Journal of Machine Learning Research 25 (2024) 1-65 Submitted 11/22; Revised 7/24; Published 10/24

An Asymptotic Study of Discriminant and Vote-Averaging Schemes for Randomly-Projected Linear Discriminants

Lama B. Niyazi LAMA.NIYAZI@KAUST.EDU.SA

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Abla Kammoun ABLA.KAMMOUN@KAUST.EDU.SA

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Hayssam Dahrouj HAYSSAM.DAHROUJ@GMAIL.COM

University of Sharjah
Sharjah, United Arab Emirates

Mohamed-Slim Alouini SLIM.ALOUINI@KAUST.EDU.SA

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Tareq Y. Al-Naffouri TAREQ.ALNAFFOURI@KAUST.EDU.SA

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Editor: Marc Schoenauer

Abstract

Modern technology has contributed to the rise of high-dimensional data in various domains such as bio-informatics, chemometrics, and face recognition. In the recent literature, random projections and, in particular, randomly-projected ensembles based on the classical Linear Discriminant Analysis (LDA), have been proposed for classification problems involving such high-dimensional data. In this work, we study the two main classes of randomly-projected LDA ensemble classifiers, namely discriminant averaging and vote averaging. Through asymptotic analysis in a growth regime where the problem dimensions are assumed to grow at constant rates to each other for a fixed ensemble size, we determine the exact mechanism through which the ensemble size affects the classification performance. Furthermore, we investigate whether projection selection truly matters in an ensemble setting, and, ultimately, derive the optimal form of the randomly-projected LDA ensemble. Motivated by these findings, we propose a framework for efficient tuning of the optimal classifier's ensemble size and projection dimension based on an estimator of the classifier probability of misclassification which is consistent under the assumed growth regime. The proposed framework is shown to outperform the existing rule-of-thumb, as well as other methods for parameter tuning, on both real and synthetic data.

Keywords: linear discriminant analysis, random projection, high-dimensional data, small sample size, classification ensembles, random matrix theory

1. Introduction

Since its inception in 1936 (Fisher, 1936), Linear Discriminant Analysis (LDA) has remained a popular choice of classifier across a wide variety of domains. This is not surprising, as the LDA classifier exhibits numerous optimality properties (Niyazi et al., 2022) and, despite its simplicity, has proven to be a powerful competitor among more sophisticated methods (Lim et al., 2000; Hand, 2006).

The contemporary emergence of high-dimensional data in the form of text, images, mass spectra, and gene microarrays, poses a challenge for conventional classification methods. While statistical inference generally suffers performance-wise from the curse of dimensionality in high-dimensional data settings, the LDA classifier, in particular, becomes intractable when data dimensionality exceeds sample size. This is due to the singularity of the sample covariance estimator under these conditions which are commonly encountered in data sets for various applications, such as chemometrics, face recognition, and tumor classification. On such data, both linear and non-linear models tend to perform comparably (Yuan et al., 2012), with the former enjoying faster training times and greater stability (Beleites and Salzer, 2008). As a result, many high-dimensional variants of LDA have been proposed over the years. For instance, one popular category of variants is based on alternative estimators of the covariance/precision matrix. This includes regularized estimators, yielding the family of Regularized-LDA (R-LDA) (Guo et al., 2007) variants, and simplified estimators, yielding variants such as diagonal LDA (Hastie et al., 2009). Principle components analysis, random projection, and other dimensionality reduction techniques, compose another category of variants which operate based on the principle of reducing data dimensionality relative to the sample size so that the sample covariance matrix is invertible. In addition to solving the singularity problem, these techniques are able to achieve significant computational savings as a result of the reduced problem dimensions. For detailed accounts of these methods and others, the reader is referred to the papers of Mai (2013) and Sharma and Paliwal (2015). In the current work, we focus on high-dimensional variants of LDA based on dimensionality reduction by random projection. In particular, we study *ensembles* of randomly-projected LDA (RP-LDA) discriminants.

Durrant and Kabán (2010) were the first to study the classifier consisting of a single RP-LDA discriminant. The single RP-LDA discriminant classifier is known to perform poorly in practice, while ensemble classifiers which combine multiple RP-LDA discriminants perform relatively well. The literature on RP-LDA ensemble classifiers can broadly be divided into two categories: discriminant-averaging ensembles (Durrant and Kabán, 2015; Peressutti et al., 2015) and vote-averaging ensembles (Schclar and Rokach, 2009; Cannings and Samworth, 2017). Discriminant-averaging ensembles average all RP-LDA discriminants followed by thresholding the result to obtain the final class prediction. Interestingly, this is equivalent to estimating the precision matrix in the LDA discriminant by the finite version of the Marzetta estimator (Marzetta et al., 2011). Vote-averaging ensembles average the class prediction corresponding to each individual RP-LDA discriminant before thresholding to obtain the final prediction. The main difference between various implementations within each category is whether the projections are subjected to a preliminary selection process for inclusion within the ensemble based on some criterion of their expected performance.

Although theoretical analyses of both categories of RP-LDA ensembles exist, these studies have their limitations and raise several key questions. Durrant and Kabán (2015) derive error bounds for the basic form of the discriminant-averaging RP-LDA ensemble classifier without selection; however, the analysis is based on an abstraction wherein the ensemble size grows to infinity, revealing the converged Marzetta estimator of the precision matrix within the classifier discriminant. This, in addition to a Gaussian data assumption, form the basis for the asymptotic analysis of the discriminant-averaging ensemble conducted by Niyazi et al. (2020a), where it is found that the ensemble behaves as a a special case of R-LDA. This result implies that the discriminant-averaging RP-LDA ensemble classifier can never outperform a properly-tuned R-LDA classifier on Gaussian data. Cannings and Samworth (2017) provide bounds on the error difference between the vote-averaging RP-LDA ensemble (with and without selection) and the Bayes error, but this bound is not a function of the number of projections in the ensemble. While these findings are useful, an analysis which takes into account the number of projections allows for a more accurate characterization of the practical performance of these classifiers.

Kabán (2017, 2020) studies the performance of the finite versus the converged version of the Marzetta estimator of the precision matrix and finds that in order to achieve a certain tolerance on the spectral norm of the difference between the two, the ensemble size must grow linearly with the data dimension. A shortcoming of this approach is that it neither provides a measure of efficacy of the finite Marzetta estimator with respect to the true measures of interest in classification, such as misclassification rate, nor does it provide practical guidelines on how to choose the number of projections. Of particular concern is the selection of an ensemble size that is sufficient to achieve satisfactory performance, yet not so large that the computational savings provided by dimensionality reduction are lost.

Another glaring gap within the literature is the lack of a thorough comparison between the discriminant averaging and vote averaging RP-LDA ensembles. An attempt at this was made by Cannings (2021); however, bearing in mind that the target data sets for these types of classifiers are small samples of high dimensional data, it must be noted that this study is extremely limited with regards to the dimensionality and variety of data considered. In any case, beyond merely comparing the two types of ensembles, one would ultimately like to know the overall best way of combining any given set of RP-LDA discriminants.

The current work addresses the aforementioned issues through a comprehensive study of randomly-projected linear discriminant ensembles by asymptotic analysis under Gaussian data assumptions using Random Matrix Theory (RMT) tools in a growth regime where the data and projection dimensions grow together. This growth regime is chosen specifically in order to more accurately represent the small-sample finite regime, where the data dimensionality is greater than the number of samples, in contrast to the classical regime, where the number of samples is much greater than the data dimensionality. The analysis yields a number of insightful results stemming from the asymptotic distributions of the discriminant-averaging and vote-averaging RP-LDA ensemble classifiers and limits of their discriminant statistics and probabilities of misclassification. The main findings are:

- The class-conditional discriminant means of the discriminant-averaging RP-LDA ensemble are asymptotically identical, regardless of ensemble size.

- The class-conditional discriminant means of the vote-averaging RP-LDA ensemble are asymptotically identical, regardless of ensemble size.

- The asymptotic class-conditional variance of the discriminant-averaging RP-LDA ensemble is a convex combination of that of the single RP-LDA discriminant and the infinite ensemble. The asymptotic variance corresponding to the single RP-LDA discriminant is shown to be strictly greater than the asymptotic variance of the infinite ensemble.

- The asymptotic class-conditional variance of the vote-averaging RP-LDA ensemble is a convex combination of that of a vote on a single RP-LDA discriminant, that is, a Bernoulli variance, and a fraction of this Bernoulli variance. Thus, the asymptotic variance corresponding to the single vote is strictly greater than the asymptotic variance of the infinite ensemble.

- Each class-conditional discriminant of the discriminant-averaging RP-LDA ensemble is asymptotically Gaussian with parameters being the limits of their corresponding exact discriminant statistics.

- Each class-conditional discriminant of the vote-averaging RP-LDA ensemble is asymptotically a normalized constantly-correlated Binomial with parameters being the limits of their corresponding exact probabilities of success and correlations between trials.

To elaborate on the above, one of the major contributions of this study is that it shows the direct effect of the RP-LDA ensemble size on its classification performance through the asymptotic class-conditional discriminant means and variances. More specifically, since the ensemble size acts to decrease the variance of the discriminant from its maximum at a single projection to its minimum when the ensemble size grows to infinity, while maintaining a constant mean separation, the misclassification rate decreases monotonically with increasing ensemble size. This is true for both the discriminant-averaging and vote-averaging RP-LDA ensemble classifiers. Additionally, access to the single RP-LDA discriminant asymptotic distribution allows for a derivation of the asymptotically optimal way of constructing the ensemble via the Neyman-Pearson lemma and Maximum A Posteriori (MAP) rule. These results reveal that, for Gaussian data, the optimal ensemble is linear in form, that is, it is a form of discriminant averaging, wherein all projections are weighted equally, implying that projection selection is asymptotically sub-optimal in the context of RP-LDA ensemble classification.

The theoretical analysis in this paper leads to several significant implications for the deployment of RP-LDA ensemble classifiers in practice, which are verified through simulations on both real and synthetic data. Firstly, our simulations suggest that there is generally no need to look beyond the basic discriminant-averaging RP-LDA ensemble classifier. Although the theoretical guarantee on which this is based assumes Gaussian data, we find that, on real data, this classifier generally performs just as well, if not better, than its immediate competitors. Secondly, as mentioned previously, classifier performance can only increase with increasing ensemble size. Thus, the infinite ensemble represents the classifier's full classification potential, and finite ensemble performance may be assessed relative to the infinite ensemble. Based on this, a framework for tuning the discriminant-averaging RP-LDA ensemble size and projection dimension is proposed. An estimator of the probability of misclassification which is consistent in the RMT growth regime is derived for use in this framework. This estimator has the advantage of greater computational efficiency compared to conventional empirically-produced estimators of the test error, such as cross-validation, as well as dispensing of the need for additional data. Different variants of the tuning algorithm are implemented on real and synthetic data and compared in terms of performance and computational complexity.

The rest of this paper is structured as follows. Section 2 sets the background for this work in terms of classification setting and the classifiers to be studied. Section 3 presents the results pertaining to the asymptotic analysis using RMT, while Section 4 translates these findings into practice. Finally, Section 5 concludes this work and discusses its limitations.

Throughout the paper, scalars are denoted by plain lower-case letters, vectors by bold lower-case letters, and matrices by bold upper-case letters. The symbol $\mathbf{I}_p$ is used to represent the $p \times p$ identity matrix, the symbol $\mathbf{1}_p$ represents the all-ones $p \times 1$ vector, and the symbol $\mathbf{0}_p$ represents the all-zeros $p \times 1$ vector. The notation $||\cdot||$ is used to symbolize the Euclidean norm when its argument is a vector and the spectral norm when its argument is a matrix. Almost-sure convergence is denoted by $\xrightarrow{\text{a.s.}}$ or $a \asymp b$ which means $a - b \xrightarrow{\text{a.s.}} 0$. As defined by Benaych-Georges and Couillet (2016), for a sequence of random square matrices $\mathbf{A}$ and $\mathbf{B}$ of size n, $\mathbf{A} \leftrightarrow \mathbf{B}$ means that $\frac{1}{n}\text{tr}\left\{\mathbf{D}\left(\mathbf{A}-\mathbf{B}\right)\right\} \xrightarrow{p} 0$ and $\mathbf{d}_1^T\left(\mathbf{A}-\mathbf{B}\right)\mathbf{d}_2 \xrightarrow{p} 0$ for all sequences $\mathbf{D}$ of deterministic $n \times n$ matrices of bounded norms and all deterministic sequences of vectors $\mathbf{d}_1$, $\mathbf{d}_2$ of bounded norms. The function $\Phi(\cdot)$ denotes the standard Gaussian Cumulative Distribution Function (CDF), $Q\left(\cdot\right)$ is the standard Gaussian complementary CDF, that is, the *Q-function* or *survival function*, namely $Q\left(\cdot\right) := 1 - \Phi(\cdot)$, and the $\sim$ symbol stands for 'distributed as'.

2. Problem Formulation

This section lays out the classification setting and formally defines the discriminant-averaging and vote-averaging RP-LDA ensemble classifiers which are studied in this work.

2.1 Classification Setting

This paper assumes a data setting where the training set, $\mathcal{T}$, as well as the test point, $\mathbf{x} \in \mathbb{R}^p$, are drawn from a Gaussian mixture model consisting of two classes $\mathcal{C}_0$ and $\mathcal{C}_1$ having means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, respectively, a common covariance $\mathbf{\Sigma}$, and prior probabilities π_0 and π_1, respectively, that is,

$$\mathbf{x}|\mathbf{x} \in \mathcal{C}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}),\ i = 0, 1, \tag{1}$$

and

$$\pi_i := P[\mathbf{x} \in \mathcal{C}_i],\ i = 0, 1.$$

Furthermore, $\mathcal{T}$ consists of a total of n labelled training points with n_i points belonging to class $\mathcal{C}_i$, $i = 0, 1$, that is, $n = n_0 + n_1$.

Estimates $\hat{\boldsymbol{\mu}}_i,\ i = 0, 1$, $\hat{\mathbf{\Sigma}}$, and $\hat{\pi}_i,\ i = 0, 1$, of $\boldsymbol{\mu}_i,\ i = 0, 1$, $\mathbf{\Sigma}$, and $\pi_i,\ i = 0, 1$, respectively, are computed from the training data. The estimates are defined as follows. Let

$\mathbf{X}_0 \in \mathbb{R}^{p\times n_0}$ and $\mathbf{X}_1 \in \mathbb{R}^{p\times n_1}$ be matrices whose columns are the individual training samples corresponding to $\mathcal{C}_0$ and $\mathcal{C}_1$ respectively, then $\hat{\boldsymbol{\mu}}_0 := \frac{1}{n_0}\mathbf{X}_0\mathbf{1}_{n_0}$, $\hat{\boldsymbol{\mu}}_1 := \frac{1}{n_1}\mathbf{X}_1\mathbf{1}_{n_1}$, $\hat{\mathbf{\Sigma}} := \frac{(n_0-1)\hat{\mathbf{\Sigma}}_0+(n_1-1)\hat{\mathbf{\Sigma}}_1}{n_0+n_1-2}$, $\hat{\pi}_0 := \frac{n_0}{n}$, and $\hat{\pi}_1 := \frac{n_1}{n}$ where $\hat{\mathbf{\Sigma}}_0 := \frac{1}{n_0-1}\left(\mathbf{X}_0 - \hat{\boldsymbol{\mu}}_0\mathbf{1}_{n_0}^T\right)\left(\mathbf{X}_0 - \hat{\boldsymbol{\mu}}_0\mathbf{1}_{n_0}^T\right)^T$ and $\hat{\mathbf{\Sigma}}_1 := \frac{1}{n_1-1}\left(\mathbf{X}_1 - \hat{\boldsymbol{\mu}}_1\mathbf{1}_{n_1}^T\right)\left(\mathbf{X}_1 - \hat{\boldsymbol{\mu}}_1\mathbf{1}_{n_1}^T\right)^T$. These estimates are used to construct various discriminants which are in turn subjected to a threshold. The thresholding determines the predicted class of the test point $\mathbf{x}$. Formally, given a discriminant $W(\mathbf{x})$, the decision rule $C(\mathbf{x})$ takes the form

$$C(\mathbf{x}) = \begin{cases} 1, & \text{if } W(\mathbf{x}) > \zeta \\ 0, & \text{otherwise,} \end{cases}$$

where $\zeta \in \mathbb{R}$ is a classifier-specific threshold, and $C(\mathbf{x}) = i,\ i = 0, 1$, indicates the class prediction $\mathcal{C}_i$. Note that there is no loss of generality with respect to the exact value of ζ chosen for the analyses of any of the classifiers considered in this work, as the main subject of analysis in each case is the discriminant.

2.2 The Single Randomly-Projected LDA Discriminant

Denote by $\mathbf{R} \in \mathbb{R}^{d\times p}$ a Gaussian projection with i.i.d. entries distributed as $\mathcal{N}(0, 1/d)$. In order to construct a randomly-projected LDA discriminant, the training data $\mathcal{T}$ is projected as $\mathbf{R}\mathbf{X}_0$ and $\mathbf{R}\mathbf{X}_1$. The sample statistic estimates are then computed based on this projected data. It is easy to show (see Niyazi et al., 2020a) that the resulting single randomly-projected LDA discriminant, denoted by $W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R})$, has the form

$$W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) = \hat{\boldsymbol{\mu}}^T \hat{\mathbf{\Sigma}}_{\mathbf{R}}^{-1}\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}, \tag{2}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{R}}^{-1} = \mathbf{R}^T(\mathbf{R}\hat{\mathbf{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}$. The discriminant is then subjected to a threshold of 0 to obtain the final classification.

Of interest in the analysis of this paper is the discriminant's behavior in terms of the mean separation between and variance of the distribution of projected points from each of the two classes. We denote the class-conditional means and variances of this particular discriminant by

$$m_i(1) := \mathbb{E}\left[W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R})\,|\mathbf{x} \in \mathcal{C}_i\right],\ i = 0, 1 \tag{3}$$

and

$$\sigma^2(1) := \text{Var}\left[W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R})\,|\mathbf{x} \in \mathcal{C}_i\right],\ i = 0, 1, \tag{4}$$

respectively, where the expectation and variance are with respect to the training data, test point, and the random projection, conditioned on the test point's class (the '1' in the argument indicates that these quantities correspond to a single random projection). As it is difficult to compute these quantities exactly, they are analyzed asymptotically in Section 3.1.

As mentioned in Section 1, the randomly-projected LDA variant based on a single random projection does not perform well in practice. The two main ensemble-based schemes that have been proposed in order to improve upon the performance of this classifier are the discriminant-averaging and vote-averaging ensembles. These are considered in the next section.

2.3 Randomly-Projected LDA Ensemble Classifiers

This section defines the discriminants and class-conditional discriminant statistics of the discriminant-averaging and vote-averaging RP-LDA ensemble classifiers which are analyzed in this paper.

2.3.1 Discriminant-Averaging Ensemble

A discriminant-averaging ensemble of RP-LDA discriminants averages multiple discriminants, each of which corresponds to an independently-realized random projection. In this paper, we focus on a particular discriminant-averaging ensemble which weights the contribution of each projection equally rather than according to some measure of how 'good' they are. In fact, we establish later in the paper that this *equally-weighted* discriminant-averaging scheme is asymptotically optimal under the data distribution assumptions detailed in Section 2.1.

Now, let us formally define the discriminant-averaging ensemble of interest. Letting $\mathbf{R}_k$ correspond to the k^{th} projection among M random projections $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M$, the discriminant-averaging scheme which assigns equal weights to each RP-LDA discriminant is constructed as (Durrant and Kabán, 2015)

$$W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) = \frac{1}{M} \sum_{k=1}^{M} W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right). \tag{5}$$

One can imagine a scheme in which the weights of $1/M$ in (5) vary for each discriminant as a function of its random projection. Furthermore, the weights may take on binary values of zero and one, thus excluding certain projections altogether. This is referred to as 'projection selection' in this paper.

Peressutti et al. (2015) employ a kind of projection selection where selection occurs through a process of generating a projection to form a single RP-LDA discriminant, followed by subjecting the resulting classifier to a predefined threshold on training error. This is repeated until a satisfactory projection is found, and until a minimum number of satisfactory discriminants are collected. There is no direct correspondence between this scheme and (5), as the total number of projections (M) is not known in advance.

In their paper, Peressutti et al. (2015) do not compare the proposed selection scheme against no selection. Moreover, they select the projections based on the same data that is used to evaluate classifier performance: the projections are chosen using the resubstitution error on the training set and the final ensemble evaluated using cross-validation on the same training set. This results in an overestimate of the performance gain that can be attributed to selection (Cawley and Talbot, 2010). We take care to avoid this bias in this work through nesting the selection within the cross-validation loops, as is detailed in Section 3.2.2.

The discriminant (5) is subjected to a threshold of 0 to obtain the final classification. The recommended projection dimension setting of this classifier according to empirical observations made by Durrant and Kabán (2015) is $d = \frac{\text{rank}\{\hat{\boldsymbol{\Sigma}}\}}{2}$. We denote the equally-weighted discriminant-averaging RP-LDA ensemble discriminant's class-conditional means and variances by

$$m_i(M) := \mathbb{E}\left[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i\right], \ i = 0, 1 \tag{6}$$

and

$$\sigma^2(M) := \mathrm{Var}\left[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i\right], \ i = 0, 1, \tag{7}$$

respectively.

Moreover, of theoretical interest in this study is the 'infinite ensemble' for which the number of randomly-projected LDA discriminants in the ensemble, each corresponding to an independent projection, grows to infinity. Its discriminant is defined as (Durrant and Kabán, 2015)

$$\begin{aligned} W_{M=\infty}(\mathbf{x}) &:= \lim_{M\to\infty} W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) \\ &= \hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}, \end{aligned} \tag{8}$$

where $\mathbb{E}_{\mathbf{R}}[\cdot]$ is the expectation with respect to the random projection $\mathbf{R}$, conditioned on the training data and test point, and $\mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]$ is, in fact, the Marzetta estimator of the precision matrix (Durrant and Kabán, 2015; Marzetta et al., 2011). This classifier sets an upper bound on finite ensemble performance, which may be approached by employing a very large number of projections. The work by Durrant and Kabán (2015) suggests that it suffices to use the discriminant in (5) with $M = 100$ to approximate (8) in practice, since, according to their simulations, there is very little empirical difference between ensembles with $M = 100$ projections versus $M = 3000$ projections. We denote the infinite ensemble discriminant's class-conditional means and variances by

$$m_i^{M=\infty} := \mathbb{E}\left[W_{M=\infty}(\mathbf{x})|\mathbf{x} \in \mathcal{C}_i\right], \ i = 0, 1$$

and

$$\sigma^2_{M=\infty} := \mathrm{Var}\left[W_{M=\infty}(\mathbf{x})|\mathbf{x} \in \mathcal{C}_i\right], \ i = 0, 1,$$

respectively. The asymptotic analysis of these quantities is presented in Section 3.1.

2.3.2 Vote-Averaging Ensemble

In contrast to the discriminant introduced in the previous section. which averages the RP-LDA discriminants, a vote-averaging ensemble discriminant averages the final class votes obtained by thresholding each RP-LDA discriminant. In terms of the set of M random projections $\{\mathbf{R}_k\}_{k=1}^M$, the equally-weighted vote-averaging ensemble discriminant is defined as (Cannings and Samworth, 2017)

$$W_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) = \frac{1}{M}\sum_{k=1}^{M} \mathbb{1}\left\{W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right) > 0\right\}. \tag{9}$$

The discriminant is subjected to a threshold of 0.5 to obtain the final classification. This threshold corresponds to a majority vote.

With the aim of exploiting observed differences in classification performance among random projections, Cannings and Samworth (2017) propose a projection selection scheme on top of the basic vote-averaging RP-LDA ensemble. They generate a number $B_1 \in \mathbb{N}$

of disjoint groups of a number $B_2 \in \mathbb{N}$ of projections each and select the projection from each group which yields the lowest error rate according to an error estimator of choice. The final set of projections is used to build the discriminant in (9) composed of a total of B_1 projections. Technically, this corresponds to (9) with a total of $B_1 \times B_2$ projections taking on binary weights of zeros and ones. Again, Cannings and Samworth (2017) do not compare their proposed selection scheme to no selection in an ensemble setting. The simulations in subsequent sections of this paper look further into both the question of whether the intuitive basis for projection selection holds in an ensemble setting, and the question of how to choose the number of projections for an ensemble.

The next section presents insights into the behavior of both of these classifiers based on asymptotic analysis using RMT tools.

3. Asymptotic Insights

This section draws several insights into the behavior of RP-LDA classifiers from the asymptotic analyses of the single RP-LDA discriminant, the discriminant-averaging RP-LDA finite ensemble discriminant, the discriminant-averaging RP-LDA infinite ensemble discriminant, and the vote-averaging RP-LDA ensemble discriminant. For two-class, p-dimensional, Gaussian data as in (1), n the number of data samples with n_i samples corresponding to each class, and d the projection dimension, the conditions under which these analyses hold are:

(a) $0 < \liminf \frac{p}{n} < \limsup \frac{p}{n} < \infty$

(b) $0 < \liminf \frac{d}{n} < \limsup \frac{d}{n} < 1$

(c) $0 < \liminf \frac{d}{p} < \limsup \frac{d}{p} < 1$

(d) $\frac{n_i}{n} \to c_i \in (0, 1),\ i = 0, 1$

(e) $\limsup_p \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 < \infty$

(f) $\limsup_p \|\boldsymbol{\Sigma}\|_2 < \infty$

(g) $\liminf_p \lambda_{\min}(\boldsymbol{\Sigma}) > 0.$

Conditions (a), (b), (c) specify the relationships between the dimensions, p, n, and d, in the growth regime considered where p, n, and d grow together at constant rates to each other. More specifically, condition (a) implies that p and n grow together with either $p > n$ or $p < n$, condition (b) implies that d and n grow together with d strictly less than n, and condition (c) implies that d and p grow together with d strictly less than p. Note that the fact that the ratios in conditions (a)-(c) are strictly greater than zero ensures that the involved dimensions grow *together*, whereas a ratio of zero would mean that the denominator grows faster than the numerator. By the same logic, condition (d) ensures that there is a sizable portion of points, n_i, $i = 0, 1$, from each class as the number of training points, n, grows. The two conditions (e) and (f) are technicalities stemming from the use of random matrix theory tools. Finally, condition (g) is necessary so that all members of the sequence, $\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T$ (which is inverted in $\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}$) as d, p and n grow, are invertible. Note that the ensemble size, M, is fixed relative to the problem dimensions n, p, and d, unless the notation $M = \infty$ is used, in which case M is allowed to diverge beforehand so that the presented asymptotic results correspond to the converged ensemble.

Remark 1 *Condition* (g) *may not hold in practice. As such, it may be relaxed to the condition that there exists at least one eigenvalue of $\boldsymbol{\Sigma}$ for which the limit infimum is bounded away from zero. This, however, comes at the cost of additional conditions on the projection dimension d. More specifically, if there are $p - k$ such eigenvalues (and therefore k eigenvalues which tend to zero), then we must have $d < p - k$ in order for all members of the sequence $\boldsymbol{R}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}^T$ to be invertible. In practice, d is tuned to optimize performance, and so one need not delve into these technicalities.*

The remainder of this paper references the term *deterministic equivalent.* This is defined in the following.

Definition 2 *(Müller and Debbah, 2016) The Deterministic Equivalent (DE) of the sequence of random variables X_n is a deterministic sequence $\bar{X}_n$ which approximates X_n in the sense that $X_n - \bar{X}_n \xrightarrow{a.s.} 0$ as $n \to \infty$.*

According to the above definition, the difference between the random variable, X_n, and the deterministic quantity, $\bar{X}_n$, converges to zero. In this way, $\bar{X}_n$, itself need not converge for it to exist. Moreover, $\bar{X}_n$ yields an approximation of X_n for every n which becomes increasingly more accurate with increasing n, in contrast to the typical limit which summarizes an entire sequence with one statistic (Müller and Debbah, 2016).

DEs are widely used in the asymptotic analysis of systems which can be modeled through random matrices such as those encountered in communication theory (Couillet and Debbah, 2011) and machine learning (Couillet and Liao, 2022). The following sections detail the results of the asymptotic analysis of the RP-LDA classifiers defined in Section 2 using DEs.

3.1 Convergence of Discriminant Statistics and Asymptotic Distributions

This section presents the convergence results of the class-conditional statistics of the discriminants (2), (5), (8), and (9), and their asymptotic distributions.

Our previous work (see Niyazi et al., 2020a) derived deterministic equivalents for the class-conditional discriminant statistics of the discriminant-averaging RP-LDA infinite ensemble. This work extends those results by deriving analogous results for the single RP-LDA classifier and the discriminant-averaging RP-LDA finite ensemble. Additionally, it is shown that the single RP-LDA discriminant, the discriminant-averaging RP-LDA finite ensemble discriminant, and the discriminant-averaging RP-LDA infinite ensemble discriminant, each conditioned on the class of the test point, are asymptotically Gaussian having parameters which are the deterministic equivalents of their respective (exact) statistics. This allows for comparison between the three classifiers, and thus an understanding of the effect of the ensemble size M on the classification. For completeness, all three sets of results are presented in what follows. The explicit expressions of the DEs are provided in the appendices.

Theorem 3 *(Single RP-LDA discriminant asymptotic distribution) Under conditions* (a) *to* (g)*, for the single RP-LDA discriminant, $W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R})$, defined in* (2) *and $i = 0, 1$,* *we have*

$$1/\bar{\sigma}(1)\left[\left(W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}\right)|\mathbf{x}\in\mathcal{C}_i\right)-\bar{m}_i(1)\right]\xrightarrow{d}\mathcal{N}(0,1),$$

where $\bar{m}_i(1)$ *and* $\bar{\sigma}^2(1)$ *(given by equations* (15) *and* (40)*, respectively) are DEs of* $m_i(1)$ *and* $\sigma^2(1)$*, respectively.*

Proof See Appendix A and Appendix D.3. ■

Theorem 4 *(Discriminant-averaging RP-LDA finite ensemble discriminant asymptotic distribution) Under conditions* (a) *to* (g)*, for the discriminant-averaging RP-LDA finite ensemble discriminant,* $W_{\text{disc-avg}}\left(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M\right)$*, defined in* (5)*, fixed* M*, and* $i=0,1$*, we have*

$$1/\bar{\sigma}(M)\left[\left(W_{\text{disc-avg}}\left(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M\right)|\mathbf{x}\in\mathcal{C}_i\right)-\bar{m}_i(M)\right]\xrightarrow{d}\mathcal{N}(0,1),$$

where $\bar{m}_i(M)$ *and* $\bar{\sigma}^2(M)$ *(given by equations* (15) *and* (44)*, respectively) are DEs of* $m_i(M)$ *and* $\sigma^2(M)$*, respectively.*

Proof See Appendix B and Appendix D.3. ■

Theorem 5 *(Discriminant-averaging RP-LDA infinite ensemble discriminant asymptotic distribution) Under conditions* (a) *to* (g)*, for the discriminant-averaging RP-LDA infinite ensemble discriminant,* $W_{M=\infty}(\mathbf{x})$*, defined in* (8) *and* $i=0,1$*, we have*

$$1/\bar{\sigma}_{M=\infty}\left[\left(W_{M=\infty}\left(\mathbf{x}\right)|\mathbf{x}\in\mathcal{C}_i\right)-\bar{m}_i^{M=\infty}\right]\xrightarrow{d}\mathcal{N}(0,1),$$

where $\bar{m}_i^{M=\infty}$ *and* $\bar{\sigma}^2_{M=\infty}$ *(given by equations* (15) *and* (45)*, respectively) are DEs of* $m_i^{M=\infty}$ *and* $\sigma^2_{M=\infty}$*, respectively.*

Proof See Appendix D.4. ■

Furthermore, Corollary 6 below, concerned with the relationships between the DEs of the class-conditional discriminant statistics, follows from Theorems 3, 4, and 5:

Corollary 6 *(Asymptotic relationships between single RP-LDA, the discriminant-averaging RP-LDA finite ensemble, and the discriminant-averaging infinite ensemble class-conditional discriminant statistics)*

$$\bar{m}_i(1)=\bar{m}_i(M)=\bar{m}_i^{M=\infty},\ i=0,1, \tag{10}$$

$$\bar{\sigma}^2(1)>\bar{\sigma}^2_{M=\infty}, \tag{11}$$

and

$$\bar{\sigma}^2(M)=\frac{1}{M}\bar{\sigma}^2(1)+\left(1-\frac{1}{M}\right)\bar{\sigma}^2_{M=\infty}. \tag{12}$$

Proof See Appendices A.1 and B.1 for the proof of (10), Appendix A.2.1 for the proof of (11), and Appendix B.2 for the proof of (12). ■

The results of Corollary 6 along with the asymptotic distributions stated in the preceding theorems reveal that the class-conditional discriminants corresponding to the single RP-LDA, discriminant-averaging RP-LDA finite ensemble, and discriminant-averaging RP-LDA infinite ensemble classifiers, each tend to a Gaussian distribution with common means across the discriminants. Furthermore, the single RP-LDA classifier discriminant has a variance strictly greater than that of the discriminant-averaging infinite ensemble classifier, while the discriminant-averaging finite ensemble classifier's variance is a convex combination of the two determined by coefficients $1/M$ and $1-1/M$, respectively. Thus, as M increases, the variance of the corresponding discriminant decreases from one extreme to another, all while maintaining a constant mean separation. In light of Corollary 6, the deterministic equivalents, $\bar{m}_i(1)$, $\bar{m}_i(M)$, and $\bar{m}_i^{M=\infty}$, of the class-conditional means are subsequently referred to by a common notation, $\bar{m}_i,\ i=0,1$.

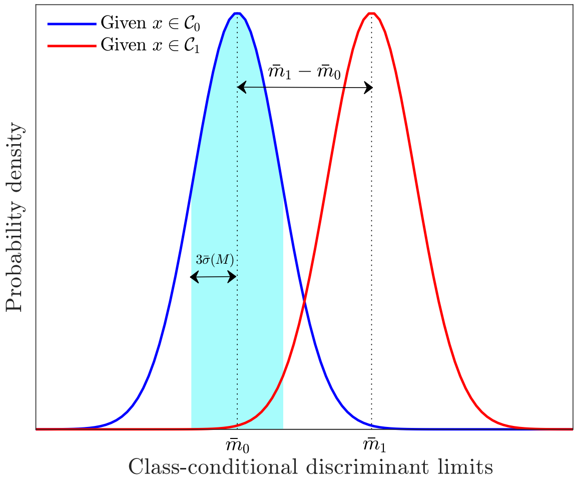

Figure 1: Class-conditional asymptotic distributions of the discriminant-averaging ensemble $M=10$.

Figure 1 shows an example of the asymptotic class-conditional distributions of the discriminant-averaging RP-LDA ensemble discriminant when $M=10$. The figure depicts the probability densities of the class-conditional discriminants, along with the mean separation, $\bar{m}_1 - \bar{m}_0$, and three standard deviations, $3\bar{\sigma}(M)$. Notice that the distributions overlap.

Figure 2 shows the same distributions depicted in Figure 1 alongside the class-conditional asymptotic distributions of the single RP-LDA discriminant and the discriminant-averaging RP-LDA infinite ensemble. Consistent with Corollary 6, the mean separation between class distributions is maintained with increasing ensemble size M while their variance decreases. This leads to less overlap between the distributions with increasing M, as quantified by

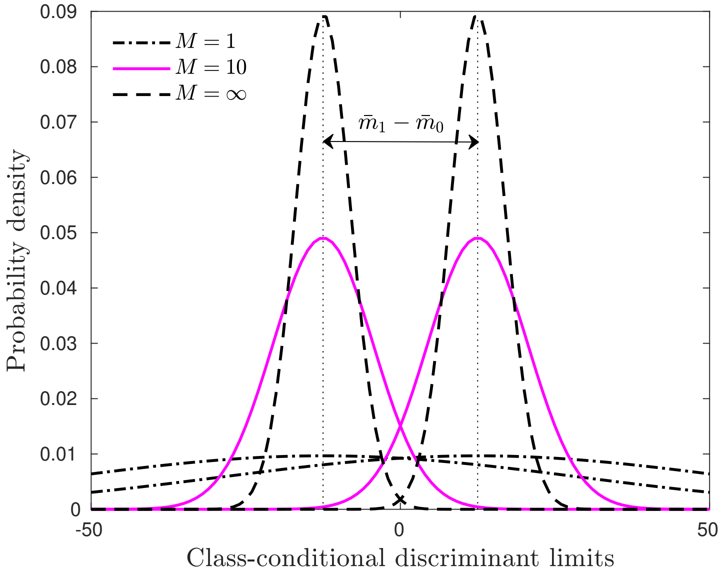

Figure 2: Class-conditional asymptotic distributions of the discriminant-averaging ensemble $M = 1$, $M = 10$, and $M = \infty$.

$2\Phi\left(-\frac{1}{2}\frac{\bar{m}_1(M)-\bar{m}_0(M)}{\sqrt{\bar{\sigma}(M)}}\right)$ (which is a decreasing sequence of M since $\bar{m}_1(M) - \bar{m}_0(M) > 0$ by Equation 15). This overlap is an indication of the ability of the discriminant to distinguish between a test point in $\mathcal{C}_0$ versus $\mathcal{C}_1$. Thus, a lower overlap implies greater discrimination and suggests a lower probability of misclassification of the classifier. To complement Figure 2, we plot the probability of misclassification DE, $\bar{\varepsilon}$ (based on the asymptotic distributions of Theorems 3 and 4), corresponding to a discriminant-averaging RP-LDA ensemble classifier as M is increased under a parameter setting where the Bayes error is about 0.05 in Figure 3. This plot shows that the asymptotic probability of misclassification is decreasing with ensemble size M. The explicit expression for $\bar{\varepsilon}$ is stated in Appendix C.1.

The final theorem in this section states that the asymptotic distribution of the vote-averaging RP-LDA ensemble discriminant is a normalized correlated binomial random variable with constant correlation between the trials. To see this, we re-write the decision rule as

$$\mathbb{1}\left\{\frac{1}{M}MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) > 0.5\right\}, \tag{13}$$

wherein the discriminant in (9) is multiplied and divided by the factor M. The following theorem formally states the distribution of the $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)$ part of the discriminant as it is expressed in (13).

Theorem 7 *(Vote-averaging RP-LDA ensemble discriminant asymptotic distribution) Under conditions* (a) *to* (g)*, for M times the vote-averaging RP-LDA finite ensemble discrim-*

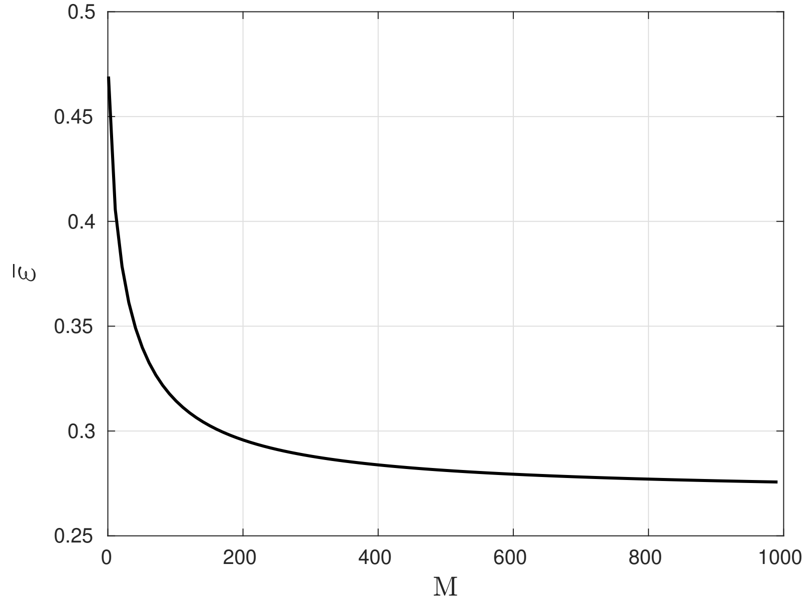

Figure 3: Plot of the probability of misclassification DE, $\bar{\varepsilon}$, of the discriminant-averaging RP-LDA ensemble classifier against increasing ensemble size M.

inant, $W_{\text{vote-avg}}\left(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M\right)$*, defined in* (9)*, fixed* M*, and* $i = 0, 1$*, we have*

$$P\left\{(MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i) > t\right\} - P\left\{\mathcal{CB}\left(M, \bar{p}_i, \bar{\rho}_i\right) > t\right\} \to 0, \;\forall t \in \mathbb{R},$$

where M *is the number of trials,*

$$\bar{p}_i = \Phi\left(\frac{\bar{m}_i}{\sqrt{\bar{\sigma}^2(1)}}\right)$$

is the asymptotic probability of success in each trial, and

$$\bar{\rho}_i = \frac{\mathcal{I}_i - \bar{p}_i^2}{\bar{p}_i\left(1 - \bar{p}_i\right)}$$

is the asymptotic correlation between each trial, where

$$\mathcal{I}_i := \int_0^\infty \int_0^\infty \frac{1}{2\pi\lambda} \exp\left(-\frac{1}{2\lambda^2}\left[\sum_{j=1}^{2}\left(\alpha_j^i\right)^2 - 2\bar{\sigma}^2_{M=\infty}\alpha_1^i\alpha_2^i\right]\right) d\alpha_1 d\alpha_2,$$

$\lambda^2 := \left(\bar{\sigma}^2(1)\right)^2 - \left(\bar{\sigma}^2_{M=\infty}\right)^2$, $\alpha_1^i := \alpha_1 - \bar{m}_i$*, and* $\alpha_2^i := \alpha_2 - \bar{m}_i$*.*

Proof See Appendix D.5. ■

From Theorem 7, one infers that $W_{\text{vote-avg}}\left(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M\right)|\mathbf{x} \in \mathcal{C}_i$ is asymptotically a correlated binomial normalized by the factor $1/M$. It is straightforward to show that the mean of this asymptotic distribution is

$$\bar{p}_i,\ i = 0, 1,$$

and its variance is

$$\frac{1}{M}\bar{p}_i\left(1 - \bar{p}_i\right) + \left(1 - \frac{1}{M}\right)\bar{\rho}_i\bar{p}_i\left(1 - \bar{p}_i\right),\ i = 0, 1.$$

Therefore, the mean separation between the asymptotic distributions is a constant $\bar{p}_1 - \bar{p}_0$ regardless of the ensemble size M and, since $\bar{\rho}_i \in [0, 1]$, each of their respective variances decrease from $\bar{p}_i\left(1 - \bar{p}_i\right)$ at $M = 1$ (a Bernoulli variance corresponding to single trial) to $\bar{\rho}_i\bar{p}_i\left(1 - \bar{p}_i\right)$ as M tends to infinity. Thus the asymptotic class-conditional discriminants of the vote-averaging ensemble exhibit similar behavior to the asymptotic class-conditional discriminants of the discriminant-averaging ensemble; the mean separation between the distributions stays constant with M while the variances decrease.

A direct comparison between the mean separations and variances of the asymptotic distributions corresponding to each classifier is quite difficult. We take an alternative approach to comparing the two schemes in the next section by showing—via the Neyman-Pearson lemma—that among **all** RP-LDA discriminant combining schemes, the equally-weighted discriminant-averaging RP-LDA ensemble is asymptotically optimal for Gaussian data. Further detail regarding the form of the asymptotic PMF of the vote-averaging RP-LDA ensemble classifier can be found in Appendix D.5 and Appendix C.3.

3.2 Asymptotically Optimal Ensemble of Randomly-Projected LDA Discriminants

By employing individual randomly-projected LDA discriminants as observations, this section constructs asymptotically optimal ensembles in terms of the Receiver Operating Characteristic (ROC) and the probability of misclassification via the Neyman-Pearson lemma and the MAP rule, respectively. These results rely on knowledge of the asymptotic joint PDF of a collection of single RP-LDA discriminants. Letting the vector of M randomly-projected LDA discriminants be denoted by $\mathbf{W} = \left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_1\right), \ldots, W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_M\right)\right]^T$, the asymptotic PDF of $\mathbf{W}|\mathbf{x} \in \mathcal{C}_i,\ i = 0, 1$, is presented in the following theorem.

Theorem 8 *(Asymptotic joint distribution of M RP-LDA discriminants) Under conditions* (a) *to* (g)*, for fixed M, the vector* $\mathbf{W}$ *conditioned on the class of the test point is such that*

$$P\left\{\left(\mathbf{W}|\boldsymbol{x} \in \mathcal{C}_i\right) > \boldsymbol{t}\right\} - P\left\{\mathcal{N}\left(\bar{\boldsymbol{\zeta}}_i, \bar{\mathbf{\Pi}}\right) > \boldsymbol{t}\right\} \to 0,\ \forall \boldsymbol{t} \in \mathbb{R}^M,\ i = 0, 1,$$

where $\bar{\boldsymbol{\zeta}}_i = \bar{m}_i\mathbf{1}_M$ *and* $\bar{\mathbf{\Pi}} = \left(\bar{\sigma}^2(1) - \bar{\sigma}^2_{M=\infty}\right)\mathbf{I}_M + \bar{\sigma}^2_{M=\infty}\mathbf{1}_M\mathbf{1}_M^T$*, that is,* $\mathbf{W}|\boldsymbol{x} \in \mathcal{C}_i$ *converges in distribution to a Gaussian random vector with expectation* $\bar{\boldsymbol{\zeta}}_i$ *and covariance* $\bar{\mathbf{\Pi}}$*.*

Proof See Appendix D.1. ■

Now, let us reconsider the classification problem in the context of hypothesis testing. Consider the null hypothesis $\mathbf{x}$ belongs to $\mathcal{C}_0$ and the alternative hypothesis $\mathbf{x}$ belongs to $\mathcal{C}_1$. For any classifier, let α be the probability of a false positive, that is, classifying the test point to $\mathcal{C}_1$ while it actually belongs to $\mathcal{C}_0$, and β the probability of false negative, that is, classifying the test point to $\mathcal{C}_0$ while it actually belongs to $\mathcal{C}_1$. The most powerful α-level test is, by definition, the test which minimizes β or, equivalently, maximizes the probability of a true positive, $1-\beta$, at a fixed α.

Based on the asymptotic PDF of Theorem 8, the asymptotically most powerful α-level test is as follows.

Theorem 9 *(Neyman-Pearson RP-LDA ensemble classifier based on the asymptotic joint distribution of a set of RP-LDA discriminants) The asymptotically most powerful α-level test is to classify $\boldsymbol{x}$ to $\mathcal{C}_1$ if*

$$W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) > \eta$$

and to $\mathcal{C}_0$ otherwise, where η is such that

$$P\left\{(W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\boldsymbol{x} \in \mathcal{C}_0) > \eta\right\} = \alpha.$$

Proof See Appendix D.2. ■

The above result shows that the classifier yielding the optimal ROC is in fact the equally-weighted discriminant-averaging RP-LDA ensemble which assigns equal weights of $1/M$ to each of the projections $\mathbf{R}_1, \ldots, \mathbf{R}_M$. This means that for classification purposes, in the context of an ensemble, the projections are asymptotically identical. Non-uniform weights, including binary weights, lead to asymptotically sub-optimal classification in this data setting. Note also that this classifier is linear in the test point.

Using the asymptotic PDF of $\mathbf{W}$, we are also able to derive the asymptotically Bayes combination of RP-LDA discriminants which minimizes the probability of misclassification. It is presented in the following theorem.

Theorem 10 *(Asymptotic MAP RP-LDA ensemble classifier) The asymptotic MAP RP-LDA ensemble classifier classifies to $\mathcal{C}_1$ when*

$$\frac{\bar{m}}{\bar{\sigma}^2(M)}\left[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) - \frac{\bar{m}_0 + \bar{m}_1}{2}\right] + \ln\frac{\pi_1}{\pi_0} > 0, \tag{14}$$

and to $\mathcal{C}_0$ otherwise.

Proof The classifier which maximizes the posterior probability $P[\mathbf{x} \in \mathcal{C}_i|\mathbf{W}]$, minimizes the probability of misclassification (Hastie et al., 2009). Maximizing the posterior probability in the two-class scenario is equivalent to the following decision rule on the ratio of posterior probabilities

$$\frac{\pi_1 f[\mathbf{W}|\mathbf{x} \in \mathcal{C}_1]}{\pi_0 f[\mathbf{W}|\mathbf{x} \in \mathcal{C}_0]} > 0,$$

where $f(\cdot)$ denotes a PDF. By Theorem 8, this ratio tends asymptotically to (14). ■

Theorem 10 shows that a slight modification of (5) yields the asymptotically lowest probability of misclassification for an RP-LDA ensemble. Note that:

- Like the asymptotic Neyman-Pearson RP-LDA ensemble classifier of Theorem 9, the asymptotic MAP RP-LDA ensemble classifier is also linear in $\mathbf{x}$ and it is easy to show that $\frac{\bar{m}_1-\bar{m}_0}{\bar{\sigma}^2(M)} > 0$. As a result, its ROC matches that of the discriminant-averaging RP-LDA ensemble classifier (which is optimal according to Theorem 9).
- This classifier corresponds to a particular operating point on the ROC of the discriminant-averaging ensemble classifier.
- When $\pi_0 = \pi_1$, it is easy to show that the asympotic MAP RP-LDA ensemble classifier has exactly the same decision rule as the discriminant-averaging RP-LDA ensemble classifier.

For the sake of completeness, the error analyses of the discriminant-averaging, asymptotic MAP, and vote-averaging RP-LDA ensembles are detailed in Appendix C, wherein deterministic equivalents of the probability of misclassification are provided for each classifier.

3.2.1 Demonstrations on Synthetic Data

This section showcases the results of Theorems 9 and 10 on synthetic data. The main goal of the simulations in this section is to show that the discriminant-averaging RP-LDA discriminant combining scheme performs at least as good as a classifier (in terms of ROC and error rate) as the discriminant-averaging-with-projection-selection, vote-averaging, and vote-averaging-with-projection-selection RP-LDA discriminant combining schemes.

Recall from Section 2.3.2 that we consider projection selection schemes which perform the selection by generating B_1 disjoint groups of B_2 projections each and select the projection from each group which yields the lowest error rate according to an error estimator of choice. The final set of projections is used to build the ensemble composed of a total of B_1 projections. So that the comparison between selection and non-selection schemes is fair, we must set $B_1 \times B_2$ in the selection schemes to M. This is because selection can be viewed as assigning weights of zeros and ones to each projection for a given set of projections, while an equally-weighted scheme assigns each projection in the same set of projections a weight of $1/M$. The total number of weighted projections in both cases must be equal; otherwise, one of the methods has the advantage of a larger initial set of projections. In the following simulations, we consider discriminant-averaging and vote-averaging ensembles with $M = 200$ and discriminant-averaging and vote-averaging plus projection selection ensembles with $B_1 = 50$ and $B_2 = 4$, so that $B_1 \times B_2 = 200$. The projections are selected using the resubstitution estimate on the training set.

For the synthetic data simulations, the data follows the Gaussian mixture model specified by (1) with

$$\boldsymbol{\mu}_0 = \frac{1}{p^{1/4}} \left[\mathbf{1}^T_{\lceil \sqrt{p} \rceil} \; \mathbf{0}^T_{p-\lceil \sqrt{p} \rceil -2} \; 2 \; 2 \right]^T, \; \boldsymbol{\mu}_1 = \mathbf{0}_p, \text{ and } \boldsymbol{\Sigma} = \frac{10}{p} \mathbf{1}_p \mathbf{1}_p^T + 0.1 \mathbf{I}_p,$$

where $\lceil x \rceil$ denotes ceil(x). The Bayes error for this distribution is 0.0401. This value is computed using Equation (11) of the paper by Niyazi et al. (2022).

We generate 500 independent realizations of the training and test sets. The problem dimensions for each realization of the training set are $n = 100$, $p = 1000$, and $\pi_0 = \pi_1 = 0.5$.

Each realization of the testing set consists of 1000 data points. All data are generated in proportion to the prior probabilities.

Figure 4 shows the ROCs of each of the four classifiers averaged over the realizations of the training and test sets by fixing the x-axis for each realization and averaging over y-axis; the (averaged) True Positive Rate (TPR) is plotted against the False Positive Rate (FPR). Here, the projection dimension of all four classifiers is set to $d = 49$ (half the rank of the sample covariance estimate). The plot shows that discriminant averaging and vote averaging perform very similarly, with discriminant averaging being slightly better. Discriminant averaging with selection and vote averaging with selection are slightly worse than the equally-weighted schemes. The shading indicates the 99% confidence intervals, which are on the order of 10^{-3}, and so are barely visible. The Area Under the Curve (AUC) values corresponding to each classifier in Figure 4 are 0.8435, 0.8340, 0.8059, and 0.7843, respectively. Pairwise one-sided t-tests between the discriminant-averaging-without-selection mean AUC and each of the other classifiers are performed. For each experiment, the null hypothesis is that the discriminant-averaging-without-selection RP-LDA ensemble classifier mean AUC and the other classifier's mean AUC (over the training sets) are equal. The alternative hypothesis is that the discriminant-averaging-without-selection RP-LDA ensemble classifier mean AUC is greater than the other classifier's mean AUC. The pairwise one-sided t-tests reject the null hypothesis at the 1% significance level for every pair of classifiers. These results are consistent with Theorem 9, in the sense that they demonstrate that discriminant averaging is optimal in terms of ROC and outperforms all other methods, whether they involve selection or not.

Figure 5 plots the average testing errors of the four classifiers against the projection dimension d over 500 realizations of the training and test set along with 99% confidence intervals (shaded). As the class priors are assumed to be equal, the discriminant-averaging RP-LDA ensemble classifier is equivalent to the asymptotic MAP RP-LDA ensemble in this case. Again, discriminant averaging and vote averaging perform similarly, with discriminant averaging being slightly better, while the selection schemes are generally worse. This is consistent with Theorem 10, since discriminant-averaging is asymptotically the MAP classifier in this equal prior scenario.

3.2.2 Demonstrations on Real Data

This section showcases the results of Theorems 9 and 10 on real data. As in the previous section, we consider discriminant-averaging and vote-averaging RP-LDA ensemble classifiers with $M = 200$ and discriminant-averaging and vote-averaging plus projection selection RP-LDA ensemble classifiers with $B_1 = 50$ and $B_2 = 4$, so that $B_1 \times B_2 = 200$, where the projections are selected using the resubstitution estimate on the training set. Here, we do not report the confidence intervals on our results as there is no unbiased estimate of the variance of the k-fold cross-validation procedure (Bengio and Grandvalet, 2003) which we use to compute the ROCs and error rates for these data sets.

We consider the colon tumor gene microarray data set provided by Alon et al. (1999), the gastrointestinal lesion colonoscopy imaging data recorded under both white light and narrow band imaging provided by Mesejo et al. (2016), both the full and reduced dimension versions of the leukemia gene microarray data set provided by Golub et al. (1999),

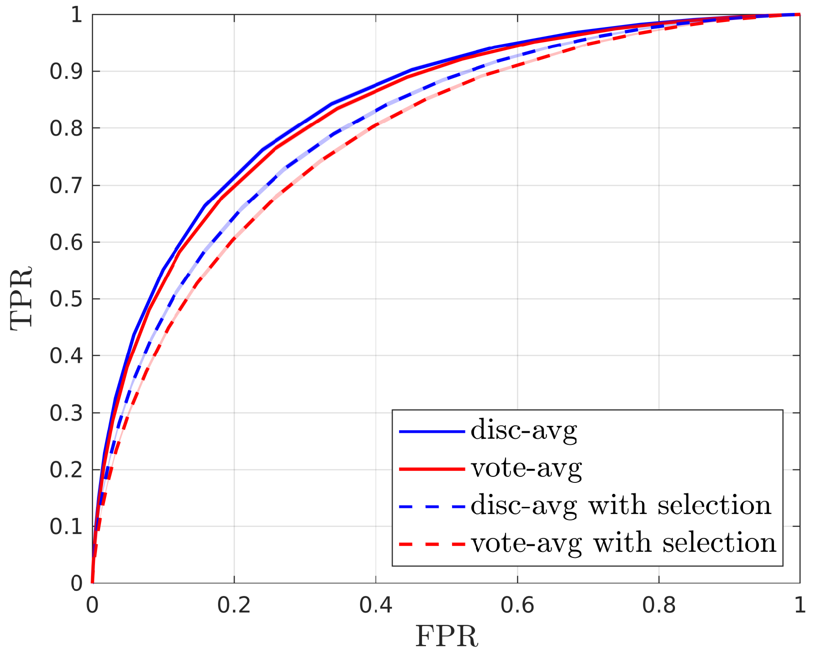

Figure 4: Average ROCs of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on Gaussian mixture model data where the shading indicates 99% confidence intervals.

both full and reduced dimension versions of the prostate cancer gene microarray data set provided by Singh et al. (2002), and 'aa' and 'ao' phoneme pairs from the data set provided by Hastie et al. (1995). These data sets are referred to as 'colon', 'gastro_WL', 'gastro_NB', 'leukemia_big', 'leukemia_small', 'prostate_full', 'prostate', and 'phoneme_aa_ao' respectively. The only preprocessing done to this data consisted of removing zero-variance predictors from the gastrointestinal lesion data sets. The number of training samples, dimensionality, and proportion of data points belonging to the majority class are listed in Table 1. Note that 'phoneme_aa_ao' actually consists of $n = 1717$ training samples, but to mimic a small sample situation where $p > n$, we randomly select a set of $n = 100$ samples for training and utilize the remaining 1617 samples for testing. The proportion of the majority class reported in Table 1 for this data set is based on the full training set. The artificially-constructed training set is sampled according to these class proportions.

Since all data sets have a relatively small number of samples, the error rates are estimated using iterated 10-fold cross validation, that is, 10 iterations of the 10-fold cross-validation estimate are computed and averaged to obtain the final estimate, with the exception of 'phoneme_aa_ao', for which we have a test set. To avoid cross-contamination between the data used for projection selection and the data used for performance evaluation, the resubstitution error computation for projection selection is nested within the cross-validation loop (as opposed to preceding the loop) so that the selection is performed using the training folds

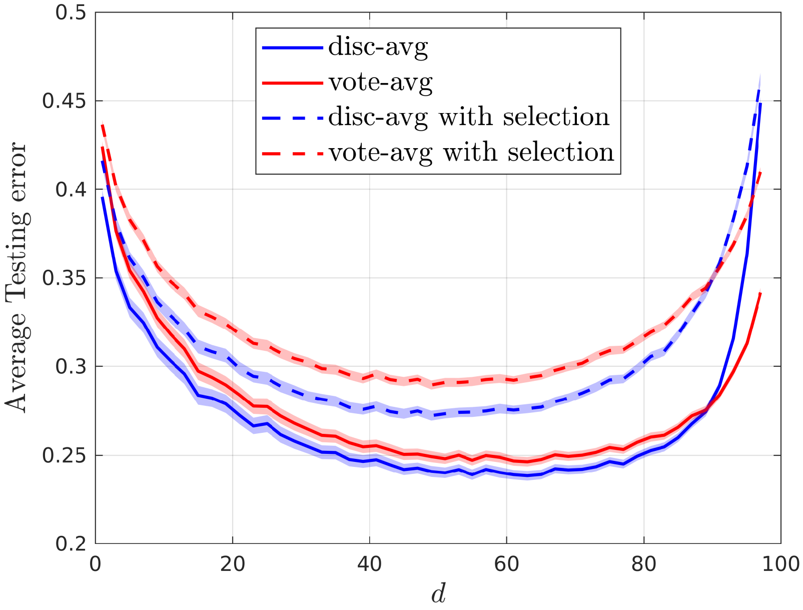

Figure 5: Average testing error of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on Gaussian mixture model data where the shading indicates 99% confidence intervals.

Data set	n	p	**Proportion of majority class**
'colon'	62	2000	0.65
'gastro_WL'	76	689	0.72
'gastro_NB'	76	689	0.72
'leukemia_small'	72	3571	0.65
'leukemia_big'	72	7128	0.65
'prostate'	102	2135	0.51
'prostate_full'	102	6032	0.51
'phoneme_aa_ao'	100	256	0.60

Table 1: Data sets and their properties

of the cross-validation procedure at each iteration, and not on the whole training set. The classifier is then evaluated on the testing fold. This is similar to the nested cross-validation procedure described by Cawley and Talbot (2010).

The AUCs corresponding to the ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers applied to each of the data sets in Table 1 are reported in Table 2. Each of the 'prostate_full' and 'phoneme_aa_ao' data sets reflect the findings of Theorem 9 in that discriminant averaging does better than all other schemes, including discriminant averaging with selection. Interestingly, vote averaging does better than vote averaging with selection on these data sets as well. On the ‘colon’ and ‘leukemia_small’ data sets, the AUCs corresponding to all four classifiers are virtually the same. For the data sets ‘gastro_WL’, ‘gastro_NB’, ‘leukemia_big’, and ‘prostate’, we have performances which are inconsistent with Theorem 9. More specifically, on these data sets, discriminant averaging performs similarly to discriminant averaging with selection, vote averaging performs similarly to vote averaging with selection, but vote averaging performs better than discriminant averaging. Such a discrepancy is not surprising as these data sets are not necessarily Gaussian and do not necessarily meet the common covariance assumption in (1). Nevertheless, looking closer at the ROCs corresponding to these data sets plotted in Figures 6-8, we observe that, within the range of practical TPR and FPR, discriminant averaging performs close to, or even better than, the remaining schemes.

Data set	**AUC**			
	disc-avg	**disc-avg with sel.**	**vote-avg**	**vote-avg with sel.**
‘colon’	0.853	0.856	0.850	0.854
‘gastro_WL’	0.813	0.812	0.834	0.833
‘gastro_NB’	0.75	0.74	0.762	0.762
‘leukemia_small’	0.9960	0.9963	0.9960	0.9957
‘leukemia_big’	0.9911	0.9900	0.9929	0.9931
‘prostate’	0.9493	0.9518	0.958	0.961
‘prostate_full’	0.787	0.765	0.770	0.74
‘phoneme_aa_ao’	0.8478	0.8409	0.8474	0.8358

Table 2: AUCs corresponding to the ROCs of the discriminant-averaging, discriminant-averaging-with-selection, vote-averaging, and vote-averaging-with-selection RP-LDA ensemble classifiers applied to real data.

Figures 9 to 14 plot the iterated 10-fold CV estimates of the error rate of the discriminant-averaging, vote-averaging, and discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the real data sets listed in Table 1 against varying projection dimension d. As explained previously, the selection process based on the resubstitution error estimate is nested within the cross-validation in order to avoid bias due to using the same data to select the projections for and evaluate the performance of the selection classifiers. Because the ‘gastro_WL’ and ‘gastro_NB’ data sets are significantly imbalanced with 72% of the data points made up by the majority class, and error rate is not the metric of interest in such cases, these data sets are omitted in this set of simulations. The proportions of the remaining data sets are close to balanced, and so for that reason it is reasonable to assume that the equally-weighted discriminant-averaging RP-LDA ensemble classifier performs similarly to the asymptotic MAP classifier as per Theorem 10.

In all figures, it can be observed that discriminant averaging and vote averaging perform similarly, while discriminant averaging with selection and vote averaging with selection perform similarly. The selection schemes exhibit slightly lower errors than the equally-weighted schemes at smaller values of d. This is inconsistent with our expectation that discriminant averaging outperforms all other schemes, and may be explained by the fact that our RMT

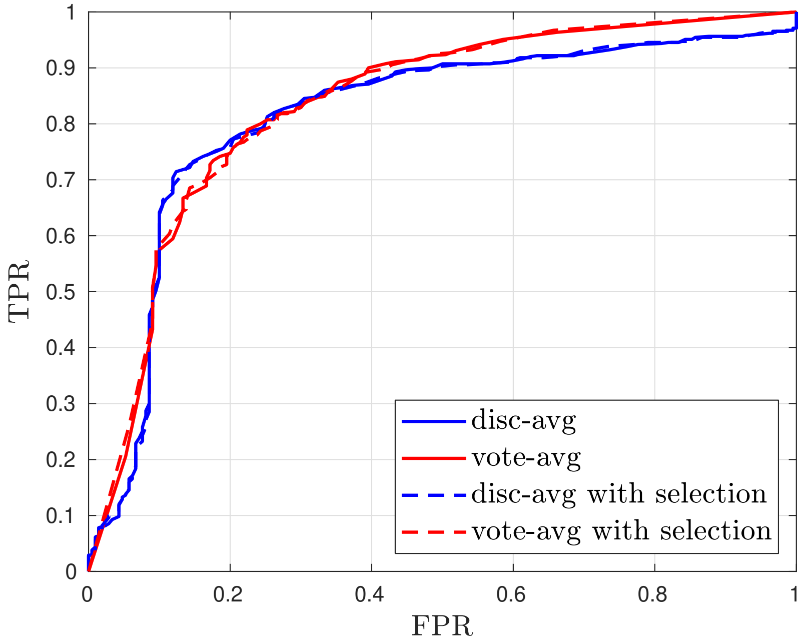

Figure 6: ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'gastro_WL' data set.

asymptotic analysis assumes d, n, and p to be in proportion to each other, whereas low values of d may constitute a different asymptotic regime. On the higher values of d, discriminant averaging and vote averaging outperform the selection schemes. In addition,the minimum error among all classifiers for each data set occurs within this range. The minimum error is achieved by vote averaging at $d = 25$ on the 'colon' data set, by discriminant averaging at $d = 27$ on the 'leukemia_big' data set, by discriminant averaging, vote averaging, and discriminant averaging with selection at $d = 43$ on the 'leukemia_small' data set, by discriminant averaging with selection at $d = 19$ on the 'prostate' data set, by vote averaging at $d = 17$ on the 'prostate_ full' data set, and by discriminant averaging at $d = 37$, vote averaging at $d = 39$ and vote averaging with selection at $d = 21$ on the 'phoneme_aa_ao' data set. Thus, it is reasonable to conclude, that on these data sets, selection generally gives no significant advantage over equally-weighted schemes, and that equally-weighted discriminant averaging, in particular, seems to perform as well as any other scheme.

To conclude, Section 3 derived asymptotic distributions of the discriminant-averaging and vote-averaging RP-LDA ensemble discriminants. It also proved that the optimal form of RP-LDA ensemble classifier under Gaussian data assumptions is the equally-weighted discriminant-averaging RP-LDA ensemble classifier. This finding is confirmed by simulations on synthetic data, as well as on real data,where it is shown that selection generally offers no additional performance advantage over equally-weighted schemes, and that discriminant averaging performs as good, if not better, than vote averaging on most data

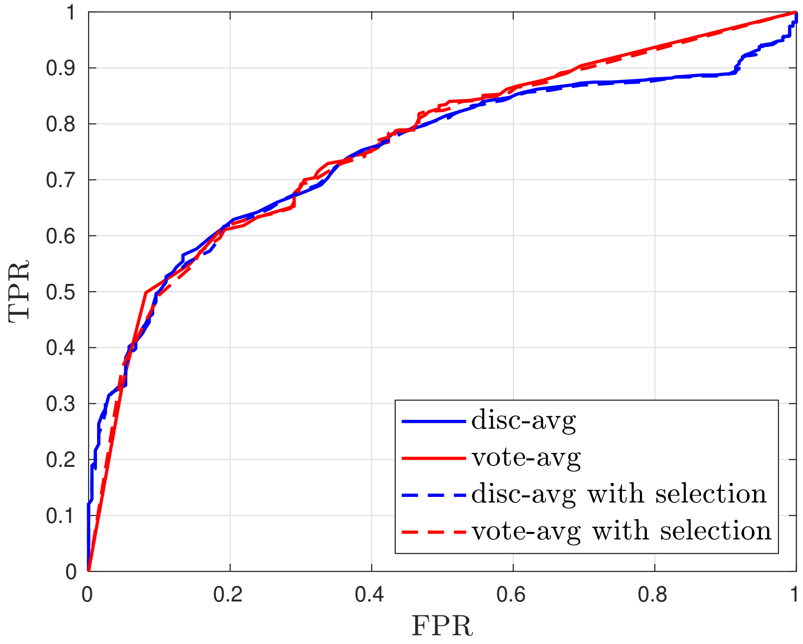

Figure 7: ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the ‘gastro_NB’ data set.

sets. Based on these findings, the next section studies the equally-weighted discriminant-averaging RP-LDA ensemble classifier from a practical perspective: choosing the number of projections M and tuning the projection dimension d through the use of G-estimators.

4. Turning Theory into Practice

In this section, we focus on practical implications of the analysis of the previous section in conjunction with G-estimators to propose a working framework for RP-LDA ensemble classification. The main lessons to take from the previous section are:

1. The optimal ensemble under Gaussian data assumptions is a linear function of the RP-LDA discriminants, that is, it is a form of discriminant averaging, as opposed to a non-linear scheme like vote averaging. As demonstrated in the previous section, both schemes perform very similarly on real data. Thus, there is no need to look beyond linear schemes.

2. Derivations under Gaussian data assumptions show that it is the number of projections which is critical for the classification performance of the discriminant-averaging RP-LDA ensemble, not the projections themselves. In fact, as evidenced by the previous section, projection selection under these assumptions may result in a performance loss. Furthermore, projection selection adds an extra cost in the form of computing error

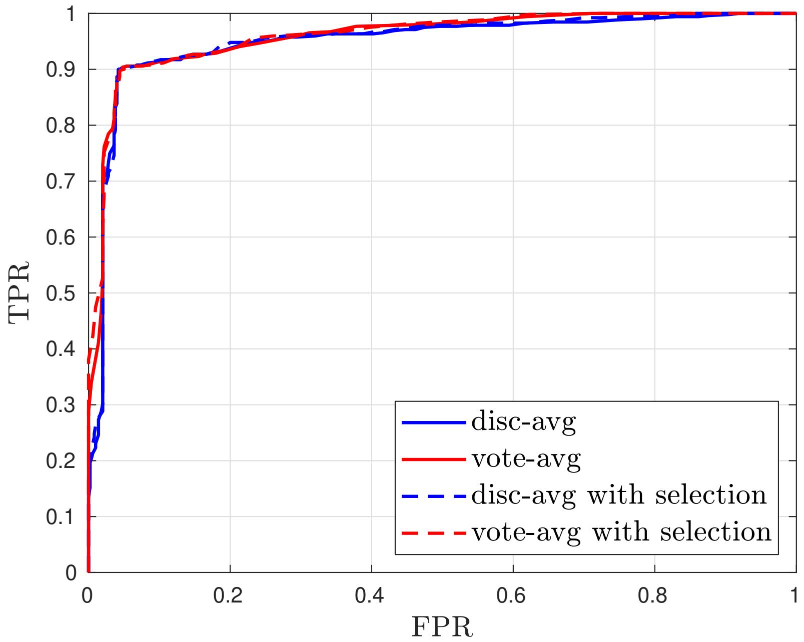

Figure 8: ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'prostate' data set.

estimators for each single RP-LDA ensemble classifier corresponding to a member of the set of projections in order to implement the selection process.

Based on these findings, this section proposes methods for the practical implementation of the equally-weighted discriminant-averaging RP-LDA ensemble classifier. We first present G-estimators of the most common classification metrics of this classifier. We then propose and demonstrate a method for tuning the number of projections M and projection dimension d on real and synthetic data.

4.1 G-estimators

A *G-estimator* of a quantity is an estimator of that quantity which is consistent in the RMT regime. This section provides G-estimators of the class-conditional discriminant statistics of the equally-weighted discriminant-averaging RP-LDA ensemble classifier, from which G-estimators of metrics such as the true positive/negative rate, false positive/negative rate, probability of misclassification, and positive/negative predictive value are constructed. This is detailed in what follows.

The main building blocks of the G-estimators of interest are G-estimators $\hat{m}_i,\ i = 0, 1$, $\hat{\sigma}^2(1)$, and $\hat{\sigma}^2_{M=\infty}$ such that

$$\hat{m}_i - \bar{m}_i \xrightarrow{a.s.} 0,\ i = 0, 1,$$

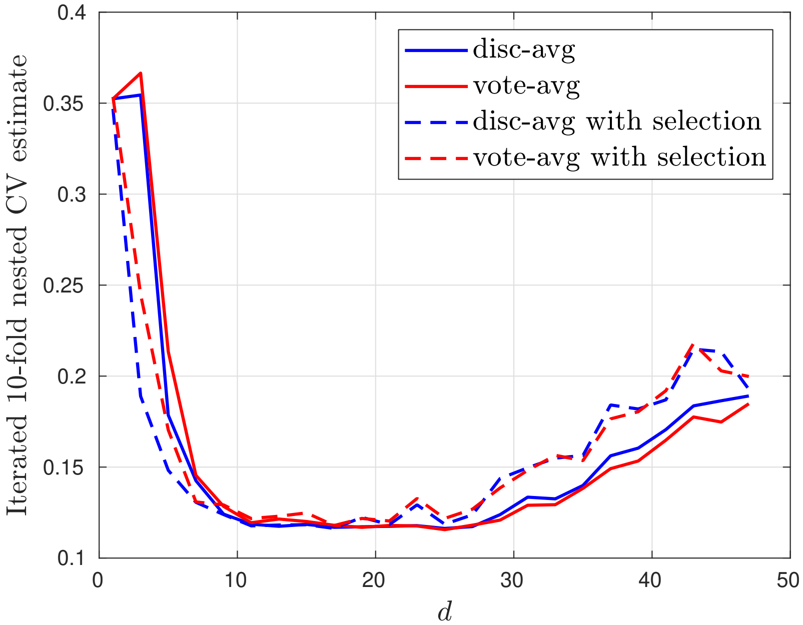

Figure 9: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'colon' data set.

which implies $\hat{m}_i \asymp m_i(1)$, $\hat{m}_i \asymp m_i(M)$, and $\hat{m}_i \asymp m_i^{M=\infty}$, that is, $\hat{m}_i$ is a G-estimator of all three classifier class-conditional means,

$$\hat{\sigma}^2(1) - \sigma^2(1) \xrightarrow{a.s.} 0,$$

and

$$\hat{\sigma}^2_{M=\infty} - \sigma^2_{M=\infty} \xrightarrow{a.s.} 0.$$

Theorem 11 presents the explicit expressions for these G-estimators. As in Section 3.1, these results are derived under conditions (a) to (g), and assume that M is fixed relative to the problem dimensions, p, n, and d, except when $M = \infty$ is specified, in which case M is allowed to diverge beforehand so that the results apply to the converged ensemble. Note that the G-estimators $\hat{m}_i,\ i = 0, 1$, and $\hat{\sigma}^2_{M=\infty}$ were first derived by Niyazi et al. (2020a). They are restated here for completeness.

Theorem 11 *(G-estimators of the class-conditional discriminant statistics of the discriminant-averaging RP-LDA ensemble classifier) Under conditions* (a) *to* (g)*, the G-estimators $\hat{m}_i,\ i = 0, 1$, $\hat{\sigma}^2(1)$, and $\hat{\sigma}^2_{M=\infty}$ are given by*

$$\hat{m}_i := (-1)^{i+1}\left[\frac{1}{2}\hat{\boldsymbol{\mu}}^T\left(\hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\mu}} - \frac{\frac{1}{n_i-1}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}{1 - \frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}\right] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0},\ i = 0, 1,$$

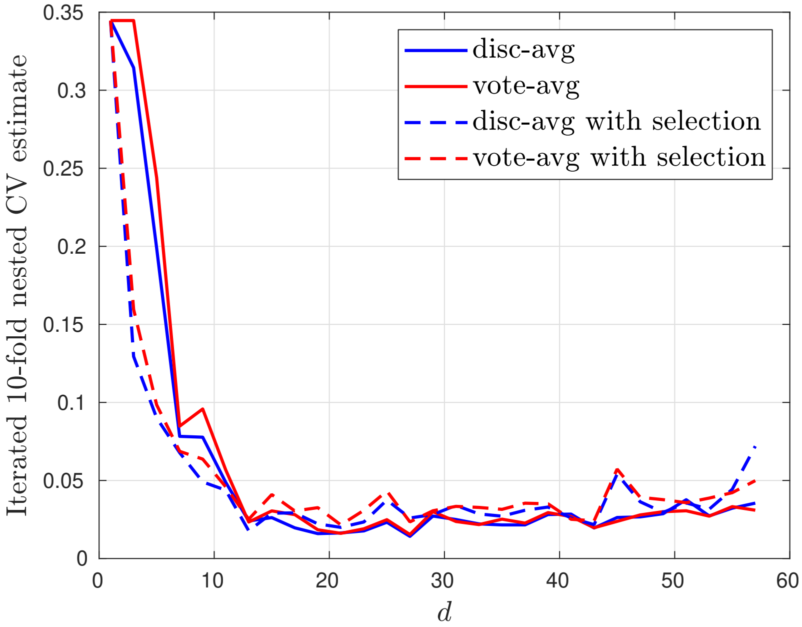

Figure 10: Iterated 10-fold CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'leukemia_big' data set.

$$\hat{\sigma}^2(1) := \left(\frac{1}{1-\frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}\right)^2 \hat{\boldsymbol{\mu}}^T\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\mu}}$$

$$+\frac{1}{\hat{\nu}^2}\frac{\left(\frac{1}{1-\frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}\right)^2\frac{1}{p}\mathrm{tr}\left\{\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}{1-\frac{1}{d}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}\hat{\boldsymbol{\mu}}^T\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-2}\hat{\boldsymbol{\mu}},$$

and

$$\hat{\sigma}^2_{M=\infty} := \left(1+\frac{\frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}{1-\frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}}\right)^2 \hat{\boldsymbol{\mu}}^T\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\hat{\boldsymbol{\mu}},$$

where $\hat{\nu}$ is such that

$$1-\frac{1}{d}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\boldsymbol{I}_p\right)^{-1}\right\}=0.$$

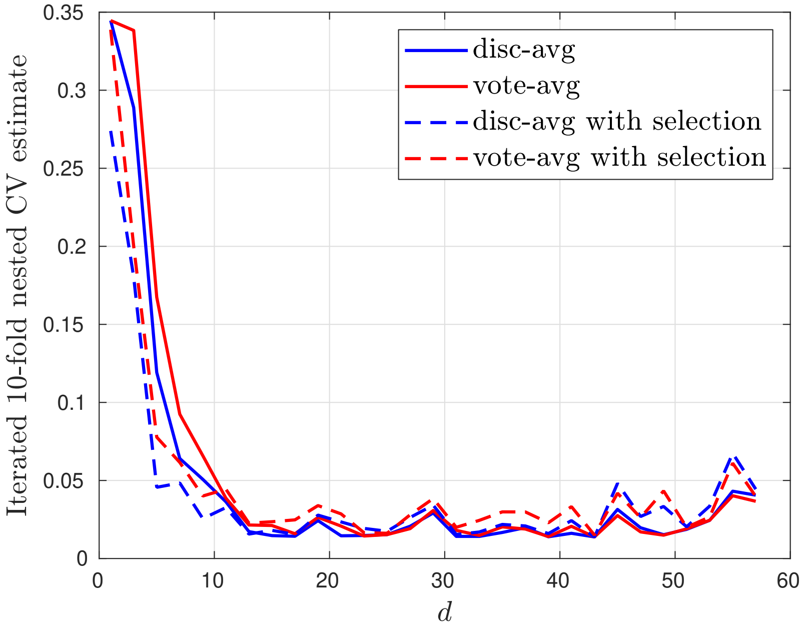

Figure 11: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'leukemia_small' data set.

Proof See Appendix E.1. ■

In addition, it can be shown that the G-estimator $\hat{\sigma}^2(M)$ of $\sigma^2(M)$, for fixed M, such that

$$\hat{\sigma}^2(M) - \sigma^2(M) \xrightarrow{a.s.} 0,$$

is simply

$$\hat{\sigma}^2(M) = \frac{1}{M}\hat{\sigma}^2(1) + \left(1 - \frac{1}{M}\right)\hat{\sigma}^2_{M=\infty}.$$

The next theorem presents the G-estimators of some common binary classification metrics of the equally-weighted discriminant-averaging RP-LDA ensemble in terms of the preceding G-estimators. Note that we take $\mathcal{C}_0$ to be the negative class and $\mathcal{C}_1$ to be the positive class.

Theorem 12 *(G-estimators of some common classification metrics of the discriminant-averaging RP-LDA ensemble classifier)*

- *TPR:* $\hat{TPR} := \Phi\left(\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)$

- *TNR:* $\hat{TNR} := \Phi\left(-\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2(M)}}\right)$

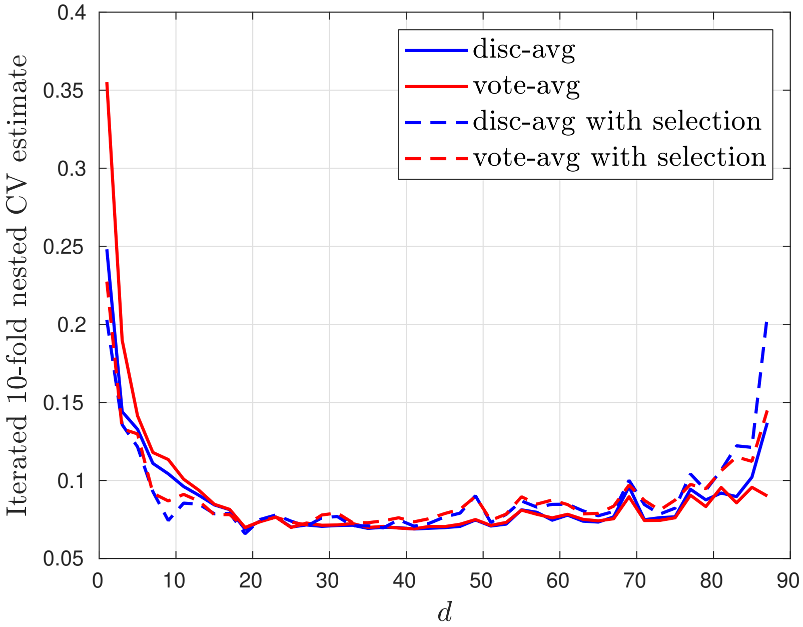

Figure 12: Iterated 10-fold nested CV estimate of the error rate of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'prostate' data set.

- *FPR:* $\hat{FPR} := \Phi\left(\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2(M)}}\right)$
- *FNR:* $\hat{FNR} := \Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)$
- *Probability of misclassification:* $\hat{\varepsilon} := \hat{\pi}_0\hat{FPR} + \hat{\pi}_1\hat{FNR}$
- *PPV:* $\hat{PPV} := \frac{\hat{\pi}_1 \hat{TPR}}{\hat{\pi}_0 \hat{FPR} + \hat{\pi}_1 \hat{TPR}}$
- *NPV:* $\hat{NPV} := \frac{\hat{\pi}_0 \hat{TNR}}{\hat{\pi}_0 \hat{TNR} + \hat{\pi}_1 \hat{FNR}}$

Proof See Appendix E.2. ■

In the next section, we propose a general procedure for tuning the parameters of the discriminant-averaging RP-LDA ensemble classifier. We also demonstrate how G-estimators may be made use of in this context.

4.2 Tuning the Discriminant-Averaging RP-LDA Ensemble Parameters

This section maps out a procedure for tuning the number of projections M and projection dimension d of the equally-weighted discriminant-averaging RP-LDA ensemble classifier.

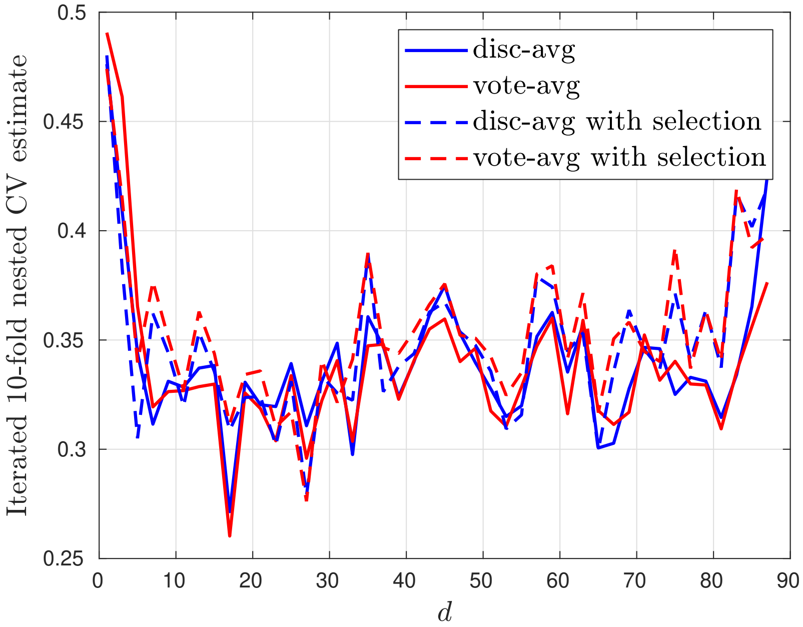

Figure 13: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'prostate_full' data set.

To begin with, note that, as shown in Section 3, for a given d, performance improves with increasing M so that the upper bound on the performance of the finite version of the discriminant-averaging RP-LDA ensemble classifier is the performance of the infinite version of the discriminant-averaging RP-LDA ensemble classifier. In other words, the ratio of infinite ensemble error to finite ensemble error is always less than or equal to one.

As M is constrained by computational efficiency, one may specify the trade-off between performance and computational efficiency as a fraction of the infinite ensemble performance, denoted by ψ. More specifically, $\psi = \frac{\text{infinite ensemble error}}{\text{finite ensemble error}}$. This idea is illustrated in Figure 15, which shows this ratio approaching 1 with increasing M. As alluded to earlier, the infinite ensemble cannot be realized practically, but must be approximated by a large number of projections. Figure 15 uses 3000 projections to approximate the infinite ensemble. As indicated on the figure, a performance of $\psi = 0.98$ is achieved at $M = 112$. Setting $M = 112$ results in significant computational savings as compared to the full set of 3000 projections which approximate the infinite ensemble. Based on this, we propose the following experimental approach to tuning M and d, which uses 5000 projections to approximate the infinite ensemble. It is important to realize that this procedure is part of the classifier training, and should be applied to the training set. The two stages of this procedure are:

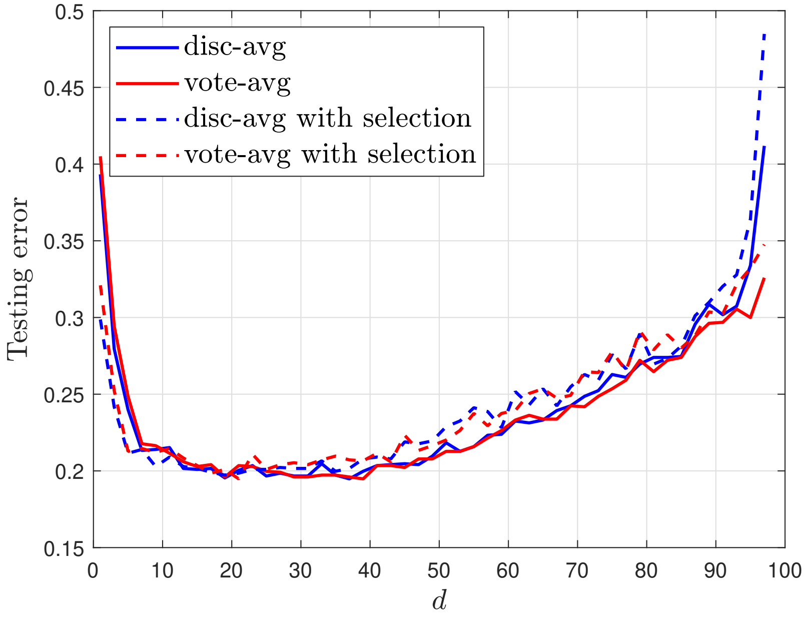

Figure 14: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection RP-LDA ensemble classifiers on the 'phoneme_aa_ao' data set.

1. Tune d for an ensemble with $M = 5000$. This approximates the optimal projection dimension for an infinite ensemble. Compute the corresponding error, which serves as the "infinite" ensemble performance benchmark.

2. Now starting at a small M (for example, $M = 100$), compute the error and the resulting ratio of the "infinite" ensemble error (from step 1) to this finite ensemble error. Check if the preset ψ is achieved. If not, increment M. Repeat in this manner until a ratio of ψ is achieved. The value of M at which ψ is achieved is the final setting of M for the finite ensemble which achieves at least ψ level of performance relative to the "infinite" ensemble.

While it is possible to add a third stage to this procedure in which d is further tuned for the particular finite ensemble obtained in the second stage, we find that this can result in a performance loss in practice, probably due to overfitting to the training data.

Algorithm 1 presents the previously outlined procedure in more detail. Here, $R(M, d)$ is any training-data-based estimate of the probability of misclassification of a discriminant-averaging RP-LDA ensemble classifier composed of M projections each having a projection dimension of d. Of course, Algorithm 1, especially lines 1-3, may be very computationally intensive depending on the choice of error estimator $R(M, d)$. The G-estimator of the infinite ensemble error derived by Niyazi et al. (2020a) and the G-estimator of the finite ensemble error derived in this work (see Section 4.1, Theorem 12) can significantly reduce

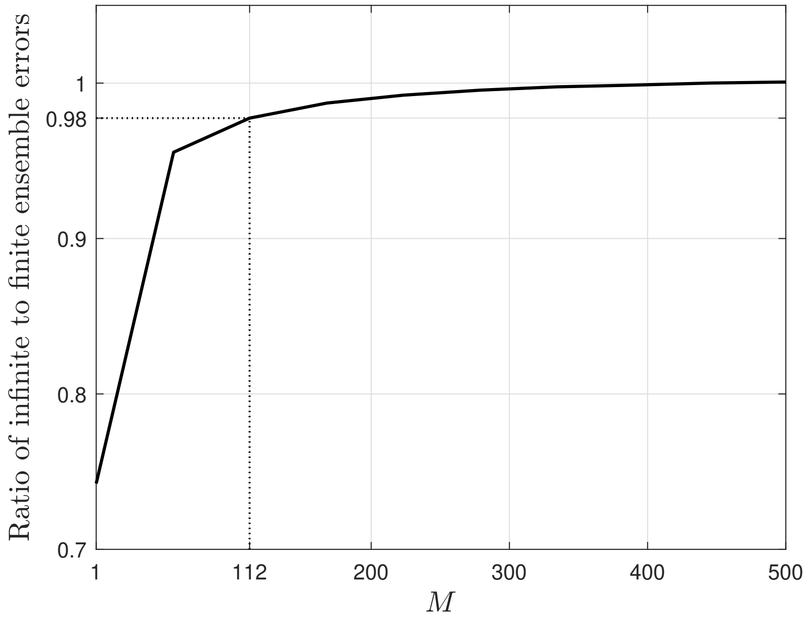

Figure 15: Ratio of "infinite" to finite discriminant-averaging RP-LDA ensemble classifier error on Gaussian mixture model data. A ratio of 0.98 is achieved at $M = 112$.

Algorithm 1 Tuning the discriminant-averaging RP-LDA ensemble parameters M and d

```
Require: $\psi < 1$
 1: for $d' = 1 : \mathrm{rank}\{\hat{\mathbf{\Sigma}}\}$ do
 2:     Compute $R\,(5000, d')$          ▷ Set $M = 5000$ to approximate an infinite ensemble
 3: end for
 4: $d \leftarrow \mathrm{argmin}_{d'}\, R\,(5000, d')$          ▷ Tune $d$ for "infinite" ensemble
 5: $R_{M=\infty} \leftarrow R(5000, d)$          ▷ Set minimum "infinite" ensemble error estimate
 6: $M' \leftarrow 100$
 7: Compute $R(M', d)$
 8: while $\frac{R_{M=\infty}}{R(M',d)} < \psi$ do          ▷ Set $M$ so that $\psi$ is satisfied
 9:     $M' \leftarrow M' + 100$
10:     Compute $R(M', d)$
11: end while
12: $M \leftarrow M'$
```

computational costs. We propose using the former for the calculation in line 2, and the latter for the calculations in lines 7 and 10. For further savings, the M tuning procedure in lines 6-12 may be bypassed through the following formula based on an approximation of the probability of misclassification using G-estimators:

$$M \approx \text{ceil}\left(\frac{\left(\hat{\sigma}^2(1) - \hat{\sigma}^2_{M=\infty}\right) W_0\left(\psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}} \exp\left(\frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}}\right)\right)}{\hat{m}_1^2 - \hat{\sigma}^2_{M=\infty} W_0\left(\psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}} \exp\left(\frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}}\right)\right)}\right),$$

where $W_0(\cdot)$ is the principal branch of the Lambert W function. This approximation is valid when the class priors are **equal**. It is derived in Appendix E.3. In what follows, we refer to this approximation as 'the heuristic'.

We now report the errors achieved by tuning the discriminant-averaging RP-LDA ensemble classifier by Algorithm 1 on both real and synthetic data. As mentioned previously, R-LDA presents an upper limit to the performance of the discriminant-averaging RP-LDA ensemble classifier; however, R-LDA works with the full data dimensions, making it more computationally-demanding than the discriminant-averaging RP-LDA ensemble classifier. The objective of the following simulations is to show that the performance of R-LDA can be approached by the discriminant-averaging RP-LDA ensemble classifier at a lower computational complexity through the proposed tuning procedure. For this reason, only R-LDA and the discriminant-averaging RP-LDA ensemble classifier (with various ways of tuning its parameters) are considered in this set of simulations. The reader is referred to the work by Durrant and Kabán (2015) for a comprehensive comparison of the discriminant-averaging RP-LDA ensemble classifier with the state of the art.

Remark 13 *We also report the complexities of R-LDA and the discriminant-averaging RP-LDA ensemble classifier at execution. They are computed as* $\mathcal{O}\left(p^3\right)$ *and* $O(M(np + d^3))$ *(Durrant and Kabán, 2015), respectively.*

We first consider synthetic data generated from the Gaussian mixture model specified at the beginning of Section 3.2.1. The testing error is evaluated on a testing set consisting of 10^5 data points from each class. All test errors are rounded to three decimal places. Table 3 presents the testing errors and parameter settings of the discriminant-averaging RP-LDA "infinite" ensemble classifier, R-LDA, and various tunings of the discriminant-averaging RP-LDA finite ensemble classifier for a given training and test set.

The R-LDA precision matrix estimator takes the form $(\hat{\mathbf{\Sigma}} + \gamma \mathbf{I}_p)^{-1}$, where γ is a positive scalar. The parameter setting of the "infinite" ensemble is determined by setting $M = 5000$ and tuning d optimally according to the iterated 10-fold CV estimate of error on the training set. Similarly, the parameter setting of R-LDA is determined by tuning γ optimally (over the interval 10^{-6} to 5 in increments of 0.01) according to the iterated 10-fold CV estimate of error on the training set. The parameter settings of the finite ensemble are determined by different variants of Algorithm 1 at $\psi = 0.95$. The first sub-row under the finite ensemble uses the iterated 10-fold CV on the training set as the error estimator $R(M, d)$. The second sub-row uses the G-estimator for the infinite and finite ensembles on the training set in place of all error estimators in Algorithm 1. The third sub-row uses the heuristic described previously to compute M directly and the G-estimators for the infinite ensemble in the second step. Finally, the fourth sub-row uses Durrant's rule-of-thumb (see Section 2) to set M and d without any need for error estimation. To reduce fluctuations due to the random projections, the 'Test error' values reported in the table are testing errors averaged over

500 sets of projections, except for the first row which is averaged over 5 sets, since each set consists of 5000 projections. The 99% confidence interval for each of these averages is also calculated and found to be on the order of 10^{-4} for all ensembles except the ensemble tuned by Durrant's rule of thumb for which it is on the order of 10^{-3}. Additionally, the tuning times (in seconds) for each of the classifiers in Table 3 is reported in Table 4.

Table 3 shows that R-LDA achieves the lowest testing error of 0.214 followed by the finite ensemble tuned by heuristic at 0.218. On the other hand, according to Table 4, R-LDA takes the longest time to tune (about 12 minutes) out of all classifiers. Another consideration is complexity at execution, that is, when the classifier is used to classify a test point. As the computational complexity of R-LDA at execution is $\mathcal{O}\left(p^3\right)$, while that of the discriminant-averaging RP-LDA ensemble classifier is $O(M(np + d^3))$ (Durrant and Kabán, 2015), using a large M may negate the computational savings gained by projection, in which case it is better to use R-LDA. For Table 3 in particular, the complexity of R-LDA is on the order of 10^9 operations, while that of the ensemble tuned by the heuristic is on the order of 10^8 operations. The heuristic also has the advantage of minimal tuning at training time as reflected in Table 4. The worst performance goes to Durrant's rule-of-thumb which yields an error of 0.245. Its complexity at execution is, however, on the order of 10^7 and, moreover, has the fastest tuning time at a fraction of a second (since all that is required is to compute the sample covariance matrix rank). The finite ensemble tuned by cross-validation also has a relatively high test error of 0.241 at a corresponding complexity at execution on the order of 10^7, although the training procedure is much more involved (around 8 minutes). The "infinite" ensemble tuned by cross-validation and the finite ensemble tuned by G-estimators yield test errors of 0.224 and 0.220, respectively, at a common complexity at execution on the order of 10^8; they exhibit higher errors in addition to offering no computational advantages over the finite ensemble tuned by the heuristic.

Classifier		**Test error**	**Parameters**
disc-avg "infinite" ensemble tuned by CV		0.224	$d = 31$
R-LDA		0.214	$\gamma = 1.03$
disc-avg finite ensemble tuned by:	CV	0.241	$M = 200$, $d = 31$
	G-estimators	0.220	$M = 400$, $d = 66$
	G-estimators + Heuristic	0.218	$M = 552$, $d = 66$
	Durrant's rule-of-thumb	0.245	$M = 100$, $d = 49$

Table 3: Table of average testing errors and parameter settings of the discriminant-averaging "infinite" ensemble classifier, where d is tuned based on cross-validation, and the discriminant-averaging RP-LDA finite ensemble classifier where M and d are tuned by Algorithm 1 based on cross-validation, G-estimators, and the heuristic on Gaussian mixture model data. Here $\psi = 0.95$. For comparison, the Bayes error is 0.0401.

Classifier		Tuning time (s)
R-LDA		700.654
disc-avg finite ensemble tuned by:	CV	498.517
	G-estimators	8.800
	G-estimator + Heuristic	0.608
	Durrant's rule-of-thumb	0.135

Table 4: Tuning time (in seconds) of each of the classifiers in Table 3.

For real data, we consider the 'phoneme_aa_ao' data set. Again, we use 5000 projections to approximate the infinite ensemble. Table 5 can be interpreted exactly as Table 3. Here $\psi = 0.99$ and, as in the synthetic data simulation, the reported testing errors are averaged over 500 trials except for the first row where the testing error is averaged over 5 trials. All test errors are rounded to three decimal places. The 99% confidence intervals of the averages are all on the order of 10^{-4}. Table 6 reports the tuning time (in seconds) for each of the classifiers in Table 5.

Classifier		Test error	Parameters
disc-avg "infinite" ensemble tuned by CV		0.206	$d = 11$
R-LDA		0.210	$\gamma = 4.76$
disc-avg finite ensemble tuned by:	CV	0.206	$M = 2500$, $d = 11$
	G-estimators	0.199	$M = 100$, $d = 31$
	G-estimators + Heuristic	0.199	$M = 82$, $d = 31$
	Durrant's rule-of-thumb	0.214	$M = 100$, $d = 50$

Table 5: Table of average testing errors and parameter settings of the discriminant-averaging RP-LDA "infinite" ensemble classifier, where d is tuned based on cross-validation, and the discriminant-averaging RP-LDA finite ensemble classifier where M and d are tuned by Algorithm 1 based on cross-validation, G-estimators, and the heuristic on the 'phoneme_aa_ao' data set. Here $\psi = 0.99$.

Table 5 shows that the two strategies of tuning a finite ensemble using the G-estimators and the heuristic achieve the lowest testing error of 0.199. Both have complexities at execution on the order of 10^6. Their tuning times, as report in Table 6 are fractions of a second. The fact that these finite ensembles perform better than the "infinite" ensemble tuned by cross-validation can be explained by the better choice of d obtained by the G-estimators ($d = 31$) as compared to cross-validation ($d = 11$). This is confirmed by computing the

Classifier		**Tuning time (s)**
R-LDA		76.010
disc-avg finite ensemble tuned by:	CV	277.467
	G-estimators	0.610
	G-estimator + Heuristic	0.073
	Durrant's rule-of-thumb	0.002

Table 6: Tuning time (in seconds) of each of the classifiers in Table 5.

testing error of an ensemble with $M = 5000$ and $d = 31$ which turns out to be 0.195. Durrant's rule-of-thumb has the worst performance on this data, yielding an error of 0.214 at a complexity (at execution time) on the order of 10^7 operations. It, however, has the least tuning time at 2 milliseconds. The "infinite" ensemble, R-LDA and the finite ensemble tuned by cross-validation have complexities at execution on the order of 10^8, 10^7, and 10^7, respectively, as well as relatively high tuning times, and so have higher errors than the G-estimator and heuristic schemes while offering no computational advantages.

The MATLAB code for the tuning framework based on the G-estimator of error and the heuristic can be found at `https://github.com/niyazil/DA-RP-ensemble_tuning`.

5. Conclusion and Limitations

In this work, we studied randomly-projected LDA ensemble classifiers. In particular, we looked into two main categories of these classifiers: discriminant-averaging and vote-averaging RP-LDA ensemble classifiers. We conducted an asymptotic analysis of the ensembles using RMT tools in a regime where the data and projection dimensions are assumed to grow at constant rates to each other. As a result, we derived their asymptotic distributions as well as limits of their class-conditional discriminant statistics.

Some important outcomes of this study include a newfound knowledge of the direct effect of ensemble size on classification performance, the optimal way of constructing an ensemble out of a set of RP-LDA discriminants, and whether or not projection selection matters for these classifiers. More specifically, we show that ensemble size improves RP-LDA ensemble classifier performance through decreasing the class-conditional discriminant variances while maintaining their mean separations with increasing ensemble size. Moreover, we find that equally-weighted discriminant-averaging is the asymptotically optimal method of combining RP-LDA discriminants for Gaussian data. In particular, a zero-one binary weight scheme—equivalent to projection selection—is not optimal, and neither are non-linear schemes such as vote-averaging. These findings are confirmed on real and synthetic data.

Following from these findings, we propose a framework for tuning the discriminant-averaging RP-LDA ensemble classifier parameters. We incorporate G-estimators derived from the RMT analysis into this framework in order to improve computational complexity at training. This is demonstrated on both real and synthetic data.

One limitation of this work is that the analysis is based entirely on a binary classifier, while in practice, a multi-class classifier may be needed for some problems. This is not an issue, as the binary classifier readily extends to a multi-class setting as done by Sifaou et al. (2020), who consider another variant of the LDA binary classifier derived under the assumption of two Gaussian classes. In general, in order to apply a binary classifier to the multi-class setting, one may consider a one-versus-the-rest approach or a one-versus-one approach (Bishop, 2006). Grouping multiple classes together as in the one-versus-the-rest approach violates the assumption of Gaussian classes (since 'the rest' is a Gaussian mixture model). Thus, the one-versus-one approach is the only viable option, but can lead to ambiguous classification (Bishop, 2006). As pointed out by Sifaou et al. (2020), ambiguities can be resolved by deciding on the class with a higher discriminant score.

Another apparent limitation of this work is the assumption of Gaussian random projections. This assumption is inherited from the literature, mainly the body of work done by Durrant and Kabán (2015) and Cannings and Samworth (2017). This is not strictly required by the results in our paper as, for our derivations, all that is needed is for the distribution of the projection to be invariant to orthogonal projection as well as to ensure that the fourth moment of this distribution is finite. Our results then hold for any such random projection. Gaussian projections are, however, desirable, as their behavior is well-understood within the field of randomized numerical linear algebra, as well as having strong performance guarantees. Yet the choice of Gaussian projections comes at the cost of other desirable properties. For example, sparse projections or sub-Gaussian projections provide various advantages over the classical Gaussian projection, such as simplicity and computational efficiency. The properties of these alternative distributions in the context of RP-LDA ensemble classification would make an interesting potential future research direction.

On another note, we require that the data itself be Gaussian in order to be able to derive an explicit expression for the probability of misclassification on top of which we build all of our results. This condition can be relaxed using a Central Limit Theorem argument. This is validated by the real data simulations where the data does not necessarily meet the Gaussian assumption.

Acknowledgments

The authors would like to thank the anonymous reviewers for their very helpful and constructive feedback. This work was supported by the King Abdullah University of Science and Technology (KAUST) Competitive Research Grant (CRG) funding program and by the KAUST Office of Sponsored Research (OSR) under Award No. OSR-CRG2019-4041.

Appendix A. Single RP-LDA Classifier Class-Conditional Discriminant Statistics

This section of the appendices derives the DEs for the single RP-LDA classifier class-conditional discriminant statistics as well as related proofs.

A.1 Means

In this section, we derive the DE of the quantity $m_i(1),\ i = 0, 1$, defined in (3). Using the law of total expectation, we have

$$\begin{aligned} m_i(1) &= \mathbb{E}_{\mathcal{T},\mathbf{R}}\left[\mathbb{E}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right]\right] \\ &= \mathbb{E}_{\mathcal{T},\mathbf{R}}\left[\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0+\hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right] \\ &= \mathbb{E}_{\mathcal{T}}\left[\hat{\boldsymbol{\mu}}^T\mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]\left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0+\hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right],\ i = 0, 1. \end{aligned}$$

Starting from the expression in the last line, we can proceed with the derivation of the DE as we would with the class-conditional mean of the discriminant-averaging RP-LDA infinite ensemble classifier discriminant. Note that this is exactly why we end up having the equivalence $\bar{m}_i(1) = \bar{m}_i^{M=\infty}$. Based on the derivation by Niyazi et al. (2020a), for $i = 0, 1$,

$$\begin{aligned} \bar{m}_i(1) = \frac{1}{2}\lim_{\beta\to 0}\tilde{\nu}_1(\beta)\Bigg[&(-1)^{i+1}\boldsymbol{\mu}^T\left(\frac{p}{n-2}g\boldsymbol{\Sigma}+\mathbf{I}_p\right)^{-1}\boldsymbol{\mu}+ \\ &\left(\frac{1}{n_0}-\frac{1}{n_1}\right)\text{tr}\left\{\boldsymbol{\Sigma}\left(\frac{p}{n-2}g\boldsymbol{\Sigma}+\mathbf{I}_p\right)^{-1}\right\}\Bigg] + \ln\frac{\pi_1}{\pi_0}, \end{aligned} \tag{15}$$

where g satisfies the system of equations defined by

$$\frac{p}{n-2}g = \frac{\lim_{\beta\to 0}\tilde{\nu}_1(\beta)}{1+\tilde{g}} \tag{16}$$

and

$$\tilde{g} = \frac{\lim_{\beta\to 0}\tilde{\nu}_1(\beta)}{n-2}\text{tr}\left\{\boldsymbol{\Sigma}\left(\mathbf{I}_p + \frac{p}{n-2}g\boldsymbol{\Sigma}\right)^{-1}\right\}, \tag{17}$$

and

$$\lim_{\beta\to 0}\tilde{\nu}_1(\beta) = \frac{\frac{p}{n-2}y^*}{1-\frac{p}{n-2}y^*\frac{1}{n-2}\text{tr}\left\{\boldsymbol{\Sigma}\left(\mathbf{I}_p+\frac{p}{n-2}y^*\boldsymbol{\Sigma}\right)^{-1}\right\}},$$

where y^* is the unique root of the function

$$h(y) = 1 - \frac{p}{d} + \frac{1}{d}\text{tr}\left\{\left(\mathbf{I}_p + \frac{p}{n-2}y\mathbf{D}_{\boldsymbol{\Sigma}}\right)^{-1}\right\},$$

which exists when $p > d$. Since $\bar{m}_i(1) = \bar{m}_i^{M=\infty}$, we denote both DEs by $\bar{m}_i$.

A.2 Variance

In this section, we derive the DE of the quantity $\sigma^2(1)$ defined in (4). By making use of the law of total variance with conditioning on the training data and projections (which are independent of the test point $\mathbf{x}$ by assumption), we have

$$\sigma^2(1) = \mathbb{E}_{\mathcal{T},\mathbf{R}}\left[\mathrm{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right]\right]+\mathrm{Var}_{\mathcal{T},\mathbf{R}}\left[\mathbb{E}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right]\right] \tag{18}$$

The second term tends almost-surely to zero, as it is decaying. This can be shown using Lemma 3.1 in the paper by Hachem et al. (2013).

Now, based on the data assumptions on $\mathbf{x}$, the inner term of the first term in (18) is exactly

$$\mathrm{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right] = \hat{\boldsymbol{\mu}}^T\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\hat{\boldsymbol{\mu}}. \tag{19}$$

We derive the DE of (19) in what follows. The first term in (18) is then the expectation of this DE by the Vitali convergence theorem, since it can be shown that (19) is a uniformly integrable sequence of random variables. This class of random variables has the property that for a sequence X_n such that $X_n \asymp X$, we also have $\mathbb{E}[X_n] \asymp \mathbb{E}[X]$.

Note that the rank of the $p\times p$ matrix $\hat{\boldsymbol{\Sigma}}$ is at most $\min\{p, n-2\}$. Therefore, $\hat{\boldsymbol{\Sigma}}$ is singular when $p > n-2$. Let $r = \mathrm{rank}\left(\hat{\boldsymbol{\Sigma}}\right)$. Then $\hat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T = \mathbf{U}_r\mathbf{D}_r\mathbf{U}_r^T$, where $\mathbf{U}_r \in \mathbb{R}^{p\times r}$ contains the r eigenvectors of $\hat{\boldsymbol{\Sigma}}$ corresponding to non-zero eigenvalues and $\mathbf{D}_r \in \mathbb{R}^{r\times r}$ contains the non-zero eigenvalues of $\hat{\boldsymbol{\Sigma}}$ along its diagonal. This is the compact form of $\hat{\boldsymbol{\Sigma}}$. Note that since $\hat{\boldsymbol{\Sigma}}$ is symmetric (and thus a normal matrix), its pseudoinverse is $\hat{\boldsymbol{\Sigma}}^+ = \mathbf{U}_r\mathbf{D}_r^{-1}\mathbf{U}_r^T$. This is made use of later in the derivation. Also note that since we are deriving a DE (which should depend only on true statistics), access to the actual value of r is forbidden as it depends on the sample covariance matrix. Nonetheless, we can make use of the fact that under the Gaussian assumptions, $r = \min\{p, n-2\}$ almost-surely. Keep in mind that $\mathbf{U}_r^T\mathbf{U}_r = \mathbf{I}_r$, while, in general, $\mathbf{U}_r\mathbf{U}_r^T \neq \mathbf{I}_p$, except when $r = p$, that is, $p \leq n-2$.

We can decompose $\mathbf{U}$ as $\mathbf{U} = [\mathbf{U}_r\ \tilde{\mathbf{U}}_r]$, where $\tilde{\mathbf{U}}_r \in \mathbb{R}^{p\times(p-r)}$ has as its columns the eigenvectors corresponding to the zero eigenvalues of $\hat{\boldsymbol{\Sigma}}$. Then $\mathbf{I}_p = \mathbf{U}\mathbf{U}^T = \mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T$. Let $\mathbf{R}_r := \mathbf{R}\mathbf{U}_r \in \mathbb{R}^{d\times r}$ and $\tilde{\mathbf{R}}_r := \mathbf{R}\tilde{\mathbf{U}}_r \in \mathbb{R}^{d\times(p-r)}$. Define the resolvent $\mathbf{Q}_2(\beta)$ as

$$\begin{aligned}\mathbf{Q}_2(\beta) &:= (\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T + \beta\mathbf{I}_d)^{-1}\\ &= (\mathbf{R}_r\mathbf{D}_r\mathbf{R}_r^T + \beta\mathbf{I}_d)^{-1},\end{aligned}$$

then

$$\begin{aligned}\mathbf{A}(\beta) &:= \mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T + \beta\mathbf{I}_d)^{-1}\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T + \beta\mathbf{I}_d)^{-1}\mathbf{R}\\ &= \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R}\left(\mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\right)\boldsymbol{\Sigma}\left(\mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\right)\mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R}\\ &= \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R} + \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}\\ &\quad + \mathbf{R}^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R} + \mathbf{R}^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}.\end{aligned}$$

Overall,

$$\mathrm{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right] = \lim_{\beta\to 0}\hat{\boldsymbol{\mu}}^T\mathbf{A}(\beta)\hat{\boldsymbol{\mu}}.$$

Now we find the DE of the term $\hat{\boldsymbol{\mu}}^T\mathbf{A}(\beta)\hat{\boldsymbol{\mu}}$ after which we take the limit as $\beta \to 0$.

$$
\begin{align*}
\hat{\boldsymbol{\mu}}^T\mathbf{A}(\beta)\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T\left(\mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\right)\mathbf{A}(\beta)\left(\mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\right)\hat{\boldsymbol{\mu}} \\
&= \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{U}_r^T\mathbf{A}(\beta)\mathbf{U}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{U}_r^T\mathbf{A}(\beta)\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\mathbf{A}(\beta)\mathbf{U}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\mathbf{A}(\beta)\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}. \tag{20}
\end{align*}
$$

We consider each term in (20) one by one. The derivations which follow use the fact that the odd moments of a zero-mean Gaussian random variable are zero. This yields asymptotic simplifications when taking the expectation with respect to $\tilde{\mathbf{R}}_r$ which is independent of $\mathbf{R}_r$ and never appears in a resolvent.

For the first term in (20), we have

$$
\begin{align*}
\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{U}_r^T\mathbf{A}(\beta)\mathbf{U}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\asymp \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \tag{21} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}. \tag{22}
\end{align*}
$$

For the second term in (20), we have

$$
\begin{align*}
\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{U}_r^T\mathbf{A}(\beta)\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\asymp \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \tag{23} \\
&\quad + \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}. \tag{24}
\end{align*}
$$

For the third term in (20), we have

$$
\begin{align*}
\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\mathbf{A}(\beta)\mathbf{U}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&\asymp \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \tag{25} \\
&\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}. \tag{26}
\end{align*}
$$

Finally, the fourth term in (20) satisfies

$$\begin{aligned}\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\mathbf{A}(\beta)\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\ &\asymp \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \qquad (27) \\ &\quad + \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}. \qquad (28)\end{aligned}$$

We derive asymptotic equivalents with respect to the projections first. Define $\mathbf{Q}_1(\beta) = \left(\mathbf{D}_r^{1/2}\mathbf{R}_r^T\mathbf{R}_r\mathbf{D}_r^{1/2} + \beta\mathbf{I}_r\right)^{-1}$. For (21), we have

$$\begin{aligned}&\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\ &= \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{D}_r^{-1/2}\mathbf{D}_r^{1/2}\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{D}_r^{1/2}\mathbf{D}_r^{-1/2}\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{D}_r^{-1/2}\mathbf{D}_r^{1/2}\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{D}_r^{1/2}\mathbf{D}_r^{-1/2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\ &= \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^+\hat{\boldsymbol{\mu}} - 2\beta\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^+\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{D}_r^{-1/2}\mathbf{Q}_1(\beta)\mathbf{D}_r^{-1/2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\ &\quad + \beta^2\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{D}_r^{-1/2}\mathbf{Q}_1(\beta)\mathbf{D}_r^{-1/2}\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{D}_r^{-1/2}\mathbf{Q}_1(\beta)\mathbf{D}_r^{-1/2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}},\end{aligned}$$

where the second-to-last line makes use of the following relation obtained from the matrix-inversion lemma:

$$\mathbf{D}_r^{1/2}\mathbf{R}_r^T(\mathbf{R}_r\mathbf{D}_r^{1/2}\mathbf{D}_r^{1/2}\mathbf{R}_r^T + \beta\mathbf{I}_d)^{-1}\mathbf{R}_r\mathbf{D}_r^{1/2} = \beta\left[\frac{1}{\beta}\mathbf{I}_r - \left(\mathbf{D}_r^{1/2}\mathbf{R}_r^T\mathbf{R}_r\mathbf{D}_r^{1/2} + \beta\mathbf{I}_r\right)^{-1}\right].$$

From the paper by Kammoun et al. (2019), we have

$$\mathbf{Q}_1(\beta) \leftrightarrow \mathbf{T}_1(\beta),$$

where

$$\begin{aligned}\mathbf{T}_1(\beta) &= \frac{1}{\beta}\left(\mathbf{I}_r + \tilde{\nu}_1(\beta)\mathbf{D}_r\right)^{-1} \\ \tilde{\mathbf{T}}_1(\beta) &= \frac{1}{\beta\left(1 + \frac{r}{d}\nu_1(\beta)\right)}\mathbf{I}_d\end{aligned}$$

and

$$\begin{aligned}\nu_1(\beta) &= \frac{1}{\beta}\frac{1}{r}\operatorname{tr}\left\{\mathbf{D}_r\left(\mathbf{I}_r + \tilde{\nu}_1(\beta)\mathbf{D}_r\right)^{-1}\right\} \\ \tilde{\nu}_1(\beta) &= \frac{1}{\beta\left(1 + \frac{r}{d}\nu_1(\beta)\right)}. \qquad (29)\end{aligned}$$

Using the above relations, we have

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\\
&\quad\asymp \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-1}\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\left(\mathbf{D}_r+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-1}\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\\
&\qquad+\left(\frac{1}{\tilde{\nu}_1(\beta)}\right)^2\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\left(\frac{\beta^2\theta(\mathbf{C})\tilde{\theta}}{1-\beta^2\theta(\mathbf{D}_r)\tilde{\theta}}\right),
\end{aligned}$$

where

$$\begin{aligned}
\beta^2\theta(\mathbf{D}_r)&=\frac{\beta^2}{r}\mathrm{tr}\left\{\mathbf{D}_r\mathbf{T}_1(\beta)\mathbf{D}_r\mathbf{T}_1(\beta)\right\}\\
&=\frac{1}{(\tilde{\nu}_1(\beta))^2}\frac{1}{r}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{D}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\},
\end{aligned}$$

$$\begin{aligned}
\beta^2\theta(\mathbf{C})&=\frac{\beta^2}{r}\mathrm{tr}\left\{\mathbf{D}_r\mathbf{T}_1(\beta)\mathbf{C}\mathbf{T}_1(\beta)\right\}\\
&=\frac{1}{r}\mathrm{tr}\left\{\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\right\}-\frac{2}{r}\mathrm{tr}\left\{\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}\mathbf{D}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\}\\
&\quad+\frac{1}{r}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{D}\right\}\\
&=\frac{1}{(\tilde{\nu}_1(\beta))^2}\frac{1}{r}\mathrm{tr}\left\{\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\}-\frac{1}{r}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\},
\end{aligned}\tag{30}$$

and

$$\tilde{\theta}=\frac{r}{d}\left(\frac{1}{\beta\left(1+\frac{r}{d}\nu_1(\beta)\right)}\right)^2=\frac{r}{d}\left(\tilde{\nu}_1(\beta)\right)^2,$$

where $\mathbf{C}:=\mathbf{D}_r^{-1/2}\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{D}_r^{-1/2}$.

Now consider (22). Let $\mathbf{a}=\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}$ and $\tilde{\mathbf{R}}_r$ have rows $\tilde{\mathbf{r}}_1,\ldots,\tilde{\mathbf{r}}_d$. We have the intermediate convergence

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}&=\sum_{i,j}a_ia_j\tilde{\mathbf{r}}_i^T\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{r}}_j\\
&\asymp\sum_i a_i^2\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\\
&=\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2^2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}.
\end{aligned}$$

We can show that

$$\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2^2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\asymp\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\}\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}\mathbf{D}_r+\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}}.$$

Using the paper by Kammoun et al. (2019), we have

$$\mathbf{Q}_2(\beta) \leftrightarrow \mathbf{T}_2(\beta), \tag{31}$$

where

$$\begin{aligned}\mathbf{T}_2(\beta) &= \frac{1}{\beta\left(1+\tilde{\nu}_2(\beta)\right)}\mathbf{I}_d \\ \tilde{\mathbf{T}}_2(\beta) &= \frac{1}{\beta}\left(\mathbf{I}_r + \nu_2(\beta)\mathbf{D}_r\right)^{-1}\end{aligned}$$

and

$$\begin{aligned}\nu_2(\beta) &= \frac{1}{\beta\left(1+\tilde{\nu}_2(\beta)\right)} \\ \tilde{\nu}_2(\beta) &= \frac{1}{\beta}\frac{1}{d}\mathrm{tr}\left\{\mathbf{D}_r\left(\mathbf{I}_r + \nu_2(\beta)\mathbf{D}_r\right)^{-1}\right\}.\end{aligned} \tag{32}$$

From (31), we have

$$\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\} \asymp \frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}.$$

Now, let's find the DE of $\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\}$. First, using the systems of equations in (29) and (32), we can show that $\nu_2(\beta) = \tilde{\nu}_1(\beta)$. Since

$$\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\} = -\frac{d\left[\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}\right]}{d\beta},$$

then

$$\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\} \asymp -\frac{d\left[\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}\right]}{d\beta},$$

that is, the limit of the derivative is the derivative of the limit. To justify this, first note that $\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}$ is a Stieltjes transform which is analytic outside the support of the spectrum of $\mathbf{R}\hat{\mathbf{\Sigma}}\mathbf{R}^T$. Since the support of the spectrum is bounded away from zero, taking $\beta \to 0$ ensures that $\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}$ is analytic. Similarly, $\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}$ is a Stieltjes transform which is analytic outside the support of the limiting spectrum of $\mathbf{R}\hat{\mathbf{\Sigma}}\mathbf{R}^T$, and since the support of the limiting spectrum is bounded away from zero, taking $\beta \to 0$ ensures that the $\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}$ is analytic. Since both $\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}$ and its limit $\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}$ are analytic, it follows that all derivatives of $\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}$ of any order converge to the corresponding derivatives of $\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}$. Then, because

$$\begin{aligned}\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\} &= \frac{1}{d}\mathrm{tr}\left\{\frac{1}{\beta(1+\tilde{\nu}_2(\beta))}\right\} \\ &= \nu_2(\beta) \\ &= \tilde{\nu}_1(\beta),\end{aligned}$$

we have

$$\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\} \asymp -\tilde{\nu}_1'(\beta).$$

We can solve the following set of equations (obtained by differentiating the system of equations (29)) for $\tilde{\nu}_1'(\beta)$:

$$\begin{aligned}
\nu_1'(\beta) &= -\frac{\tilde{\nu}_1'(\beta)}{\beta}\frac{1}{r}\mathrm{tr}\left\{\mathbf{D}_r\left(\mathbf{I}_r+\tilde{\nu}_1(\beta)\mathbf{D}_r\right)^{-1}\mathbf{D}_r\left(\mathbf{I}_r+\tilde{\nu}_1(\beta)\mathbf{D}_r\right)^{-1}\right\}\\
&\quad -\frac{1}{\beta^2}\frac{1}{r}\mathrm{tr}\left\{\mathbf{D}_r\left(\mathbf{I}_r+\tilde{\nu}_1(\beta)\mathbf{D}_r\right)^{-1}\right\}\\
&= -\tilde{\nu}_1'(\beta)\beta\theta(\mathbf{D}_r)-\frac{1}{\beta}\nu_1(\beta)\\
\tilde{\nu}_1'(\beta) &= -\frac{r}{d}\frac{\nu_1'(\beta)}{\beta}\frac{1}{\left(1+\frac{r}{d}\nu_1(\beta)\right)^2}-\frac{1}{\beta^2\left(1+\frac{r}{d}\nu_1(\beta)\right)}\\
&= -\nu_1'(\beta)\beta\tilde{\theta}-\frac{1}{\beta}\tilde{\nu}_1(\beta),
\end{aligned}$$

from which

$$\tilde{\nu}_1'(\beta)=\frac{\nu_1(\beta)\tilde{\theta}-\frac{1}{\beta}\tilde{\nu}_1(\beta)}{1-\beta^2\theta(\mathbf{D}_r)\tilde{\theta}}.$$

So, overall we have

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\\
&\quad\asymp -\frac{\tilde{\nu}_1'(\beta)}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\frac{1}{d}\mathrm{tr}\left\{\mathbf{T}_2(\beta)\right\}\mathbf{D}_r+\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\\
&\quad= -\frac{\tilde{\nu}_1'(\beta)}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\tilde{\nu}_1(\beta)\mathbf{D}_r+\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}}\\
&\quad= -\frac{\tilde{\nu}_1'(\beta)}{\left(\tilde{\nu}_1(\beta)\right)^2}\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}}.
\end{aligned}$$

Applying the same techniques to (23), we can show that

$$\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\asymp\tilde{\nu}_1(\beta)\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-1}\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}.$$

For (24), we have

$$\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\asymp 0.$$

The terms (25) and (26) are just the transpose of (23) and (24). For (27), we have

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\\
&\quad\asymp -\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\frac{\tilde{\nu}_1'(\beta)}{(\tilde{\nu}_1(\beta))^2}\frac{1}{d}\mathrm{tr}\left\{\boldsymbol{\Sigma}\mathbf{U}_r\left(\mathbf{D}_r+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\right\},
\end{aligned}$$

and for the final term (28), we have

$$\begin{aligned}&\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\\&\quad=\sum_{i,j,k,l}\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{k,l}\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_l^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_j.\end{aligned}$$

It is easy to see that only three cases survive asymptotically in this summation:

1. $i=j=k=l$;
2. $i=k,\ j=l,\ i\neq j$;
3. $i=j,\ k=l,\ i\neq k$.

For the first case,

$$\begin{aligned}&\sum_{i=j=k=l}\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{k,l}\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_l^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_j\\&\quad=\sum_i\left(\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\right)^2\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,i}\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_i\\&\quad\asymp\left[\frac{2}{d^2}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\}+\left(\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}\right)^2\right]\sum_i\left(\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\right)^2\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,i}\\&\quad\asymp\left(\tilde{\nu}_1(\beta)\right)^2\sum_i\left(\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\right)^2\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,i},\end{aligned}\tag{33}$$

where the third line uses the expectation of the term $\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_i$. For the second case,

$$\begin{aligned}&\sum_{i=k,\ j=l,\ i\neq j}\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{k,l}\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_l^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_j\\&\quad=\sum_{i\neq j}\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,j}\tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_j^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_j\\&\quad\asymp\left(\frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2(\beta)\right\}\right)^2\sum_{i\neq j}\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,j}\\&\quad\asymp\left(\tilde{\nu}_1(\beta)\right)^2\sum_{i\neq j}\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j\left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,j}.\end{aligned}\tag{34}$$

For the third case, we have

$$\begin{aligned}
&\sum_{i=j,\ k=l,\ i\neq k} \left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i \left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_j \left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{k,l} \tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_l^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_j \\
&\quad= \sum_{i\neq k} \left(\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\right)^2 \left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{k,k} \tilde{\mathbf{r}}_i^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T\mathbf{Q}_2(\beta)\tilde{\mathbf{r}}_i \\
&\quad\asymp \frac{1}{d^2}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\} \sum_{i\neq k} \left(\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\right)^2 \left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{k,k} \\
&\quad= \frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\}\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} - \frac{1}{d^2}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\}\sum_i \left(\left[\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\right]_i\right)^2 \left[\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right]_{i,i} \\
&\quad\asymp \frac{1}{d}\mathrm{tr}\left\{\mathbf{Q}_2^2(\beta)\right\}\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\quad\asymp -\tilde{\nu}_1'(\beta)\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}. \qquad (35)
\end{aligned}$$

Combining (33), (34), and (35), we have overall

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&\qquad\asymp (\tilde{\nu}_1(\beta))^2\,\hat{\boldsymbol{\mu}}\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} - \tilde{\nu}_1'(\beta)\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}. \qquad (36)
\end{aligned}$$

Combining the above derivations starting from (20) to (36), we have

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^T\mathbf{A}(\beta)\hat{\boldsymbol{\mu}} &\asymp \hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-1}\mathbf{U}_r^T\boldsymbol{\Sigma}\mathbf{U}_r\left(\mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-1}\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&+ \left[\left(\frac{\beta^2\theta(\mathbf{C})\tilde{\theta}}{1-\beta^2\theta(\mathbf{D}_r)\tilde{\theta}}\right) - \tilde{\nu}_1'(\beta)\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\right]\frac{1}{(\tilde{\nu}_1(\beta))^2}\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\hat{\boldsymbol{\mu}} \\
&+ 2\tilde{\nu}_1(\beta)\hat{\boldsymbol{\mu}}^T\mathbf{U}_r\left(\mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_r\right)^{-1}\mathbf{U}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} \\
&- \hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}\frac{\tilde{\nu}_1'(\beta)}{(\tilde{\nu}_1(\beta))^2}\frac{1}{d}\mathrm{tr}\left\{\boldsymbol{\Sigma}\mathbf{U}_r\left(\mathbf{D}_r + \frac{1}{\tilde{\nu}}_1(\beta)\mathbf{I}_r\right)^{-2}\mathbf{U}_r^T\right\} \\
&+ (\tilde{\nu}_1(\beta))^2\,\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}} - \tilde{\nu}_1'(\beta)\frac{1}{d}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\}\hat{\boldsymbol{\mu}}^T\tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T\hat{\boldsymbol{\mu}}.
\end{aligned}$$

Through a series of manipulations in which we express everything in terms of $\mathbf{D}$ instead of $\mathbf{D}_r$ and also by using the relation

$$\tilde{\nu}_1'(\beta) = -\frac{(\tilde{\nu}_1(\beta))^2}{1-\beta^2\theta(\mathbf{D}_r)\tilde{\theta}}$$

obtained through the system of equations (29), and by expressing (30) as

$$\beta^2\theta(\mathbf{C}) = \beta^2\theta(\mathbf{C}') - \frac{1}{r}\mathrm{tr}\left\{\tilde{\mathbf{U}}_r^T\boldsymbol{\Sigma}\tilde{\mathbf{U}}_r\right\},$$

where

$$\beta^2\theta(\mathbf{C}') = \frac{1}{(\tilde{\nu}_1(\beta))^2}\frac{1}{r}\mathrm{tr}\left\{\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\},$$

we have the simplification

$$\begin{aligned}\hat{\boldsymbol{\mu}}^T\mathbf{A}(\beta)\hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T\mathbf{U}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\mathbf{U}^T\hat{\boldsymbol{\mu}}\\ +\left(\frac{\beta^2\theta(\mathbf{C}')\tilde{\theta}}{1-\beta^2\theta(\mathbf{D}_r)\tilde{\theta}}\right)\frac{1}{(\tilde{\nu}_1(\beta))^2}\hat{\boldsymbol{\mu}}^T\mathbf{U}\left(\mathbf{D}+\frac{1}{\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-2}\mathbf{U}^T\hat{\boldsymbol{\mu}}.\end{aligned} \tag{37}$$

Now what must be done is to remove the randomness from the training. This appears in $\hat{\boldsymbol{\mu}}$, $\mathbf{D}$, and in the current definition of $\tilde{\nu}_1(\beta)$.

First, we derive $\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)$ in such a way that it depends only on the true statistics. Using the equations in (29), it can be shown that

$$1-\frac{p}{d}+\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\frac{1}{d}\mathrm{tr}\left\{\left(\hat{\mathbf{\Sigma}}+\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\}=0 \tag{38}$$

Using the fact that $\hat{\mathbf{\Sigma}} = \frac{1}{n-2}\mathbf{\Sigma}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Sigma}^{1/2}$ for some $\mathbf{Z}\in\mathbb{R}^{p\times(n-2)}$ with i.i.d. standard Gaussian entries and by eigendecomposing $\mathbf{\Sigma}$ as $\mathbf{\Sigma}=\mathbf{V}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{V}^T$, we have

$$\begin{aligned}\frac{1}{d}\mathrm{tr}\left\{\left(\hat{\mathbf{\Sigma}}+\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\} &= \frac{1}{d}\mathrm{tr}\left\{\left(\frac{1}{n-2}\mathbf{\Sigma}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Sigma}^{1/2}+\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\}\\ &\sim \frac{1}{d}\mathrm{tr}\left\{\left(\frac{1}{n-2}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}+\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\}\end{aligned}$$

From the paper by Kammoun et al. (2019), we have

$$\mathbf{W}(\gamma)\leftrightarrow\mathbf{E}(\gamma),$$

where

$$\mathbf{W}(\gamma)=\left(\frac{1}{n-2}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}-\gamma\mathbf{I}_p\right)^{-1},$$

$$\mathbf{E}(\gamma)=-\frac{1}{\gamma}\left(\mathbf{I}_p+\frac{p}{n-2}g(\gamma)\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1},$$

$$\frac{p}{n-2}g(\gamma)=-\frac{1}{\gamma}\frac{1}{1+\tilde{g}(\gamma)},$$

and

$$\tilde{g}(\gamma)=-\frac{1}{\gamma}\frac{1}{p}\mathrm{tr}\left\{\frac{p}{n-2}\mathbf{D}_{\mathbf{\Sigma}}\left(\mathbf{I}_p+\frac{p}{n-2}g(\gamma)\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\},$$

from which it follows that

$$\frac{1}{d}\mathrm{tr}\left\{\left(\hat{\mathbf{\Sigma}}+\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\mathbf{I}_p\right)^{-1}\right\}\asymp\frac{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}{d}\mathrm{tr}\left\{\left(\mathbf{I}_p+\frac{p}{n-2}g\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\}.$$

Let $g := g\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}(\beta)}\right)$ and $\tilde{g} := \tilde{g}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}(\beta)}\right)$. We now have

$$1-\frac{p}{d}+\frac{1}{d}\mathrm{tr}\left\{\left(\mathbf{I}_p+\frac{p}{n-2}g\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\}\asymp 0. \tag{39}$$

Using (39), the quantity g can be solved for as the unique root y^* of the monotonically decreasing function

$$\begin{aligned}h(y)&=1-\frac{p}{d}+\frac{1}{d}\mathrm{tr}\left\{\left(\mathbf{I}_p+\frac{p}{n-2}y\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\}\\&=1-\frac{p}{d}+\frac{1}{d}\sum_{i=1}^{p}\frac{1}{1+\frac{p}{n-2}\lambda_i(\mathbf{\Sigma})y},\end{aligned}$$

which exists when $p>d$. It can be shown that $g\asymp y^*$. Then combining (16) and (17), we can solve for $\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)$ in terms of g, and so we have

$$\lim_{\beta\to 0}\tilde{\nu}_1(\beta)=\frac{\frac{p}{n-2}y^*}{1-\frac{p}{n-2}y^*\frac{1}{n-2}\mathrm{tr}\left\{\mathbf{D}_{\mathbf{\Sigma}}\left(\mathbf{I}_p+\frac{p}{n-2}y^*\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\}}.$$

By dealing with the randomness from the sample covariance in (37) using similar techniques, followed by taking the limit as $\beta\to 0$, we obtain

$$\begin{aligned}\lim_{\beta\to 0}\hat{\boldsymbol{\mu}}^T\mathbf{A}(\beta)\hat{\boldsymbol{\mu}}&=\hat{\boldsymbol{\mu}}^T\mathbf{R}^T(\mathbf{R}\hat{\mathbf{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{\Sigma}\mathbf{R}^T(\mathbf{R}\hat{\mathbf{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\hat{\boldsymbol{\mu}}\\&\asymp\left(\frac{1}{1-\Omega}\right)\hat{\boldsymbol{\mu}}^T\mathbf{V}\mathbf{E}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{E}\mathbf{V}^T\hat{\boldsymbol{\mu}}+\end{aligned}$$

$$\frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\left[\hat{\boldsymbol{\mu}}^T\mathbf{V}\mathbf{E}^2\mathbf{V}^T\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}}^T\mathbf{V}\mathbf{E}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{E}\mathbf{V}^T\hat{\boldsymbol{\mu}}\left(\frac{\left(\frac{\frac{p}{n-2}g}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)^2\frac{1}{n-2}\mathrm{tr}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{E}^2}{1-\Omega}\right)\right]\times$$

$$\left(\frac{\left(\frac{1}{1-\Omega}\right)\frac{1}{d}\mathrm{tr}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{E}^2}{1-\frac{p}{d}+\frac{2}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\frac{1}{d}\mathrm{tr}\mathbf{E}-\frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\left[\frac{1}{d}\mathrm{tr}\mathbf{E}^2+\frac{1}{d}\mathrm{tr}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{E}^2\left(\frac{\left(\frac{\frac{p}{n-2}g}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)^2\frac{1}{n-2}\mathrm{tr}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{E}^2}{1-\Omega}\right)\right]}\right),$$

where

$$\Omega = \left(\frac{\frac{p}{n-2} g}{\lim\limits_{\beta \to 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \operatorname{tr}\left\{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \right\}.$$

The final step is to remove the randomness coming from the sample means in $\hat{\boldsymbol{\mu}}$. Using

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0},$$

where $\mathbf{Z}_i \in \mathbb{R}^{p \times n_i}$, $i = 0, 1$, has i.i.d. $\mathcal{N}(0, 1)$ entries, and taking the expectation over $\mathbf{Z}_i \mathbf{1}$, $i = 0, 1$, while making use of the fact that $\frac{\mathbf{Z}_i \mathbf{1}}{n_i} \sim \mathcal{N}\left(\mathbf{0}_p, \frac{1}{n_i} \mathbf{I}_p \right)$, $i = 0, 1$, we have

$$\begin{aligned}
&\bar{\sigma}^2(1) = \left(\frac{1}{1 - \Omega} \right) \left[\boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{V}^T \boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \operatorname{tr}\left\{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \right\} \right] + \\
&\quad \frac{1}{\left(\lim\limits_{\beta \to 0} \tilde{\nu}_1(\beta) \right)^2} \Bigg[\boldsymbol{\mu}^T \mathbf{V} \mathbf{E}^2 \mathbf{V}^T \boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \operatorname{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2 + \\
&\left(\boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{V}^T \boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \operatorname{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \right) \left(\frac{\left(\frac{\frac{p}{n-2} g}{\lim\limits_{\beta \to 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \operatorname{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2}{1 - \Omega} \right) \Bigg] \times \\
&\left(\frac{\left(\frac{1}{1 - \Omega} \right) \frac{1}{d} \operatorname{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2}{1 - \frac{p}{d} + \frac{2}{\lim\limits_{\beta \to 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \operatorname{tr} \mathbf{E} - \frac{1}{\left(\lim\limits_{\beta \to 0} \tilde{\nu}_1(\beta) \right)^2} \left[\frac{1}{d} \operatorname{tr} \mathbf{E}^2 + \frac{1}{d} \operatorname{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2 \left(\frac{\left(\frac{\frac{p}{n-2} g}{\lim\limits_{\beta \to 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \operatorname{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2}{1 - \Omega} \right) \right]} \right).
\end{aligned} \tag{40}$$

A.2.1 Proof that the Single RP-LDA Discriminant Variance is Asymptotically Greater than that of the Infinite Ensemble

We simply prove that $\bar{\sigma}^2(1) > \bar{\sigma}^2_{M=\infty}$. Using the expressions in (40) and (45), we have

$$\bar{\sigma}^2(1) - \bar{\sigma}^2_{M=\infty} = \frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\left[\boldsymbol{\mu}^T\mathbf{V}\mathbf{E}^2\mathbf{V}^T\boldsymbol{\mu} + \left(\frac{1}{n_0}+\frac{1}{n_1}\right)\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}^2\right\} + \right.$$

$$\left.\left(\boldsymbol{\mu}^T\mathbf{V}\mathbf{E}\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}\mathbf{V}^T\boldsymbol{\mu} + \left(\frac{1}{n_0}+\frac{1}{n_1}\right)\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}\right\}\right)\left(\frac{\left(\frac{\frac{p}{n-2}g}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)^2\frac{1}{n-2}\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}^2\right\}}{1-\Omega}\right)\right]\times$$

$$\left(\frac{\left(\frac{1}{1-\Omega}\right)\frac{1}{d}\mathrm{tr}\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}^2}{1-\frac{p}{d}+\frac{2}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\frac{1}{d}\mathrm{tr}\mathbf{E}-\frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\left[\frac{1}{d}\mathrm{tr}\mathbf{E}^2+\frac{1}{d}\mathrm{tr}\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}^2\left(\frac{\left(\frac{\frac{p}{n-2}g}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)^2\frac{1}{n-2}\mathrm{tr}\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{E}^2}{1-\Omega}\right)\right]}\right). \tag{41}$$

Now we must show that each of the constituent terms of (41) is positive. The term $1-\Omega$ fits the form of the term $1 - t^2\gamma_n(t)\tilde{\gamma}_n(t)$ in the paper by Hachem et al. (2008) in which it was shown to be positive. Additionally, all traces and quadratic terms in the first and second lines of (41) are positive since the matrices involved are positive definite. What remains is the denominator of the fraction in the last line. This term comes from taking the asymptotic limit of the term $1 - \lim\limits_{\beta\to 0}\beta^2\theta(\mathbf{D}_r)\tilde{\theta}$ which can be expressed as

$$1 - \lim_{\beta\to 0}\beta^2\theta(\mathbf{D}_r)\tilde{\theta} = 1 - \frac{p}{d} + \frac{2}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\frac{1}{d}\mathrm{tr}\left\{\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\right\}$$

$$-\frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\frac{1}{d}\mathrm{tr}\left\{\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\right\}. \tag{42}$$

Using (38), equation (42) simplifies to

$$\begin{aligned}1-\lim_{\beta\to 0}\beta^2\theta(\mathbf{D}_r)\tilde{\theta} &= \frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\frac{1}{d}\mathrm{tr}\left\{\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\right\}\\ &\quad-\frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\frac{1}{d}\mathrm{tr}\left\{\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\right\}.\end{aligned}$$

Let $\mathbf{G} := \lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\mathbf{D}+\mathbf{I}_p$. Then, using the relation $\mathbf{A}^{-1}-\mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$ with $\mathbf{A}^{-1} = \frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)$ and $\mathbf{B}^{-1} = \frac{1}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)\right)^2}\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)\mathbf{W}\left(-\frac{1}{\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)}\right)$, we have

$$\begin{aligned}1-\lim_{\beta\to 0}\beta^2\theta(\mathbf{D}_r)\tilde{\theta} &= \frac{1}{d}\mathrm{tr}\left\{\mathbf{G}^{-1}\left(\mathbf{G}^2-\mathbf{G}\right)\mathbf{G}^{-2}\right\}\\ &= \frac{1}{d}\sum_{i=1}^{p}\frac{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)d_i+1\right)^2-\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)d_i+1\right)}{\left(\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)d_i+1\right)^3}\\ &> 0,\end{aligned}$$

since $\lim\limits_{\beta\to 0}\tilde{\nu}_1(\beta)d_i+1>1$, where d_i is the i^{th} entry along the diagonal of the diagonal matrix $\mathbf{D}$.

Appendix B. Discriminant-Averaging RP-LDA Ensemble Classifier Class-Conditional Discriminant Statistics

This section of the appendices derives the DEs for the discriminant-averaging RP-LDA ensemble classifier class-conditional discriminant statistics.

B.1 Means

In this section, we derive the DE of the quantity $m_i(M),\ i=0,1$, defined in (6). By the law of total expectation, we have

$$\begin{aligned}m_i(M) &= \mathbb{E}_{\mathcal{T},\mathbf{R}}\left[\mathbb{E}\left[W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right]\right]\\ &= \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_{\mathcal{T},\mathbf{R}}\left[\mathbb{E}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}\right]\right]\\ &\asymp \frac{1}{M}\sum_{k=1}^{M}\bar{m}_i(1)\\ &= \bar{m}_i(1),\ i=0,1,\end{aligned}$$

where the convergence in the second-to-last line is proven in Appendix A.1. Thus, $\bar{m}_i(M) = \bar{m}_i(1)$ and we denote both DEs by $\bar{m}_i$.

B.2 Variance

In this section, we derive the DE of the quantity $\sigma^2(M)$, defined in (7). By the law of total variance,

$$\begin{aligned}\sigma^2(M) = \mathbb{E}_{\mathcal{T},\mathbf{R}}&\left[\mathrm{Var}\left[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}\right]\right] \\ &+ \mathrm{Var}_{\mathcal{T},\mathbf{R}}\left[\mathbb{E}\left[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}\right]\right]. \end{aligned} \tag{43}$$

For a similar reason to that in Appendix A.2, the second term in (43) is asymptotically zero. Considering the inner term of the first term, we have

$$\begin{aligned}&\mathrm{Var}\left[W_{\text{disc-avg}}\left(\mathbf{x}, \{\mathbf{R}_k\}_{k=1}^M\right) \Big| \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M\right] \\ &= \mathrm{Var}\left[\frac{1}{M}\sum_{k=1}^{M} W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right) \Bigg| \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M\right] \\ &= \frac{1}{M^2}\sum_{k=1}^{M} \mathrm{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}_k\right] \\ &\quad + \frac{1}{M^2}\sum_{k\neq j}^{M} \mathrm{Cov}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right), W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_j\right) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}_k, \mathbf{R}_j\right] \\ &\asymp \frac{1}{M}\bar{\sigma}^2(1) + \frac{M-1}{M^2}\bar{\sigma}^2_{M=\infty} \\ &= \frac{1}{M}\bar{\sigma}^2(1) + \left(1 - \frac{1}{M}\right)\bar{\sigma}^2_{M=\infty}, \end{aligned} \tag{44}$$

where the convergence in the second-to-last line follows from the proof in Appendix A.2 and also the fact that

$$\begin{aligned}&\mathrm{Cov}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right), W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_j\right) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}\right] \\ &\quad = \hat{\boldsymbol{\mu}}^T \mathbf{R}_k^T (\mathbf{R}_k \hat{\boldsymbol{\Sigma}} \mathbf{R}_k^T)^{-1} \mathbf{R}_k \boldsymbol{\Sigma} \mathbf{R}_j^T (\mathbf{R}_j \hat{\boldsymbol{\Sigma}} \mathbf{R}_j^T)^{-1} \mathbf{R}_j \hat{\boldsymbol{\mu}} \\ &\quad \asymp \hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}}\left[\mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R}\right] \boldsymbol{\Sigma} \mathbb{E}_{\mathbf{R}}\left[\mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R}\right] \hat{\boldsymbol{\mu}} \\ &\quad \asymp \bar{\sigma}^2_{M=\infty}. \end{aligned}$$

The exact expression of $\bar{\sigma}^2_{M=\infty}$ is derived by Niyazi et al. (2020a) as

$$\bar{\sigma}^2_{M=\infty} = \frac{1}{1-\Omega}\left[\boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{V} \boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1}\right) \mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}\right\}\right]. \tag{45}$$

Appendix C. RP-LDA Ensemble Classifier Error Analysis

This section of the appendices derives the DEs for the probabilities of misclassification of the discriminant-averaging, asymptotic MAP, and vote-averaging RP-LDA ensemble classifiers.

C.1 Discriminant-Averaging RP-LDA Ensemble Classifier

The expected probability of misclassification DE of the discriminant-averaging RP-LDA ensemble classifier composed of M RP-LDA discriminants is

$$\bar{\varepsilon} := \pi_0 \Phi\left(\frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2(M)}}\right) + \pi_1 \Phi\left(-\frac{\bar{m}_1}{\sqrt{\bar{\sigma}^2(M)}}\right), \tag{46}$$

where $M = 1, 2, \ldots$. This statement claims the convergence of the expected probability of misclassification of the discriminant-averaging RP-LDA ensemble (over training and projections) to the probability of misclassification computed using the distribution of the asymptotic discriminant stated in Theorem 4. This follows from the convergence in distribution in Theorem 4 of this paper and Lemma 2.11 in the book by Vaart (1998). Note that the convergence is not in the probabilistic sense; the probability of misclassification is conditioned on the training and random projections before applying Lemma 2.11 to obtain its limit. The limit of the expected probability of misclassification over the training and random projections (46) is then simply the expectation of the first limit. This follows by the bounded convergence theorem since the probability measure is upper bounded by 1.

C.2 Asymptotic MAP RP-LDA Ensemble Classifier

The expected probability of misclassification DE of the asymptotic MAP RP-LDA ensemble classifier composed of M RP-LDA discriminants is

$$\pi_0 \Phi\left(\frac{-\frac{1}{2}\frac{(\bar{m}_1-\bar{m}_0)^2}{\bar{\sigma}^2(M)} + \ln\frac{\pi_1}{\pi_0}}{\sqrt{\frac{(\bar{m}_1-\bar{m}_0)^2}{\bar{\sigma}^2(M)}}}\right) + \pi_1 \Phi\left(\frac{-\frac{1}{2}\frac{(\bar{m}_1-\bar{m}_0)^2}{\bar{\sigma}^2(M)} - \ln\frac{\pi_1}{\pi_0}}{\sqrt{\frac{(\bar{m}_1-\bar{m}_0)^2}{\bar{\sigma}^2(M)}}}\right),$$

where $M = 1, 2, \ldots$. This statement claims the convergence of the expected probability of misclassification of the asymptotic MAP RP-LDA ensemble (over training and projections) to the probability of misclassification computed using the distribution of the asymptotic discriminant, which can be derived easily from Theorem 4. This result then follows from this convergence in distribution and Lemma 2.11 in the book by Vaart (1998). Again, the convergence is not in the probabilistic sense.

C.3 Vote-Averaging RP-LDA Ensemble Classifier

As stated in Theorem 7, the asymptotic distribution of the vote-averaging RP-LDA ensemble class-conditional discriminant times M is a correlated binomial having M trials, probability of success $\bar{p}_i$, $i = 0, 1$, and constant correlation $\bar{\rho}_i$ between trial outcomes. In this case, however, knowing the distribution is not enough to determine the asymptotic probability of misclassification. This is because the *correlated* binomial PMF is not uniquely specified by the correlation coefficient(s) between and probability of success of each trial. Additional information pertaining to the conditional correlations is needed. Consequently, multiple models for correlated binomials based on varying assumptions have been proposed.

One of these models is Moody's model (Witt, 2004). Moody's correlated binomial model makes the assumption that the conditional correlations of the outcomes of any two trials given that any subset of the others are all successes is constant. To test this model, we generate the empirical PMF of the vote-averaging RP-LDA ensemble classifier. This model fits the empirical PMF well, at least up to $M = 35$, beyond which numerical issues occur which hinder the accurate computation of Moody's PMF. This seems to suggest that the constant conditional correlation condition holds for our setup.

Through our own numerical investigation, we find that this condition, in fact, does not hold, although the conditional correlations are close enough that the corresponding conditional probabilities of success end up being very close to those predicted by Moody's model. Since Moody's PMF is characterized by these conditional probabilities, this might explain why we have a close match.

Denote by $p_i(k)$ the probability that $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i$ asymptotically takes the value k. The asymptotic conditional PMF of the vote-averaging RP-LDA ensemble discriminant (times M) according to Moody's model is then

$$p_i(k) = \begin{cases} 1 + \sum_{j=1}^{M} (-1)^j \binom{M}{j} \prod_{i=1}^{j} \bar{p}_i, & \text{for } k = 0 \\ \binom{M}{k} \sum_{j=0}^{M-k} \left[(-1)^j \binom{M-k}{j} \prod_{l=1}^{j+k} \bar{p}_i^{(l)} \right], & \text{for } k = 1, \dots, M \\ 0, & \text{otherwise,} \end{cases}$$

where $\bar{p}_i^{(j)} = 1 - (1 - \bar{p}_i)(1 - \bar{\rho}_i)^{j-1}$, $j = 2, \dots, M$. Based on this, the asymptotic probability of misclassification of the vote-averaging RP-LDA ensemble discriminant with a threshold of 0.5 is

$$\pi_0 \sum_{k > M/2} p_0(k) + \pi_1 \sum_{k \leq M/2} p_1(k).$$

Appendix D. Proof of Asymptotic Distributions and Optimal Ensemble Construction

This section of the appendices derives the joint asymptotic distributions of a set of RP-LDA discriminants as well as the asymptotic distributions of the class-conditional discriminants of the discriminant-averaging RP-LDA finite ensemble classifier, the discriminant-averaging RP-LDA infinite ensemble classifier, and the vote-averaging RP-LDA ensemble classifier. It also derives the Neyman-Pearson RP-LDA ensemble classifier.

D.1 Asymptotic Joint Distribution of Multiple RP-LDA Discriminants

In this section, we prove the asymptotic joint distribution of M single RP-LDA discriminants as stated in Theorem 8.

Recall that $\mathbf{W} = \left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_1\right), \dots, W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_M\right)\right]^T$ and let

$$\hat{\mathbf{\Sigma}}_{\mathbf{R}_k}^{-1} := \mathbf{R}_k^T (\mathbf{R}_k \hat{\mathbf{\Sigma}} \mathbf{R}_k^T)^{-1} \mathbf{R}_k, \; k = 1, \dots, M.$$

Conditioned on $\{\mathbf{R}_k\}_{k=1}^{M}$ and the training set $\mathcal{T}$ and for $\mathbf{x} \in \mathcal{C}_i$, $\mathbf{W}$ is a Gaussian vector (through $\mathbf{x}$) with

$$\begin{aligned}\boldsymbol{\zeta}_i &:= \mathbb{E}\left[\mathbf{W}\middle|\{\mathbf{R}_k\}_{k=1}^{M}, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i\right] \\ &= \begin{bmatrix} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_1}^{-1} \left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \\ \vdots \\ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_M}^{-1} \left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \end{bmatrix}, \ i = 0, 1,\end{aligned}$$

and covariance $\boldsymbol{\Pi}$ with entries

$$\operatorname{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right) | \mathbf{R}_k, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i\right] = \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \hat{\boldsymbol{\mu}}, \ k = 1, \ldots, M,$$

along the diagonal, and

$$\begin{aligned}&\operatorname{Cov}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right), W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_j\right) | \mathbf{R}_k, \mathbf{R}_j, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i\right] \\ &= \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_j}^{-1} \hat{\boldsymbol{\mu}}, \ j, k = 1, \ldots, M, \ j \neq k,\end{aligned}$$

off the diagonal. From the derivations in Appendix A and Appendix B, we know that

$$\begin{aligned}\boldsymbol{\zeta}_i &\asymp \bar{\boldsymbol{\zeta}}_i \\ &= \bar{m}_i \mathbf{1}_M, \ i = 0, 1,\end{aligned}$$

$$\operatorname{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right) | \mathbf{R}_k, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i\right] \asymp \bar{\sigma}^2(1), \ \forall k,$$

and

$$\operatorname{Cov}\left[W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_k\right), W_{\text{RP-LDA}}\left(\mathbf{x}, \mathbf{R}_j\right) | \mathbf{R}_k, \mathbf{R}_j, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i\right] \asymp \bar{\sigma}_{M=\infty}^2, \ \forall j \neq k.$$

Since M is fixed, then $\boldsymbol{\Pi}$ defined above converges pointwise, and so we also have

$$\begin{aligned}\boldsymbol{\Pi} &\asymp \bar{\boldsymbol{\Pi}} \\ &= \begin{bmatrix} \bar{\sigma}^2(1) & \bar{\sigma}_{M=\infty}^2 & \cdots & \bar{\sigma}_{M=\infty}^2 \\ \bar{\sigma}_{M=\infty}^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \bar{\sigma}_{M=\infty}^2 \\ \bar{\sigma}_{M=\infty}^2 & \cdots & \bar{\sigma}_{M=\infty}^2 & \bar{\sigma}^2(1) \end{bmatrix} \\ &= \left(\bar{\sigma}^2(1) - \bar{\sigma}_{M=\infty}^2\right) \mathbf{I}_M + \bar{\sigma}_{M=\infty}^2 \mathbf{1}_M \mathbf{1}_M^T.\end{aligned}$$

Now we prove that $\mathbf{W}$ converges in distribution to a Gaussian random vector through its characteristic function.

Denote the characteristic function of $\mathbf{W}$ given $\mathbf{x} \in \mathcal{C}_i$ by $\phi_{\mathbf{W},i}(\boldsymbol{\omega})$. Then

$$\begin{aligned}
\phi_{\mathbf{W},i}(\boldsymbol{\omega}) &= \mathbb{E}\left[\exp\left(j\boldsymbol{\omega}^T\mathbf{W}\right)\middle|\mathbf{x} \in \mathcal{C}_i\right] \\
&= \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M,\mathcal{T}}\left[\mathbb{E}\left[\exp\left(j\boldsymbol{\omega}^T\mathbf{W}\right)\middle|\{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i\right]\right] \\
&= \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M,\mathcal{T}}\left[\exp\left(j\boldsymbol{\zeta}_i^T\boldsymbol{\omega} - \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\Pi}\boldsymbol{\omega}\right)\right] \\
&= \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M,\mathcal{T}}\left[\exp\left(j(\boldsymbol{\zeta}_i - \bar{\boldsymbol{\zeta}}_i)^T\boldsymbol{\omega}\right)\exp\left(-\frac{1}{2}\boldsymbol{\omega}^T\left(\boldsymbol{\Pi} - \bar{\boldsymbol{\Pi}}\right)\boldsymbol{\omega}\right)\exp\left(j\bar{\boldsymbol{\zeta}}_i^T\boldsymbol{\omega} - \frac{1}{2}\boldsymbol{\omega}^T\bar{\boldsymbol{\Pi}}\boldsymbol{\omega}\right)\right] \\
&\asymp \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M,\mathcal{T}}\left[\exp\left(j\bar{\boldsymbol{\zeta}}_i^T\boldsymbol{\omega} - \frac{1}{2}\boldsymbol{\omega}^T\bar{\boldsymbol{\Pi}}\boldsymbol{\omega}\right)\right] \\
&= \exp\left(j\bar{\boldsymbol{\zeta}}_i^T\boldsymbol{\omega} - \frac{1}{2}\boldsymbol{\omega}^T\bar{\boldsymbol{\Pi}}\boldsymbol{\omega}\right),
\end{aligned}$$

where the third line follows from the fact that, conditioned on the projections and training, the discriminants are jointly Gaussian, and the second-to-last line is justified through the dominated convergence theorem by the fact that characteristic functions are bounded. The final line reveals a Gaussian characteristic function with mean $\bar{\boldsymbol{\zeta}}_i$ and covariance $\bar{\boldsymbol{\Pi}}$, thus the vector $\mathbf{W}$ given $\mathbf{x} \in \mathcal{C}_i$ is asymptotically Gaussian.

D.2 Neyman-Pearson RP-LDA Ensemble Classifier

This section derives the Neyman-Pearson RP-LDA ensemble classifier of Theorem 9 which is based on the asymptotic joint distribution of a set of RP-LDA discriminants.

According to the Neyman-Pearson lemma, for simple hypotheses, the test which rejects the null hypothesis for large values of the ratio of likelihood of observations under the alternative hypothesis to the likelihood of observations under the null hypothesis is the most powerful α-level test. The likelihoods in this case are the joint PDFs of the RP-LDA discriminants under each hypothesis. Although we do not know the exact joint distributions, we do know that the discriminants are asymptotically Gaussian, as stated in Theorem 8. So, asymptotically, the likelihood ratio statistic is

$$\frac{\exp\left(-\frac{1}{2}\left(\mathbf{W} - \bar{\zeta}_1\right)^T\bar{\boldsymbol{\Pi}}^{-1}\left(\mathbf{W} - \bar{\zeta}_1\right)\right)}{\exp\left(-\frac{1}{2}\left(\mathbf{W} - \bar{\zeta}_0\right)^T\bar{\boldsymbol{\Pi}}^{-1}\left(\mathbf{W} - \bar{\zeta}_0\right)\right)} = \exp\left(\bar{m}\mathbf{1}_M^T\bar{\boldsymbol{\Pi}}^{-1}\left(\mathbf{W} - \frac{\bar{m}_0 + \bar{m}_1}{2}\mathbf{1}_M\right)\right), \quad (47)$$

where $\bar{m} := \bar{m}_1 - \bar{m}_0$. We can further simplify this by taking advantage of the special structure of $\boldsymbol{\Pi}^{-1}$. Using the matrix inversion lemma (see Lemma 21 of the tutorial of Müller and Debbah, 2016) and recalling that

$$\bar{\sigma}^2(M) = \frac{1}{M}\bar{\sigma}^2(1) + \left(1 - \frac{1}{M}\right)\bar{\sigma}^2_{M=\infty},$$

we have

$$\bar{\boldsymbol{\Pi}}^{-1} = \frac{1}{\bar{\sigma}^2(1) - \bar{\sigma}^2_{M=\infty}}\left[\mathbf{I}_M - \frac{\bar{\sigma}^2_{M=\infty}}{M\bar{\sigma}^2(M)}\mathbf{1}_M\mathbf{1}_M^T\right].$$

The log of the likelihood ratio statistic in (47) then simplifies to

$$\frac{\bar{m}}{M\bar{\sigma}^2(M)}\sum_{k=1}^{M}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right)-\frac{\bar{m}_0+\bar{m}_1}{2}\right]=\frac{\bar{m}}{\bar{\sigma}^2(M)}W_{\text{disc-avg}}\left(\mathbf{x},\{\mathbf{R}_k\}_{k=1}^{M}\right)-\frac{\bar{m}_1^2-\bar{m}_0^2}{2\bar{\sigma}^2(M)}. \tag{48}$$

For $\bar{m}>0$, (48) is an increasing function of $W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)$. Thus, rejecting the null hypothesis for large values of (48) is equivalent to rejecting the null hypothesis for large values of $W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)$. This condition can easily be verified using the definitions of $\bar{m}_i,\ i=0,1$, in Appendix A.1. The most powerful α-level test according to the Neyman-Pearson lemma is then to classify $\mathbf{x}$ to $\mathcal{C}_1$ if

$$W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)>\eta$$

and to $\mathcal{C}_0$ otherwise, where η is such that

$$P\left\{\left(W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)\,|\,\mathbf{x}\in\mathcal{C}_0\right)>\eta\right\}=\alpha,$$

or, equivalently (asymptotically),

$$\eta=\bar{m}_0+\sqrt{\bar{\sigma}^2(M)}Q^{-1}\left(\alpha\right),$$

where $Q^{-1}(\cdot)$ is the inverse Q-function.

D.3 Asymptotic Distribution of the Discriminant-Averaging RP-LDA Finite Ensemble Classifier Class-Conditional Discriminants

The asymptotic distribution of the single RP-LDA discriminant follows trivially from the proof of the joint asymptotic distribution of M RP-LDA discriminants in Appendix D.1, by setting $M=1$.

For the general case, using the fact that

$$W_{\text{disc-avg}}\left(\mathbf{x},\{\mathbf{R}_k\}_{k=1}^{M}\right)\Big|\,\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\{\mathbf{R}_k\}_{k=1}^{M}=\frac{1}{M}\sum_{k=1}^{M}W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right)\Bigg|\,\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\{\mathbf{R}_k\}_{k=1}^{M}$$

is Gaussian with mean

$$\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1}\left(\boldsymbol{\mu}_i-\frac{\hat{\boldsymbol{\mu}}_0+\hat{\boldsymbol{\mu}}_1}{2}\right)+\ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\asymp\bar{m}_i$$

and variance

$$\begin{aligned}&\frac{1}{M^2}\sum_{k=1}^{M}\operatorname{Var}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}_k\right]\\&\quad+\frac{1}{M^2}\sum_{k\neq j}^{M}\operatorname{Cov}\left[W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right),W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_j\right)|\mathbf{x}\in\mathcal{C}_i,\mathcal{T},\mathbf{R}_k,\mathbf{R}_j\right]\\&\asymp\frac{1}{M}\bar{\sigma}^2(1)+\left(1-\frac{1}{M}\right)\bar{\sigma}^2_{M=\infty},\end{aligned}$$

the asymptotic distribution can be proven by convergence of the relevant characteristic function as in Appendix D.1.

D.4 Asymptotic Distribution of the Discriminant-Averaging RP-LDA Infinite Ensemble Classifier Class-Conditional Discriminants

Using the fact that

$$W_{M=\infty}(\mathbf{x})|\,\mathbf{x}\in\mathcal{C}_i,\mathcal{T} = \hat{\boldsymbol{\mu}}^T\mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]\left(\mathbf{x}-\frac{\hat{\boldsymbol{\mu}}_0+\hat{\boldsymbol{\mu}}_1}{2}\right)+\ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\bigg|\,\mathbf{x}\in\mathcal{C}_i,\mathcal{T}$$

is Gaussian with mean

$$\hat{\boldsymbol{\mu}}^T\mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]\left(\boldsymbol{\mu}_i-\frac{\hat{\boldsymbol{\mu}}_0+\hat{\boldsymbol{\mu}}_1}{2}\right)+\ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\asymp\bar{m}_i$$

and variance

$$\hat{\boldsymbol{\mu}}^T\mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{R}}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1}\right]\hat{\boldsymbol{\mu}}\asymp\bar{\sigma}^2_{M=\infty},$$

the asymptotic distribution can be proven by convergence of the relevant characteristic function as in Appendix D.1.

D.5 Asymptotic Distribution of the Vote-Averaging RP-LDA Ensemble Classifier Class-Conditional Discriminants

The class-conditional discriminant times M,

$$MW_{\text{vote-avg}}(\mathbf{x},\mathbf{R}_1,\dots,\mathbf{R}_M)|\mathbf{x}\in\mathcal{C}_i=\sum_{k=1}^{M}\mathbb{1}\left\{W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right)\right\}|\mathbf{x}\in\mathcal{C}_i,$$

is clearly a sum of correlated Bernoullis. The probability of success for each Bernoulli and the correlations between Bernoullis vary through their random projections. Thus, $MW_{\text{vote-avg}}(\mathbf{x},\mathbf{R}_1,\dots,\mathbf{R}_M)|\mathbf{x}\in\mathcal{C}_i$ is a correlated Binomial random variable with varying probability of success for each trial and varying correlations between trials.

The PMF of $MW_{\text{vote-avg}}(\mathbf{x},\mathbf{R}_1,\dots,\mathbf{R}_M)|\mathbf{x}\in\mathcal{C}_i$, given by

$$P\left\{(MW_{\text{vote-avg}}(\mathbf{x},\mathbf{R}_1,\dots,\mathbf{R}_M)|\mathbf{x}\in\mathcal{C}_i)=m\right\},\ m=0,1,\dots,M,$$

can be obtained exactly as a function of the underlying discriminants

$$\left\{W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_k\right)|\mathbf{x}\in\mathcal{C}_i\right\}_{k=1}^{M}$$

by summing over all probabilities where exactly m of the single RP-LDA discriminants are greater than zero. For example,

$$\begin{aligned}
&P\left\{(MW_{\text{vote-avg}}(\mathbf{x},\mathbf{R}_1,\dots,\mathbf{R}_M)|\mathbf{x}\in\mathcal{C}_i)=1\right\}=\\
&\quad P\left\{W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_1\right)>0,W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_2\right)<0,\dots,W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_M\right)<0|\mathbf{x}\in\mathcal{C}_i\right\}\\
&\quad+P\left\{W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_1\right)<0,W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_2\right)>0,W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_3\right)<0,\dots|\mathbf{x}\in\mathcal{C}_i\right\}+\dots\\
&\quad+P\left\{W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_1\right)<0,\dots,W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_{M-1}\right)<0,W_{\text{RP-LDA}}\left(\mathbf{x},\mathbf{R}_M\right)>0|\mathbf{x}\in\mathcal{C}_i\right\}.
\end{aligned}\tag{49}$$

Moreover, the corresponding CDF is simply a cumulative sum of the PMF.

We have from Theorem 8 that the class-conditional joint distribution of M single RP-LDA discriminants converges to a Gaussian with mean $\bar{\boldsymbol{\zeta}}_i$ and covariance $\bar{\boldsymbol{\Pi}}$. Thus the PMF, and, as a result, the CDF of $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i$, can be computed asymptotically based on the limiting distribution. Formally, let $\bar{\mathbf{W}}_i = [\bar{W}_{1,i}, \ldots, \bar{W}_{M,i}]^T$ denote a Gaussian with $\bar{\boldsymbol{\zeta}}_i$ and covariance $\bar{\boldsymbol{\Pi}}$. Since

$$\lim_{n,p,d\to\infty} P\left\{(MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i) = m\right\} - P\left\{\sum_{k=1}^{M} \mathbb{1}\left\{\bar{W}_{k,i} > 0\right\} = m\right\} = 0$$

through the fact that the left-hand side can be expressed as the limit on a sum of probabilities involving single RP-LDA discriminants (as in (49)) and also the underlying convergence in distribution of these discriminants shown in Theorem 8, then $\forall x \in \mathbb{R}$,

$$\begin{aligned}\lim_{n,p,d\to\infty} & P\left\{(MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i) \leq x\right\} \\ & - P\left\{\sum_{k=1}^{M} \mathbb{1}\left\{\bar{W}_{k,i} > 0\right\} \leq x\right\} = 0, \end{aligned} \tag{50}$$

which is convergence in distribution.

The term $\sum_{k=1}^{M} \mathbb{1}\left\{\bar{W}_{k,i} > 0\right\}$ is a correlated Binomial consisting of M trials. It is straightforward to compute the probability of success of its trials and correlations between its trials as a function of the distribution of $\bar{W}_i$. Because of the structure of $\bar{\boldsymbol{\zeta}}_i$ and $\bar{\boldsymbol{\Pi}}$, the probabilities of success and correlations are constants denoted $\bar{p}_i$ and $\bar{\rho}_i$. It is easy to show that $\bar{\rho}_i$ is always positive.

Note that (50) gives a way to approximate the PMF of $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i$ (and thus the PMF of $W_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \ldots, \mathbf{R}_M)|\mathbf{x} \in \mathcal{C}_i$ which is obtained by simply dividing the PMF arguments by M). Since $\left\{\bar{W}_{k,i}\right\}_{k=1}^{M}$ are identically distributed, computing the asymptotic PMF becomes a counting problem. These computations, however, involve numerical integration. This can become restrictive when M is large, and for that reason we propose the approximation of the asymptotic PMF by Moody's correlated Binomial PMF in Appendix C.3.

Appendix E. Derivation of G-estimators

This section of the appendices derives the G-estimators of the most common metrics of binary classification. These rely on building blocks $\hat{m}_i,\ i = 0, 1,\ \hat{\sigma}^2(1)$, and $\hat{\sigma}^2_{M=\infty}$. As $\hat{m}_i,\ i = 0, 1$ and $\hat{\sigma}^2_{M=\infty}$ were derived in detail by Niyazi et al. (2020a), we consider only $\hat{\sigma}^2(1)$ in the current work. Section E.1 derives $\hat{\sigma}^2(1)$, while Section E.2 proves Theorem 12. Additionally, Section E.3 derives the approximation of the infinite to finite discriminant-averaging RP-LDA ensemble classifier error ratio used to solve for M in the heuristic introduced in Section 4.2.

E.1 Derivation of $\hat{\sigma}^2(1)$

The first step is to derive the quantity $\lim_{\beta \to 0} \tilde{\nu}(\beta)$ as a function of the training (as opposed to true statistics as was done in Appendix A). From the system of equations (29), we have

$$1 - \frac{p}{d} + \frac{1}{\lim_{\beta \to 0} \tilde{\nu}(\beta)} \frac{1}{d} \mathrm{tr} \left\{ \left(\hat{\mathbf{\Sigma}} + \frac{1}{\lim_{\beta \to 0} \tilde{\nu}(\beta)} \mathbf{I}_p \right)^{-1} \right\} = 0. \tag{51}$$

The trace term on the left-hand side can be rewritten as

$$\begin{aligned} \frac{1}{\lim_{\beta \to 0} \tilde{\nu}(\beta)} \frac{1}{d} \mathrm{tr} \left\{ \left(\hat{\mathbf{\Sigma}} + \frac{1}{\lim_{\beta \to 0} \tilde{\nu}(\beta)} \mathbf{I}_p \right)^{-1} \right\} &= \frac{1}{d} \mathrm{tr} \left\{ \left(\lim_{\beta \to 0} \tilde{\nu}(\beta) \mathbf{D} + \mathbf{I}_p \right)^{-1} \right\} \\ &= \frac{1}{d} \sum_{i=1}^{p} \frac{1}{1 + \lim_{\beta \to 0} \tilde{\nu}(\beta) \lambda_i(\hat{\mathbf{\Sigma}})}. \end{aligned}$$

Now consider the monotonically decreasing function,

$$f(x) = 1 - \frac{p}{d} + \frac{1}{d} \sum_{i=1}^{p} \frac{1}{1 + x \lambda_i(\hat{\mathbf{\Sigma}})}.$$

As $x \to 0$, $f(x) \to 1$ and as $x \to \infty$, $f(x) \to 1 - \frac{p}{d} < 0$, when $p > d$, which is the typical use case. Therefore, $f(x)$ has a unique root, x^*, and $\lim_{\beta \to 0} \tilde{\nu}(\beta) = x^*$. Since (51) can be rewritten as

$$1 - \frac{1}{d} \mathrm{tr} \left\{ \hat{\mathbf{\Sigma}} \left(\hat{\mathbf{\Sigma}} + \frac{1}{\lim_{\beta \to 0} \tilde{\nu}(\beta)} \mathbf{I}_p \right)^{-1} \right\} = 0,$$

then overall, the G-estimator of $\lim_{\beta \to 0} \tilde{\nu}(\beta)$, denoted $\hat{\nu}$, is such that

$$1 - \frac{1}{d} \mathrm{tr} \left\{ \hat{\mathbf{\Sigma}} \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\} = 0.$$

Now taking the limit as β goes to zero on the intermediate convergence in (37) and replace $\lim_{\beta \to 0} \tilde{\nu}(\beta)$, denoted $\hat{\nu}$ by its G-estimator $\hat{\nu}$, we have

$$\begin{aligned} \sigma^2(1) \asymp\; & \hat{\boldsymbol{\mu}}^T \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \mathbf{\Sigma} \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}} \\ & + \frac{1}{\hat{\nu}^2} \frac{\frac{1}{d} \mathrm{tr} \left\{ \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \mathbf{\Sigma} \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}}{1 - \frac{1}{d} \mathrm{tr} \left\{ \hat{\mathbf{\Sigma}} \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\mathbf{\Sigma}} \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \hat{\boldsymbol{\mu}}^T \left(\hat{\mathbf{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-2} \hat{\boldsymbol{\mu}}. \end{aligned}$$

Only two terms involve the true statistic $\boldsymbol{\Sigma}$, while the remaining terms are functions of the sample statistics. These two terms can be estimated as

$$\left(\frac{1}{1-\frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\right\}}\right)^2 \hat{\boldsymbol{\mu}}^T\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\mu}}$$
$$\asymp \hat{\boldsymbol{\mu}}^T\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\mu}}$$

and

$$\left(\frac{1}{1-\frac{1}{n-2}\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\right\}}\right)^2 \frac{1}{p}\mathrm{tr}\left\{\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\right\}$$
$$\asymp \frac{1}{p}\mathrm{tr}\left\{\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\frac{1}{\hat{\nu}}\mathbf{I}_p\right)^{-1}\right\}.$$

The proof uses the same techniques used in Section B of Appendix B of the paper by Niyazi et al. (2020b). The same growth regime assumptions stated at the beginning of Section 3 apply here.

E.2 Proof of Theorem 12

First we derive the exact probabilities as follows:

$$\begin{aligned} TPR &= P\left\{W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)>0|\mathcal{T},\{\mathbf{R}_k\}_{k=1}^M,\mathbf{x}\in\mathcal{C}_1\right\}\\ &= \Phi\left(\frac{m_1}{\sqrt{\sigma^2(M)}}\right),\end{aligned}$$

$$\begin{aligned} TNR &= P\left\{W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)<0|\mathcal{T},\{\mathbf{R}_k\}_{k=1}^M,\mathbf{x}\in\mathcal{C}_0\right\}\\ &= \Phi\left(-\frac{m_0}{\sqrt{\sigma^2(M)}}\right),\end{aligned}$$

$$\begin{aligned} FPR &= P\left\{W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)>0|\mathcal{T},\{\mathbf{R}_k\}_{k=1}^M,\mathbf{x}\in\mathcal{C}_0\right\}\\ &= \Phi\left(\frac{m_0}{\sqrt{\sigma^2(M)}}\right),\end{aligned}$$

and

$$\begin{aligned} FNR &= P\left\{W_{\text{disc-avg}}(\mathbf{x},\mathbf{R}_1,\ldots,\mathbf{R}_M)<0|\mathcal{T},\{\mathbf{R}_k\}_{k=1}^M,\mathbf{x}\in\mathcal{C}_1\right\}\\ &= \Phi\left(-\frac{m_1}{\sqrt{\sigma^2(M)}}\right).\end{aligned}$$

We are then able to substitute the G-estimators for each of the quantities m_0, m_1, and $\sigma^2(M)$ in the above expressions by a similar argument to that presented for Lemma 2 in the paper by Niyazi et al. (2020a). The G-estimators for the following quantities are derived in a similar fashion using the G-estimators of the above quantities:

$$\varepsilon = \pi_0 \mathrm{FPR} + \pi_1 \mathrm{FNR},$$

$$\mathrm{PPV} = \frac{\pi_1 \mathrm{TPR}}{\pi_0 \mathrm{FPR} + \pi_1 \mathrm{TPR}},$$

and

$$\mathrm{NPV} = \frac{\pi_0 \mathrm{TNR}}{\pi_0 \mathrm{TNR} + \pi_1 \mathrm{FNR}}.$$

E.3 Derivation of the Heuristic Approximation

Let $\hat{\varepsilon}_{M=\infty}$ and $\hat{\varepsilon}(M)$ denote the G-estimators of the probabilities of misclassification of the infinite and finite discriminant-averaging RP-LDA ensemble classifiers, respectively, where the latter consists of M randomly-projected LDA discriminants. Then,

$$\frac{\hat{\varepsilon}_{M=\infty}}{\hat{\varepsilon}(M)} = \frac{\hat{\pi}_0 \Phi\left(\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2_{M=\infty}}}\right) + \hat{\pi}_1 \Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2_{M=\infty}}}\right)}{\hat{\pi}_0 \Phi\left(\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2(M)}}\right) + \hat{\pi}_1 \Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)}.$$

By assuming equal priors, $\pi_0 = \pi_1$, $n_0 = n_1$, $\hat{\pi}_0 = \hat{\pi}_1$, and $\hat{m}_0 = -\hat{m}_1$. Then

$$\begin{aligned} \frac{\hat{\varepsilon}_{M=\infty}}{\hat{\varepsilon}(M)} &= \frac{\Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2_{M=\infty}}}\right)}{\Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)} \\ &= \frac{Q\left(\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2_{M=\infty}}}\right)}{Q\left(\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)}. \end{aligned} \tag{52}$$

The approximation,

$$Q(x) \approx \frac{1/\sqrt{2\pi}\exp\left(-x^2/2\right)}{x}, x > 0,$$

follows from the right-hand side of the inequality,

$$\frac{x}{1+x^2} 1/\sqrt{2\pi}\exp\left(-x^2/2\right) < Q(x) < \frac{1/\sqrt{2\pi}\exp\left(-x^2/2\right)}{x},\ x > 0,$$

which becomes tighter with increasing x (Borjesson and Sundberg, 1979). Applying this inequality to (52) with $x := \frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2_{M=\infty}}}$ and $y := \frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}$, we obtain

$$\frac{\hat{\varepsilon}_{M=\infty}}{\hat{\varepsilon}(M)} \approx \frac{1/\sqrt{2\pi}\exp\left(-x^2/2\right)/x}{1/\sqrt{2\pi}\exp\left(-y^2/2\right)/y}.$$

Setting this to ψ and solving for y (which is a function of the desired M), we have

$$y^{-1}\exp\left(-y^2/2\right) = \frac{x^{-1}\exp\left(-x^2/2\right)}{\psi}.$$

Squaring and inverting both sides of this equation yields

$$y^2\exp\left(y^2\right) = \psi^2 x^2 \exp\left(x^2\right),$$

which can be solved for $y^2 = \frac{\hat{m}_1^2}{\hat{\sigma}^2(M)}$ by applying the principal branch of the Lambert W function, $W_0(\cdot)$, to both sides (since they are positive). Then

$$y^2 = \frac{\hat{m}_1^2}{\hat{\sigma}^2(M)} = W_0\left(\psi^2 x^2 \exp\left(x^2\right)\right).$$

By making use of the fact that $\hat{\sigma}^2(M) = \frac{1}{M}\hat{\sigma}^2(1) + \left(1 - \frac{1}{M}\right)\hat{\sigma}^2_{M=\infty}$ and $x^2 = \frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}}$, while solving for M, we have

$$M \approx \operatorname{ceil}\left(\frac{\left(\hat{\sigma}^2(1) - \hat{\sigma}^2_{M=\infty}\right) W_0\left(\psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}} \exp\left(\frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}}\right)\right)}{\hat{m}_1^2 - \hat{\sigma}^2_{M=\infty} W_0\left(\psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}} \exp\left(\frac{\hat{m}_1^2}{\hat{\sigma}^2_{M=\infty}}\right)\right)}\right).$$

References

Uri Alon, Naama Barkai, Daniel A. Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

Claudia Beleites and Reiner Salzer. Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry*, 390(5): 1261–1271, 2008.

Florent Benaych-Georges and Romain Couillet. Spectral analysis of the Gram matrix of mixture models. *ESAIM: Probability and Statistics*, 20:217–237, 2016.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, 16, 2003.

Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006. ISBN 9780387310732.

Per Ola Borjesson and Carl-Erik W. Sundberg. Simple approximations of the error function Q(x) for communications applications. *IEEE Transactions on Communications*, 27(3): 639–643, 1979.

Timothy I. Cannings. Random projections: Data perturbation for classification problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(1), 2021.

Timothy I. Cannings and Richard J. Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11: 2079–2107, 2010.

Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

Robert J. Durrant and Ata Kabán. A bound on the performance of LDA in randomly projected data spaces. In *2010 20th International Conference on Pattern Recognition*, pages 4044–4047. IEEE, 2010.

Robert J. Durrant and Ata Kabán. Random projections as regularizers: Learning a linear discriminant from fewer observations than dimensions. *Machine Learning*, 99(2):257–286, 2015.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Todd R Golub, Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller, Mignon L. Loh, James R. Downing, Mark A. Caligiuri, Clara D. Bloomfield, and Eric S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Yaqian Guo, Trevor Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

Walid Hachem, Oleksiy Khorunzhiy, Philippe Loubaton, Jamal Najim, and Leonid Pastur. A new approach for mutual information analysis of large dimensional multi-antenna channels. *IEEE Transactions on Information Theory*, 54(9):3987–4004, 2008.

Walid Hachem, Philippe Loubaton, Jamal Najim, and Pascal Vallet. On bilinear forms based on the resolvent of large random matrices. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 49(1):36–63, 2013.

David J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1): 1–14, 2006.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. ISBN 9780387848846.

Ata Kabán. On compressive ensemble induced regularisation: How close is the finite ensemble precision matrix to the infinite ensemble? In *International Conference on Algorithmic Learning Theory*, pages 617–628. PMLR, 2017.

Ata Kabán. Sufficient ensemble size for random matrix theory-based handling of singular covariance matrices. *Analysis and Applications*, 18(05):929–950, 2020.

Abla Kammoun, Luca Sanguinetti, Merouane Debbah, and Mohamed-Slim Alouini. Asymptotic analysis of RZF in large-scale MU-MIMO systems over rician channels. *IEEE Transactions on Information Theory*, 65(11):7268–7286, 2019.

Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40(3):203–228, 2000.

Qing Mai. A review of discriminant analysis in high dimensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(3):190–197, 2013.

Thomas L. Marzetta, Gabriel H. Tucci, and Steven H. Simon. A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Transactions on Information Theory*, 57(9):6256–6271, 2011.

Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on Medical Imaging*, 35(9):2051–2063, 2016.

Axel Müller and Mérouane Debbah. Random matrix theory tutorial - Introduction to deterministic equivalents. *Traitement du Signal*, 33(2-3):223–248, 2016.

Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri. Asymptotic analysis of an ensemble of randomly projected linear discriminants. *IEEE Journal on Selected Areas in Information Theory*, 1(3):914–930, 2020a.

Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri. Asymptotic analysis of an ensemble of randomly projected linear discriminants, 2020b. URL `https://arxiv.org/abs/2004.08217`.

Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri. Weight vector tuning and asymptotic analysis of binary linear classifiers. *IEEE Open Journal of Signal Processing*, 2022.

Devis Peressutti, Wenjia Bai, Thomas Jackson, Manav Sohal, Aldo Rinaldi, Daniel Rueckert, and Andrew King. Prospective identification of CRT super responders using a motion atlas and random projection ensemble learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 493–500. Springer, 2015.

Alon Schclar and Lior Rokach. Random projection ensemble classifiers. In *International Conference on Enterprise Information Systems*, pages 309–316. Springer, 2009.

Alok Sharma and Kuldip K. Paliwal. Linear discriminant analysis for the small sample size problem: An overview. *International Journal of Machine Learning and Cybernetics*, 6 (3):443–454, 2015.

Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini. High-dimensional linear discriminant analysis classifier for spiked covariance model. *Journal of Machine Learning Research*, 21:1–24, 2020.

Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2): 203–209, 2002.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. ISBN 9780511802256.

Gary Witt. Moody's correlated binomial default distribution. *Moody's Investor Service, Special Report, August*, 2004.

Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. Recent advances of large-scale linear classification. *Proceedings of the IEEE*, 100(9):2584–2603, 2012.