# ASYMPTOTIC PERFORMANCE OF LINEAR DISCRIMINANT ANALYSIS WITH RANDOM PROJECTIONS

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

[*] CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia
[†] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27707

## ABSTRACT

We investigate random projections in the context of randomly projected linear discriminant analysis (LDA). We consider the case in which the data of dimension $p$ is randomly projected onto a lower dimensional space before being fed to the classifier. Using fundamental results from random matrix theory and relying on some mild assumptions, we show that the asymptotic performance in terms of probability of misclassification approaches a deterministic quantity that only depends on the data statistics and the dimensions involved. Such results permits to reliably predict the performance of projected LDA as a function of the reduced dimension $d < p$ and thus helps to determine the minimum $d$ to achieve a certain desired performance. Finally, we validate our results with finite-sample settings drawn from both synthetic data and the popular MNIST dataset.

## 1. INTRODUCTION

Linear discriminant analysis or in short LDA is a popular supervised classification technique that dates back to Fisher [1] where the basic idea was to find the classifier that maximizes the ratio between inter-class distance and intra-class variance. When the data arise from the Gaussian distribution and the data statistics are perfectly known LDA is known to be the Bayes classifier [2, 3, 4]. Modern machine learning however involve data in high dimensional spaces which makes conventional classification techniques generally inefficient resulting in what is called *the curse of dimensionality* [5]. Dimensionality reduction is one of the promising solutions as it permits to downscale the data dimension assuming that only a subset of features are relevant to classification. A popular example of such dimensionality reduction techniques is principal component analysis (PCA) which projects the data onto the subspace spanned by the eigenvectors of the covariance matrix relative to the highest eigenvalues. However, PCA is more suitable for data reconstruction than for classification since the principal components are chosen to maximize the data variance and may be not necessarily the best discriminative directions from a classification point of view.

An alternative way to perform dimensionality reduction consists in using random projections that randomly project the data onto a lower dimensional space [6, 7, 8]. The classification is then performed on the projected data, resulting in a substantial computational savings. From a performance point view, random projections have shown good generalization performance as discussed in the analyis of [6] through Vapnik–Chervonenkis type bounds on the generalization error of linear classifiers with Gaussian projections. Other works obtained some performance guarantees for randomly-projected classifiers under some assumptions on the data structure such as sparsity [9] or separability [10].

In this paper, we consider randomly-projected LDA when data is assumed to arise from the multivariate Gaussian distribution. We investigate the performance for general random projection matrices satisfying some finite moment assumptions. The analysis is carried out when both the data dimension $p$ and the reduced dimension $d$ tends to infinity while their ratio is fixed, i.e., $d/p \to$ constant $\in (0, 1)$ . Based on fundamental results from random matrix theory and on some assumptions controlling the data statistics and the projection matrix, we show that the classification risk converges to a universal limit that describes in a closed form fashion the performance in terms of the statistics and the dimensions involved. The result permits to examine the fundamental limits of projected LDA under known statistics. In the simulation results, we show that this assumption is not limiting since accurate predictions can be made for real data in which the data statistics are not known.

The remainder of this paper is organized as follows. In section 2, we give a brief overview of projected-LDA. In section 3, we provide our main theoretical results and conclude the paper in section 4 by making some conclusions and investigating some possible future research directions.

## 2. LINEAR DISCRIMINANT ANALYSIS WITH RANDOM PROJECTIONS

### 2.1. Linear discriminant analysis

We consider binary classification of data points arising from the multivariate Gaussian distribution. For a datum $\boldsymbol{x} \in \mathbb{R}^p$, we say $\boldsymbol{x} \in \mathcal{C}_i$, $i \in \{0, 1\}$, if and only if $\boldsymbol{x}$ writes as

$$\boldsymbol{x} = \boldsymbol{\mu}_i + \mathbf{C}^{1/2}\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, $\boldsymbol{\mu}_i \in \mathbb{R}^p$ are respectively the class means and $\mathbf{C} \in \mathbb{R}^{p \times p}$ is a symmetric non-negative matrix representing

the data covariance matrix common to both classes. We denote by $\pi_i \in [0,1]$, the prior probability of $\boldsymbol{x}$ to belong to class $\mathcal{C}_i$, for $i \in \{0,1\}$. Since both classes have the same covariance matrix, it is well known in this setting that linear discriminant analysis (LDA) is the classifier that maximizes the posterior probability $\mathbb{P}[\mathcal{C}_i|\boldsymbol{x}]$ among all classifiers. In that sense, LDA is the Bayes rule classifier which uses the following score as classification metric [2, 3]

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

Then, the classification rule is given by

$$\begin{cases} \boldsymbol{x} \in \mathcal{C}_0 & \text{if } W_{\text{LDA}}(\boldsymbol{x}) > 0 \\ \boldsymbol{x} \in \mathcal{C}_1 & \text{otherwise.} \end{cases} \tag{3}$$

One important metric to evaluate the performance of a given classifier is the probability of misclassification that we denote by $\epsilon$. To compute this probability, we need to compute the conditional probability of misclassification which is related to the error the classifier makes on data sampled from a certain class. Formally speaking, the conditional probability of misclassification for LDA is given by

$$\epsilon_i^{\text{LDA}} = \mathbb{P}\left[(-1)^i W_{\text{LDA}}(\boldsymbol{x}) < 0 | \boldsymbol{x} \in \mathcal{C}_i\right],\ i \in \{0,1\}. \tag{4}$$

Then, the total probability of misclassification is evaluated as follows

$$\epsilon^{\text{LDA}} = \pi_0 \epsilon_0^{\text{LDA}} + \pi_1 \epsilon_1^{\text{LDA}}. \tag{5}$$

Again, relying on the Gaussian assumption of the data, $\epsilon_i^{\text{LDA}}$ can be evaluated in closed form as [11, 12]

$$\epsilon_i^{\text{LDA}} = \Phi\left[\frac{-\frac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{C}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) + (-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{C}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}}\right] \tag{6}$$

Then, the total probability of misclassification reduces to

$$\epsilon^{\text{LDA}} = \Phi\left[-\frac{1}{2}\sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{C}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}\right],$$

in the case of equal priors.

### 2.2. Random projections

As mentioned in the introduction, random projection is a popular technique for dimensionality reduction, which merely consists in projecting at random the feature vectors onto a lower-dimensional space. The use of such random projections is originally motivated by the Johnsonn–Lindenstrauss Lemma with an abundant literature (see [13, 14, 8] and references therein) that states that 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

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \leq \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|^2 \leq (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2, \tag{7}$$

for all $i, j \in [n]$. For $d$ larger enough than $\log n$, it is possible to still preserve pairwise distances after randomly projecting the data onto a $d-$dimensional space. Such an observation is the main driver behind the projected LDA classifier and tends to suggest that the projected LDA classifier, consisting in projecting data onto a $d$-dimensional space prior to applying the LDA classifier would probably present comparable performance with the classical LDA, while allowing substantial computational savings, [15]. However, to satisfactorily address this question, it is important to carry out a complete analysis that investigates the probability of misclassification of the projected LDA classifier. In the sequel, we denote by $\mathbf{W} \in \mathbb{R}^{d\times p}$ with $d < p$, a random linear map with $i.i.d$ entries having zero mean and variance $1/p$. If we let $p$ to grow large enough, then $\mathbf{W}\mathbf{W}^\top \approx \mathbf{I}_d$. This means that asymptotically $\mathbf{W}$ behaves as a projection matrix. In the following, we conduct a large dimensional analysis of projected-LDA where we show under some mild assumptions that the probability of misclassification converges in probability to a deterministic quantity from which one can investigate the performance loss due to dimensionality reduction.

## 3. MAIN RESULTS

### 3.1. Technical assumptions

In this section, we present our theoretical results on the asymptotic performance of projected-LDA under the growth regime in which the data dimension $p$ and the reduced dimension $d$ grow large with $d < n$. We show that in this case the probability of misclassification converges to a deterministic quantity that describes in a closed form fashion the performance in terms of the problem's statistics and dimensions. More formally, the following assumptions are considered.

**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***: $\limsup_p \|\mathbf{C}\| < \infty$.

These assumptions are standard in the context of random matrix theory and are go in line with the increase in data dimensions met in the big data paradigm. The assumption on the Euclidean distance between the means is technically required since it allows to obtain non trivial results (perfect misclassification or a classification performance that one would get by simply relying on the priors for example) on the performance and thus permits to obtain non trivial conclusions. The same argument applies for the covariance matrix

scaling since a covariance matrix with infinite spectral norm would lead to a poor performance. Moreover, as we will show later, the assumption on the covariance matrix $\mathbf{C}$ is technically important as it allows to use fundamental results on resolvent random matrices [16, 17].

**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 [16]. There exists* $\epsilon > 0$*, such that* $\mathbb{E}\left|Z_{i,j}\right|^{4+\epsilon} < \infty$.

As can be inferred from Assumption 2, the only assumption we require for the projection matrix is that it possesses entries with finite fourth moments without any further restrictions on their distribution.

### 3.2. Asymptotic performance

We would like to investigate the performance of LDA when the data are randomly projected with $\mathbf{W}$. Before doing so, we recall that the random projection procedure is simply given by

$$\boldsymbol{x} \to \mathbf{W}\boldsymbol{x},\ \boldsymbol{x} \in \mathbb{R}^p, \tag{8}$$

where $\mathbf{W}$ satisfies Assumption 2. From (2), it is easy to derive the LDA score after projection that we denote by $W_{\text{P-LDA}}$ as

$$\begin{aligned} &W_{\text{P-LDA}}\left(\boldsymbol{x}\right) \\ &= \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^\top \mathbf{W}^\top \left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1} \mathbf{W}\boldsymbol{\mu} + \log\frac{\pi_0}{\pi_1}. \end{aligned}$$

Then for Gaussian data, the conditional probability of misclassification writes as

$$\epsilon_i^{\text{P-LDA}} = \mathbb{P}\left[(-1)^i W_{\text{P-LDA}}\left(\boldsymbol{x}\right) < 0 | \ \boldsymbol{x} \in \mathcal{C}_i, \mathbf{W}\right],\ i \in \{0,1\}, \tag{9}$$

where we also condition on the projection matrix $\mathbf{W}$. For Gaussian data[1], $\epsilon_i^{\text{P-LDA}}$ can be evaluated as

$$\begin{aligned} \epsilon_i^{\text{P-LDA}} = \Phi\Bigg[ &-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top\mathbf{W}^\top\left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1}\mathbf{W}\boldsymbol{\mu}} \\ &+ \frac{(-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top\mathbf{W}^\top\left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1}\mathbf{W}\boldsymbol{\mu}}}\Bigg]. \end{aligned} \tag{10}$$

In the following, we provide the main result of the paper related to the derivation of the asymptotic performance of projected-LDA in terms of the probability of misclassification. The main result is summarized in the following proposition.

[1]This assumption can be relaxed in the large dimensional setting we are considering because of the linear structure in the classifier which allows the LDA score to asymptotically behave as a Gaussian random variable. However, for simplicity we stick to the Gaussian assumption for the rest of the paper.

**Proposition 1** (Asymptotic performance)**.** *Under Assumptions 1 and 2, then for* $i \in \{0,1\}$ *the conditional probability of misclassification in* (10) *converges in probability to a non trivial deterministic limit given by*

$$\begin{aligned} &\epsilon_i^{\text{P-LDA}} - \Phi\left[\frac{-\frac{1}{2}\boldsymbol{\mu}^\top\left(\mathbf{C}+\delta_d\mathbf{I}_p\right)^{-1}\boldsymbol{\mu} + (-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top\left(\mathbf{C}+\delta_d\mathbf{I}_p\right)^{-1}\boldsymbol{\mu}}}\right] \\ &\to_{prob.} 0, \end{aligned} \tag{11}$$

*where* $\delta_d$ *is such that*

$$\delta_d \operatorname{tr}\left(\mathbf{C}+\delta_d\mathbf{I}_p\right)^{-1} = p - d. \tag{12}$$

*Proof.* The proof relies on computing a *deterministic equivalent* for the quantity $\boldsymbol{\mu}^\top\mathbf{W}^\top\left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1}\mathbf{W}\boldsymbol{\mu}$ when Assumptions 1 and 2 are satisfied. We start by writing

$$\begin{aligned} &\boldsymbol{\mu}^\top\mathbf{W}^\top\left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1}\mathbf{W}\boldsymbol{\mu} \\ &= \lim_{t\downarrow 0}\frac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}^\top\left(\frac{1}{p}\mathbf{Z}\mathbf{C}\mathbf{Z}^\top + t\mathbf{I}_d\right)^{-1}\mathbf{Z}\boldsymbol{\mu} \\ &= \lim_{t\downarrow 0}\frac{1}{p}\tilde{\boldsymbol{\mu}}^\top\mathbf{C}^{1/2}\mathbf{Z}^\top\left(\frac{1}{p}\mathbf{Z}\mathbf{C}\mathbf{Z}^\top + t\mathbf{I}_d\right)^{-1}\mathbf{Z}\mathbf{C}^{1/2}\tilde{\boldsymbol{\mu}} \\ &\overset{(a)}{=} \lim_{t\downarrow 0}\frac{1}{p}\tilde{\boldsymbol{\mu}}^\top\mathbf{C}^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}^{1/2}\left(\frac{1}{p}\mathbf{C}^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}^{1/2} + t\mathbf{I}_p\right)^{-1}\tilde{\boldsymbol{\mu}} \\ &= \|\tilde{\boldsymbol{\mu}}\|^2 - \lim_{t\downarrow 0} t\tilde{\boldsymbol{\mu}}^\top\left(\frac{1}{p}\mathbf{C}^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}^{1/2} + t\mathbf{I}_p\right)^{-1}\tilde{\boldsymbol{\mu}} \\ &= \|\tilde{\boldsymbol{\mu}}\|^2 - \lim_{t\downarrow 0}\tilde{\boldsymbol{\mu}}^\top\left(\frac{1}{tp}\mathbf{C}^{1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}^{1/2} + \mathbf{I}_p\right)^{-1}\tilde{\boldsymbol{\mu}}, \end{aligned}$$

where $\tilde{\boldsymbol{\mu}} = \mathbf{C}^{-1/2}\boldsymbol{\mu}$ and $(a)$ follows from the Woodbury matrix identity. Theorem 1 in [17] is of special interest to us since 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 $\mathbf{Q}\left(t\right) \in \mathbb{R}^{p\times p}$ in the sense that

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

for all deterministic $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^p$ with uniformly bounded Euclidean norms and $t > 0$. $\mathbf{Q}\left(t\right)$ is a deterministic matrix given by $\mathbf{Q}\left(t\right) = \left(\mathbf{I}_p + \frac{\frac{d}{tp}}{1+\frac{d}{tp}\delta(t)}\mathbf{C}\right)^{-1}$, where $\delta(t)$ satisfies $\delta(t) = \frac{1}{d}\operatorname{tr}\mathbf{C}\mathbf{Q}\left(t\right)$. As $t$ approaches 0, $\delta(t)$ approaches $\delta_d$ given by (12) and $\mathbf{Q}\left(t\right)$ approaches $\left(\mathbf{I}_p + \frac{1}{\delta_d}\mathbf{C}\right)^{-1}$. By Assumption 1, it is easy to verify that $\|\tilde{\boldsymbol{\mu}}\|^2$ is uniformly bounded. Therefore, by simple application of the previous results and due to the analycity with respect to $t$ of the above functionals in a neighborhood of 0,

we get

$$\boldsymbol{\mu}^\top \mathbf{W}^\top \left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1} \mathbf{W}\boldsymbol{\mu} - \Bigg[\boldsymbol{\mu}^\top \mathbf{C}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{C}^{-1/2}\left(\frac{1}{\delta_d}\mathbf{C} + \mathbf{I}_p\right)^{-1} \mathbf{C}^{-1/2}\boldsymbol{\mu}\Bigg] \to_{prob.} 0.$$

By simple manipulations, we get

$$\boldsymbol{\mu}^\top \mathbf{W}^\top \left(\mathbf{W}\mathbf{C}\mathbf{W}^\top\right)^{-1} \mathbf{W}\boldsymbol{\mu} - \boldsymbol{\mu}^\top \left(\mathbf{C} + \delta_d \mathbf{I}_p\right)^{-1} \boldsymbol{\mu} \to_{prob.} 0.$$

Finally, the proof is concluded by simple application of the continuous mapping theorem. □

The first implication of Proposition 1 is that $\mathbf{W}$ impacts the performance only through the reduced dimension $d$ which can be explained by the absence of structure in the projection matrix as detailed in Assumption 2. Moreover, the result given by proposition 1 is universal in the sense that regardless of the distribution of the projection matrix, the performance will converge to a universal limit that only depends on the data statistics and the dimensions involved. To clearly gauge the performance loss incurred by dimensionality reduction, we consider the case of equal priors which leads to

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

which as expected worse than $\Phi\left[-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top \mathbf{C}^{-1}\boldsymbol{\mu}}\right]$, the performance that we would get if we retain the full dimension $p$ since $\delta_d > 0$. In the case where $\mathbf{C} = \mathbf{I}_p$, it is also easy to show that

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

as compared to $\Phi\left[-\frac{1}{2}\|\boldsymbol{\mu}\|\right]$ with full dimension.

### 3.3. Experiments

We evaluate the performance of randomly projected-LDA using two types of random matrices. The first one belongs to the class of Gaussian random matrices where the entries are zero mean unit variance and all moments are finite which satisfies Assumption 2. The second type is given by the class of Bernoulli random matrices where the entries are generated as follows

$$Z_{i,j} = 1 - 2B_{i,j}, \;\; B_{i,j} \sim \textbf{Bernoulli}(1/2),$$

which satisfies Assumption 2. The top two sub-figures in Figure 1 are obtained for equal priors ($\pi_0 = \pi_1$) Gaussian generated data as in (1) with $p = 800$, $\boldsymbol{\mu}_0 = \mathbf{0}_p$, $\boldsymbol{\mu}_1 = \frac{3}{\sqrt{p}}\mathbf{1}_p$ and $\mathbf{C} = \{0.4^{|i-j|}\}_{i,j}$. The empirical performance is obtained by evaluating the misclassification rate over $10^4$ testing samples. The bottom sub-figures are obtained for the popular MNIST dataset

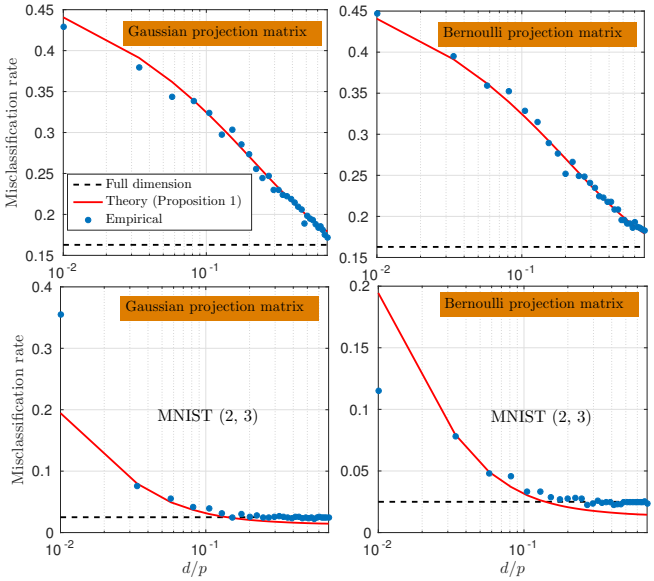


**Fig. 1**. Misclassification rate of randomly-projected LDA.

[18] where $\mathcal{C}_0$ is taken to be the digit 2 whereas $\mathcal{C}_1$ is given by digit 3. For MNIST data, we obtain the data statistics by relying on sample estimates computed from the training data. As we can see in Figure 1, the prediction obtained by proposition 1 is highly accurate for Gaussian data especially when $d$ is relatively large to comply with Assumption 1. Although our derivations heavily rely on the Gaussian assumption of the data, the theoretical formula obtained in Proposition 1 is able to give relatively accurate predictions on the performance for MNIST data as well. Finally, the universality property discussed earlier is also verified since both Gaussian and Bernoulli random projections yield almost the same classification performance especially for Gaussian data.

## 4. CONCLUSIONS AND FUTURE WORKS

The paper investigated the performance of projected-LDA in terms of the probability of misclassification. Under mild assumptions on the data statistics and the projection matrix and using fundamental results from random matrix theory, we showed that the performance asymptotically converges to a deterministic limit that relates the performance in terms of the problem's statistics and dimensions. Numerical results have been provided to support our theoretical claim for both synthetic and real data. A possible future extension of the present work is to consider the performance with estimated statistics. The analysis can also be extended to investigate the performance quadratic discriminant analysis and other model based classification algorithms.


# Acknowledgment

The authors thank Vahid Tarokh for valuable discussions.

## 5. REFERENCES


[1] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[2] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The Elements of Statistical Learning*, Springer, 2009.

[4] J. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165–175, 1989.

[5] P. J. Bickel and E. Levina, "Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and Some Alternatives when There are More Variables than Observations," *Bernoulli*, vol. 10, pp. 989–1010, 2004.

[6] R. J. Durrant and A. Kaban, "Sharp Generalization Error Bounds for Randomly-Projected Classifiers," *J. Mach. Learn. Res.*, vol. 28, pp. 693–701, 2013.

[7] B. McWilliams, C. Heinze, N. Meinshausen, G. Krummenacher, and H. P. Vanchinathan, "LOCO: Distributing Ridge Regression with Random Projections," *arXiv e-prints, 1406.3469v2*, 2014.

[8] Timothy I. Cannings and Richard J. Samworth, "Random-Projection Ensemble Classification," *https://arxiv.org/pdf/1504.04595.pdf*, 2017.

[9] R. Calderbank, S. Jafarpour, and R Schapire, "Compressed Learning: Universal Sparse Dimensionality Reduction and Learning in the Measurement Domain," *Technical Report, Rice University*, 2009.

[10] R.I. Arriaga and S. Vempala, "An Algorithmic Theory of Learning: Robust Concepts and Random Projection," *IEEE 40th Annual Symposium on Foundations of Computer Science (FOCS 1999)*, 1999.

[11] A. Zollanvari and E. R. Dougherty, "Generalized Consistent Error Estimator of Linear Discriminant Analysis," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2804–2814, June 2015.

[12] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. AlNaffouri, and M.-S. Alouini, "A Large Dimensional Study of Regularized Discriminant Analysis Classifiers," *https://arxiv.org/pdf/1711.00382.pdf*, 2017.

[13] Nir Ailon and Edo Liberty, "An Almost Optimal Unrestricted Fast Johnson-Lindenstrauss Transform," *ACM Transactions on Algorithms (TALG)*, vol. 9, 2013.

[14] S. Dasgupta and A. Gupta, "An elementary proof of the Johnson–Lindenstrauss Lemma," *Random Struct. Alg.*, vol. 22, pp. 60–65, 2002.

[15] R. J. Durrant and A. Kaban, "Sharp Generalization Error Bounds for Randomly-Projected Classifiers," *J. Mach. Learn. Res.*, vol. 28, pp. 693–701, 2013.

[16] W. Hachem, P. Loubaton, and J. Najim, "Deterministic Equivalents For Certain Functionals Of Large Random Matrices," *The Annals of Applied Probability*, vol. 17, no. 3, pp. 3987–4004, 2007.

[17] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, "A New Approach for Mutual Information Analysis of Large Dimensional Multi-Antenna Channels," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 3987–4004, Sept 2008.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-Based Learning Applied To Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.