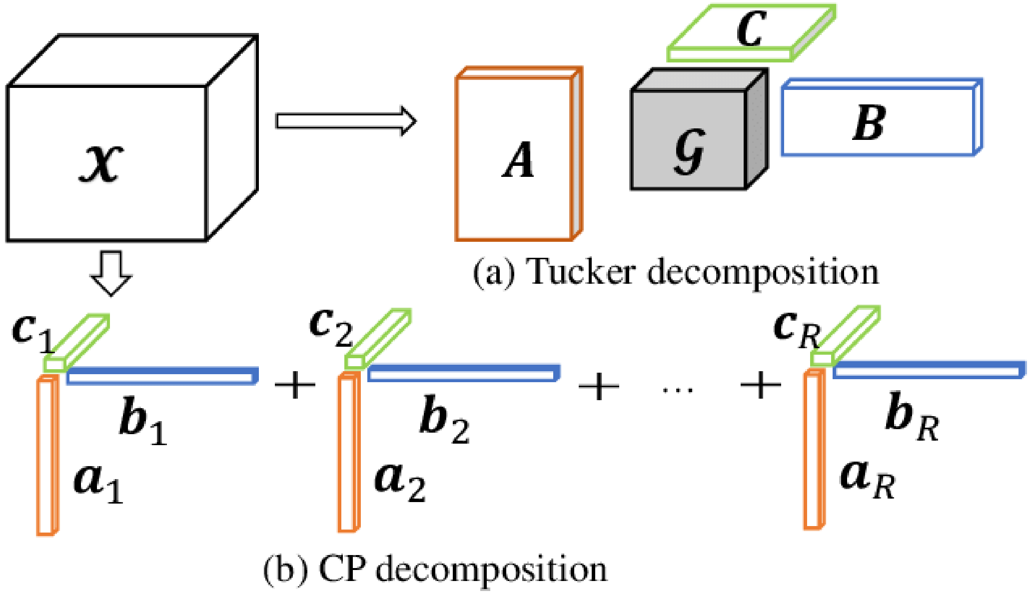
- CANDECOMP/PARAFAC (CP).
- Low-rank Tucker.



Current Investigations and Future Work

Study the benefits of imposing similar low-rank structures in the multi-dimensional LR setting:

Develop algorithms that meet the minimax lower bounds.



Current Investigations and Future Work

Study the benefits of imposing similar low-rank structures in the multi-dimensional LR setting:

Test the performance of these algorithms on practical data.