Asymptotic Performance of Linear Discriminant Analysis with Random Projections

Khalil Elkhalil, Abla Kammoun*, Robert Calderbank†, Tareq Y. Al-Naffouri* and Mohamed-Slim Alouini**

* King Abdullah University of Science and Technology (KAUST)
† Duke University

Introduction

Motivation

- The design of LDA heavily relies on the data covariance matrix which becomes ill conditioned in the large data regime.
- Most analysis focus on regularization techniques to overcome the high dimensionality effect on the estimation of the covariance matrix.
- Dimensionality reduction is an effective technique to get around high dimensionality but most analysis relies on bounds on the performance which might be loose in certain cases.
- Random projection is a common way to perform dimensionality reduction with some guarantees on the pairwise distances between data points (the Johnsonn-Lindenstrauss Lemma) but little can be told regarding the classification performance.

Contributions

- We consider LDA when data is randomly projected and arise from the multivariate distribution.
- We investigate the classification performance for general random projection matrices satisfying some finite moments assumptions.
- We carry out the analysis when both the data dimension p and the reduced dimension d grow large simultaneously at the same rate, i.e. $d/p \to c \in (0,1)$.
- Under some mild assumptions controlling the data statistics, we show that the classification risk converges to a universal limit that describes in closed form fashion the performance in terms of the statistics and the dimensions involved.
- The obtained results permits to analytically quantify the performance loss due to projection which allows to carefully choose the reduced dimension in order to achieve a certain desirable performance.

LDA with Random Projections

LDA

- For a data point $\boldsymbol{x} \in \mathbb{R}^p$, we say that $\boldsymbol{x} \in \mathcal{C}_i$ iff $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{C})$.
- When the data is Gaussian, LDA is a Bayes classifier in the sense it maximizes $\mathbb{P}[\mathcal{C}_i|\boldsymbol{x}]$ for $i \in \{0,1\}$. The LDA score is

$$W_{\text{LDA}}(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^\top \mathbf{C}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) + \log\frac{\pi_0}{\pi_1} \underset{\mathcal{C}_1}{\overset{\mathcal{C}_0}{\gtrless}} 0. \qquad (1)$$

- The conditional probability of misclassification is given by $\epsilon_i^{\text{LDA}} = \mathbb{P}\left[(-1)^i W_{\text{LDA}} < 0 | \boldsymbol{x} \in \mathcal{C}_i\right]$.
- Relying on the Gaussian assumption, we have

$$\epsilon_i^{\text{LDA}} = \Phi\left[\frac{-\frac{1}{2}\boldsymbol{\mu}^\top \mathbf{C}^{-1}\boldsymbol{\mu} + (-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top \mathbf{C}^{-1}\boldsymbol{\mu}}}\right]. \qquad (2)$$

Random Projections

Random projection consists in the following operation.

$$\mathbb{R}^p \to \mathbb{R}^d$$
$$\boldsymbol{x} \longmapsto \mathbf{W}\boldsymbol{x}.$$

Johnsonn-Lindenstrauss Lemma

For a given n data points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ in $\mathbb{R}^p$, $\epsilon \in (0,1)$ and $d > \frac{8\log n}{\epsilon^2}$, there exists a linear map $f: \mathbb{R}^p \to \mathbb{R}^d$ such that

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \leq \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|^2 \leq (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2, \qquad (3)$$

for all $i, j \in [n]$

What about the classification risk ?
Conditioning on the projection matrix $\mathbf{W}$, we have

$$\epsilon_i^{\text{P-LDA}} = \Phi\left[-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top \mathbf{W}^\top (\mathbf{W}\mathbf{C}\mathbf{W}^\top)^{-1}\mathbf{W}\boldsymbol{\mu}} + \frac{(-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top \mathbf{W}^\top (\mathbf{W}\mathbf{C}\mathbf{W}^\top)^{-1}\mathbf{W}\boldsymbol{\mu}}}\right]. \qquad (4)$$

Technical Assumptions

Assumption 1. (Growth rate)

As $p, d \to \infty$ we assume the following

- **Data scaling**: $0 < \liminf \frac{d}{p} \leq \limsup \frac{d}{p} \leq 1$,
- **Mean scaling**: Let $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, $\limsup_p \|\boldsymbol{\mu}\| < \infty$.
- **Covariance scaling**: $\limsup_p \|\mathbf{C}\| < \infty$.

Assumption 2. (Projection matrix)

We shall assume that the projection matrix $\mathbf{W}$ writes as $\mathbf{W} = \frac{1}{\sqrt{p}}\mathbf{Z}$, where the entries $Z_{i,j}$ ($1 \leq i \leq d$, $1 \leq j \leq p$) of $\mathbf{Z}$ are centered with unit variance and independent identically distributed random variables satisfying the following moment assumption.
There exists $\epsilon > 0$, such that $\mathbb{E}|Z_{i,j}|^{4+\epsilon} < \infty$.

Main Results

A Fundamental Result in RMT

under Assumptions 1 and 2, it allows to construct a deterministic equivalent of $\left(\frac{1}{tp}\mathbf{C}^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}^{1/2} + \mathbf{I}_p\right)^{-1}$ denoted by $\mathbf{Q}(t) \in \mathbb{R}^{p\times p}$ in the sense that

$$\boldsymbol{a}^\top\left(\frac{1}{tp}\mathbf{C}^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}^{1/2} + \mathbf{I}_p\right)^{-1}\boldsymbol{b} - \boldsymbol{a}^\top\mathbf{Q}(t)\boldsymbol{b} \to_{prob.} 0,$$

for all deterministic $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^p$ with uniformly bounded Euclidean norms and $t > 0$. $\mathbf{Q}(t)$ is a deterministic matrix given by $\mathbf{Q}(t) = \left(\mathbf{I}_p + \frac{\frac{d}{tp}}{1+\frac{d}{tp}\delta(t)}\mathbf{C}\right)^{-1}$, where $\delta(t)$ satisfies $\delta(t) = \frac{1}{d}\operatorname{tr}\mathbf{C}\mathbf{Q}(t)$.

Proposition 1. (Asymptotic Performance)

Under Assumptions 1 and 2, then for $i \in \{0,1\}$ the conditional probability of misclassification in (4) converges in probability to a non trivial deterministic limit given by

$$\epsilon_i^{\text{P-LDA}} - \Phi\left[\frac{-\frac{1}{2}\boldsymbol{\mu}^\top(\mathbf{C} + \delta_d\mathbf{I}_p)^{-1}\boldsymbol{\mu} + (-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top(\mathbf{C} + \delta_d\mathbf{I}_p)^{-1}\boldsymbol{\mu}}}\right] \to_{prob.} 0, \qquad (5)$$

where δ_d is such that

$$\delta_d \operatorname{tr}(\mathbf{C} + \delta_d\mathbf{I}_p)^{-1} = p - d. \qquad (6)$$

Special cases

- Equal priors, i.e. $\pi_0 = \pi_1$.

$$\epsilon^{\text{P-LDA}} - \Phi\left[-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top(\mathbf{C} + \delta_d\mathbf{I}_p)^{-1}\boldsymbol{\mu}}\right] \to_{prob.} 0.$$

- Equal priors and $\mathbf{C} = \mathbf{I}_p$.

$$\epsilon^{\text{P-LDA}} - \Phi\left[-\frac{1}{2}\sqrt{d/p}\,\|\boldsymbol{\mu}\|\right] \to_{prob.} 0.$$

As expected, there is a performance loss due to projection and it is analytically characterized by Proposition 1. Conversely, for a given desired performance $\bar{\epsilon}$, we can determine the minimum d such that $\epsilon^{\text{P-LDA}} \leq \bar{\epsilon}$.

Experiments

We consider Gaussian and Bernoulli projection matrices generated as follows.

- Gaussian: $W_{i,j} \sim_{\text{i.i.d}} \mathcal{N}(0, 1/p)$.
- Bernoulli: $W_{i,j} = \left\{\frac{1}{\sqrt{p}}(1 - 2B_{i,j})\right\}$ where $B_{i,j} \sim_{\text{i.i.d}}$ **Bernoulli** $(1/2)$.

Synthetic data

The data is generated using the Gaussian distribution with the following parameters.

- $p = 800$.
- $\boldsymbol{\mu}_0 = \mathbf{0}_p$ and $\boldsymbol{\mu}_1 = \frac{3}{\sqrt{p}}\mathbf{1}_p$.
- $\mathbf{C} = \{0.4^{|i-j|}\}_{i,j}$.

MNIST data

- $\mathcal{C}_0$ is taken to be the digit 2 whereas $\mathcal{C}_1$ is given by digit 3.
- We obtain the data statistics by relying on sample estimates computed from the training data.

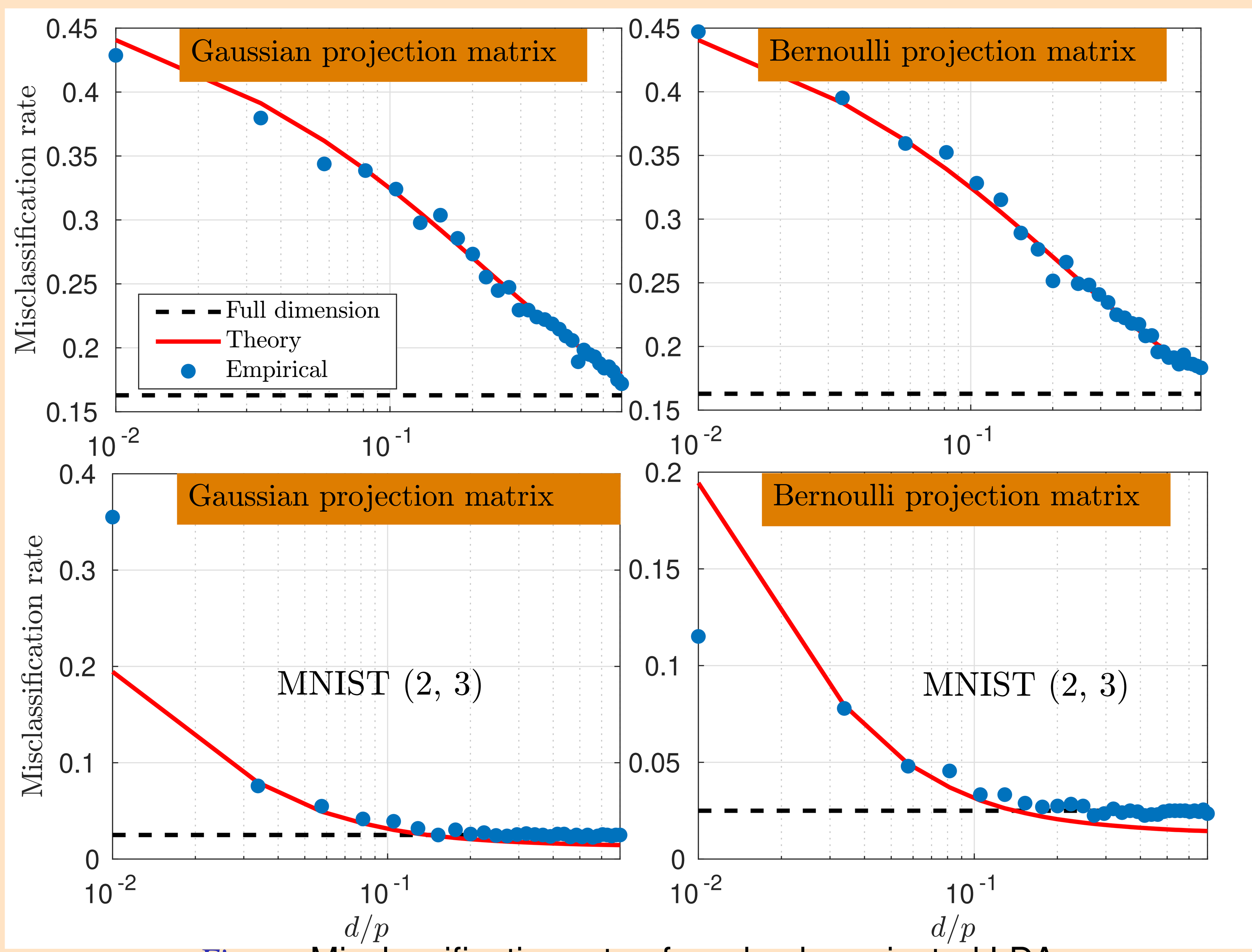

Figure: Misclassification rate of randomly-projected LDA.