Asymptotic Performance Analysis of the Randomly-Projected RLDA Ensemble Classifier

Thesis by

Lama B. Niyazi

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

June, 2019

EXAMINATION COMMITTEE PAGE

The thesis of Lama B. Niyazi is approved by the examination committee.

Committee Chairperson: Mohamed-Slim Alouini
Committee Co-Chair: Tareq Y. Al-Naffouri
Committee Members: Abla Kammoun, Hayssam Dahrouj

ABSTRACT

Asymptotic Performance Analysis of the Randomly-Projected RLDA Ensemble Classifier

Lama B. Niyazi

Reliability and computational efficiency of classification error estimators are critical factors in classifier design. In a high-dimensional data setting where data is scarce, the conventional method of error estimation, cross-validation, can be very computationally expensive. In this thesis, we consider a particular discriminant analysis type classifier, the Randomly-Projected RLDA ensemble classifier, which operates under the assumption of such a 'small sample' regime. We conduct an asymptotic study of the generalization error of this classifier under this regime, which necessitates the use of tools from the field of random matrix theory. The main outcome of this study is a deterministic function of the true statistics of the data and the problem dimension that approximates the generalization error well for large enough dimensions. This is demonstrated by simulation on synthetic data. The main advantage of this approach is that it is computationally efficient. It also constitutes a major step towards the construction of a consistent estimator of the error that depends on the training data and not the true statistics, and so can be applied to real data. An analogous quantity for the Randomly-Projected LDA ensemble classifier, which appears in the literature and is a special case of the former, is also derived. We motivate its use for tuning the parameter of this classifier by simulation on synthetic data.

ACKNOWLEDGEMENTS

I owe the completion of this work to many people whose help, support, and influence, I must acknowledge here.

First of all, thanks to my advisors, Professor Slim and Professor Tareq, for their support and guidance since I was an undergraduate and for allowing me many opportunities to learn and grow. I want to thank Dr. Abla for her patience and dedication in mentoring me and for teaching me random matrix theory. I would also like to thank Professor Hayssam for facilitating my entry into research and KAUST, Professor Omar Kittaneh for his encouragement, support, and love of math, and Dr. Mohamed Saad and Dr. Anas for teaching me so much over the course of my senior project. Also thanks to Bayan Al-Oquibi for introducing me to LaTeX over 3 years ago.

I am grateful to Professor Ahmed Sultan for having been my teacher and for his sincere advice to me during a difficult time. I also want to express my thanks to the students of the stochastic processes class of Spring 2019, for which I was a teaching assistant. I learned so much from interacting with you all.

To Somayah, Sarah, Wafa, Nojood, Sumyyah, Horiya, Hebatallah, Sondos, Zainab, Mashail, and Khadija: I am so blessed to have had each of you in my life and I hope to always cherish our friendship. I want to thank my siblings, my mother, and my father for their support and sacrifice, my mother especially for her love and purity, and my father for his strength of character and love of learning and knowledge. I want to thank my uncle for his kindheartedness and selflessness and my grandmother, may Allah have mercy on her, for her love and constant prayers. To anyone who has ever wished me well, I wish you back a thousand times better.

This is by the Grace of my Lord and all praise is to Him.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

In this chapter, we introduce this work by giving a background on error estimation for classifiers and the issues of error estimation under certain settings. Through this, we motivate an asymptotic analysis of a certain classifier using random matrix theory. We state the objectives of this work and lay out some notation and definitions towards the end of the chapter.

1.1 Background and Objectives

A critical measure of any classifier's performance is how well it generalizes to unseen data. In the absence of an exact knowledge of training set distributions, its generalization error can only be estimated from data. The quality of this error estimate is not only important for it to serve as a reliable indicator of the performance of the classifier, but also to be able to tune the classifier parameters, which are selected on the basis of minimizing generalization error. Obtaining reliable estimators of the generalization error is therefore an essential part of classifier design .

In practice, a part of the data is partitioned for testing from which an estimate of the classifier's generalization error may be computed. The estimate obtained from the testing data can be very good; it is both unbiased and its variance decays with an increasing number of test samples [1]. It is when data is scarce that the issue of error estimation becomes apparent. This is especially true of high dimensional data, as the curse of dimensionality necessitates an exponential increase in samples with increasing dimensions of the data [2]. In such cases, small-sample approaches are resorted to, in which the training data is recycled for error estimation. Within the machine learning community, the most popular among these approaches is cross-validation, but the cross-validation approach suffers from inadequacies, mainly high variance of the resulting estimator [3] and high computational cost.

In this thesis, we study the performance of a particular discriminant analysis type classifier which we refer to as the Randomly-Projected Regularized Linear Discriminant Analysis (RP-RLDA) ensemble classifier. To do this, we conduct an asymptotic analysis of its generalization error. The main outcome of the analysis is a deterministic function of the true statistics of the training data and problem dimension that is intended to serve as an approximation of the generalization error. This 'deterministic equivalent' (DE) is relatively quick to compute and is shown, by simulation on synthetic data, to approximate the empirical error very well.

As we shall see in Chapter 3, the RP-RLDA ensemble classifier, assumes a high-dimensional data setting, where the number of training samples is close to the dimensionality of the data. Asymptotically, this corresponds to the regime where the the number of training samples and the dimensionality of the data grow at a constant rate to each other. The field of random matrix theory facilitates the derivation of limits of expressions involving random matrices whose dimensions are subject to this regime. For our problem, given a closed form expression for the generalization error of a classifier, random matrix theory can be applied to yield an unbiased and deterministic limit of this error, in terms of the true statistics of the data. This idea is not new. The first application of random matrix theory to the family of discriminant analysis classifiers is Zollanvari's work [4] in which he derives a deterministic limit for the two-class Linear Discriminant Analysis (LDA) classifier in terms of the true statistics of the data. He then uses this result to construct a consistent estimator of the error that makes use of the training data. A comparison of the estimator thus derived shows favorable performance against traditional estimators such as bolstered resubstitution, bootstrap, and cross-validation. This is followed by similar work in which consistent error estimators are derived using the same approach for a variety of classifiers. For example, Elkhalil's work [5] considers the asymptotic analysis of the Regularized Linear Discriminant Analysis (RLDA) and Regularized Quadratic Discriminant Analysis (RQDA) classifiers. Similarly, Yang studies the Regularized Discriminant Analysis (RDA) classifier in [6]. In this thesis, we apply this methodology to obtain a deterministic equivalent for the error of the RP-RLDA ensemble classifier. This is the first step to constructing a consistent estimator of the error; the next step is left for future work. The closest work to the present one is that of Durrant [7] in which he proposes and analyzes the Randomly-Projected LDA (RP-LDA) ensemble, a special case of the RP-RLDA ensemble. Durrant derives bounds on the RP-LDA ensemble classifier's generalization error, whereas we derive a DE for the generalization error of the more general case of this classifier. We also recover the DE for the generalization error of the RP-LDA ensemble classifier as a limiting case of ours, and use it to tune its parameter.

The main objectives of this work are to

- derive a deterministic equivalent of the generalization error of the RP-RLDA ensemble classifier under the small-sample regime and motivate its use as an approximation of the generalization error in the finite scenario by simulation on synthetic data

- derive a deterministic equivalent of the generalization error of the RP-LDA ensemble classifier as a limiting case of the RP-RLDA ensemble classifier

- motivate the use of the deterministic equivalent to tune the parameter of the RP-LDA ensemble classifier by simulation on synthetic data

1.2 Some Notation and Definitions

Throughout this work, 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. The operator $\text{diag}(\cdot)$, when applied to a matrix, outputs its diagonal elements as a vector, and when applied to a vector, outputs a matrix with the vector elements along its diagonal. The operator $\lceil\cdot\rceil$ rounds its argument up to the nearest integer. The symbols $\mathcal{O}(\cdot)$ and $o(\cdot)$ are the standard Big-O and Little-O notation. Almost-sure convergence is denoted by $\xrightarrow{\text{a.s.}}$. The function $\Phi(\cdot)$ denotes the standard Gaussian CDF. Finally, for a matrix $\mathbf{A}$, the resolvent $\mathbf{Q}$ is defined as the matrix $\mathbf{Q}(z) = (\mathbf{A} - z\mathbf{I})^{-1}$, $\forall z \in \mathbb{C}$ except for the eigenvalues of $\mathbf{A}$.

Chapter 2

The Randomly Projected RLDA Ensemble Classifier

This chapter introduces the *Randomly Projected RLDA Ensemble Classifier*. First, the classification setting for this problem and the assumptions on the data for which the classifier is designed are established. Section 2.2 then introduces the concepts of random projections and randomly-projected classifiers and motivates their use for dimensionality reduction, particularly for discriminant analysis type classifiers. Finally, Section 2.3 presents the classifier that will be the focus of our analysis in Chapter 3.

2.1 Classification Setting and Assumptions

For the current work, we consider binary classification under a supervised setting. We assume the following setup.

A data point $\mathbf{x} \in \mathbb{R}^p$ belongs to one of two classes $\mathcal{C}_0$ and $\mathcal{C}_1$. Conditioned on its class, $\mathbf{x}$ is assumed to be Gaussian distributed, with distinct means and a common covariance between the two classes, as follows:

$$\begin{aligned} \mathbf{x}|\mathbf{x} \in \mathcal{C}_0 &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}\right) \\ \mathbf{x}|\mathbf{x} \in \mathcal{C}_1 &\sim \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right) \end{aligned} \tag{2.1}$$

We are given a training set of n instances of training data distributed as (2.1). The training set consists of pairs of data points and their labels. More formally, the set of n training data points $\mathcal{T}$ is defined as

$$\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$$

where $\mathbf{x}_i$ is a data point and $\mathbf{y}_i \in \{0,1\}$ is its corresponding label. This set contains n_0 and n_1 sample points from $\mathcal{C}_0$ and $\mathcal{C}_1$ respectively. We assume the true statistics of the data $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$ are unknown and therefore must be estimated from the training data. This is implicit in the construction of the classifier. We denote by $h(\mathbf{x}_q)$ the perfect decision rule (with no misclassification) to which any classifier must aspire for a query point $\mathbf{x}_q$ also distributed as (2.1). In non-trivial cases, the decision rules for the classifiers which follow can only be approximations to this function and therefore are accordingly denoted by $\hat{h}(\mathbf{x}_q)$ and an identifying subscript.

2.2 Random Projections and Randomly-Projected Classifiers

Random projection is a non-adaptive dimensionality reduction technique in which data points are multiplied by a matrix $\mathbf{R} \in \mathbb{R}^{d \times p}$, with $d < n$, whose entries are generated i.i.d. from a zero-mean Gaussian distribution [7]. In particular, we define our random projection matrix as having i.i.d. entries $R_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$, $\forall i, j$. We specify a normalized variance of $\frac{1}{d}$ for each entry in order to facilitate the application of random matrix theory results later.

Denote by $\mathbf{X}_0 \in \mathbb{R}^{p \times n_0}$ the matrix having the vectors $\mathbf{x}_i \in \mathcal{C}_0$ as its successive columns, and similarly denote by $\mathbf{X}_1 \in \mathbb{R}^{p \times n_1}$ the matrix having the vectors $\mathbf{x}_i \in \mathcal{C}_1$ as its successive columns. The term *randomly-projected classifier* refers to the classifier being trained on data that has been randomly projected onto the column space of $\mathbf{R}$; instead of being learned on $\mathbf{X}_0$ and $\mathbf{X}_1$, the classifier is learned on $\mathbf{R}\mathbf{X}_0$ and $\mathbf{R}\mathbf{X}_1$. For example, the Linear Discriminant Analysis (LDA) classifier with the the following decision rule

$$\hat{h}_{\text{LDA}}(\mathbf{x}_q) := \mathbb{1}\left\{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} > 0\right\} \tag{2.2}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, is learned on $\mathbf{X}_0$ and $\mathbf{X}_1$ by computing the maximum likelihood estimates $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\pi}_0$, and $\hat{\pi}_1$ of the true statistics $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$, and prior probabilities π_0 and π_1, which are unknown. These estimates are the sample means, pooled sample covariance matrix, and the prior probability estimates respectively, defined as

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_0 &= \frac{1}{n_0}\mathbf{X}_0\mathbf{1} \\
\hat{\boldsymbol{\mu}}_1 &= \frac{1}{n_1}\mathbf{X}_1\mathbf{1} \\
\hat{\boldsymbol{\Sigma}} &= \frac{(n_0 - 1)\hat{\boldsymbol{\Sigma}}_0 + (n_1 - 1)\hat{\boldsymbol{\Sigma}}_1}{n_0 + n_1 - 2} \\
\hat{\pi}_0 &= \frac{n_0}{n} \\
\hat{\pi}_1 &= \frac{n_1}{n}
\end{aligned} \tag{2.3}$$

where $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n_0 - 1}\left(\mathbf{X}_0 - \boldsymbol{\mu}_0\mathbf{1}^T\right)\left(\mathbf{X}_0 - \boldsymbol{\mu}_0\mathbf{1}^T\right)^T$ and $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1 - 1}\left(\mathbf{X}_1 - \boldsymbol{\mu}_1\mathbf{1}^T\right)\left(\mathbf{X}_1 - \boldsymbol{\mu}_1\mathbf{1}^T\right)^T$. Projecting the training data as $\mathbf{R}\mathbf{X}_0$ and $\mathbf{R}\mathbf{X}_1$ results in the following statistic estimates as a function of the old estimates

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_0^{\text{RP}} &= \mathbf{R}\hat{\boldsymbol{\mu}}_0 \\
\hat{\boldsymbol{\mu}}_1^{\text{RP}} &= \mathbf{R}\hat{\boldsymbol{\mu}}_1 \\
\hat{\boldsymbol{\Sigma}}^{\text{RP}} &= \mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T
\end{aligned} \tag{2.4}$$

The resulting decision rule for the Randomly-Projected LDA (RP-LDA) classifier, obtained by simply plugging the new estimates into the LDA decision rule, is

$$\hat{h}_{\text{RP-LDA}}(\mathbf{x}_q) := \mathbb{1}\left\{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} > 0\right\} \quad (2.5)$$

RP-LDA may be especially useful when working in a small-sample regime for which $n \leq p$. In that case, $\hat{\boldsymbol{\Sigma}}$ is singular. Typically, LDA is adapted to this regime either by substituting the pseudoinverse of $\hat{\boldsymbol{\Sigma}}$ for $\hat{\boldsymbol{\Sigma}}^{-1}$, with poor performance, or by adding a regularization term $\gamma\mathbf{I}$, $\gamma > 0$, to $\hat{\boldsymbol{\Sigma}}$ before inverting, yielding the popular variant Regularized Linear Discriminant Analysis (RLDA) [8]. Randomly projecting the training data achieves a similar effect when $d \leq \text{rank}(\hat{\boldsymbol{\Sigma}})$ (noting that $\text{rank}(\hat{\boldsymbol{\Sigma}}) \leq n - 2$). Under this condition, $\hat{\boldsymbol{\Sigma}}^{\text{RP}}$ is guaranteed to be invertible since $\mathbf{R}$ is almost surely of rank d [7]. But random projection can be applied to more than just discriminant analysis type classifiers in which an ill-conditioned estimate of the covariance is involved; there exists a result in the field of compressed sensing that states that, under some sparsity conditions and conditions on $\mathbf{R}$, a vector $\mathbf{x}$ can be perfectly reconstructed from $\mathbf{R}\mathbf{x}$, its reduced dimension representation [7]. For classification, this translates to no loss of information between the original and projected data. Another result is the Johnson-Lindenstrauss lemma which states that, with high probability, vector norms are preserved under projection for d greater than a functional threshold of n [7]. The implication here is that distances between training vectors are preserved under random projection. Both these results motivate the use of randomly projected classifiers, as they suggest that, at the very least, randomly projected classifiers can maintain the accuracy of their unprojected counterparts, while reaping the benefits of a lower computational complexity because of the reduced dimensions of the data involved. In fact, performance often deteriorates in practice, at least in the case of discriminant analysis type classifiers. This is why we look into ensembles of randomly-projected classifiers in the next section.

2.3 The RP-RLDA Ensemble Classifier

An ensemble of supervised classifiers consists of multiple classifiers learned on $\mathcal{T}$ and combined in some way as to reach a net decision. The literature has shown, both empirically and theoretically, that such strategies achieve better accuracy than single models [9]. When it comes to single randomly-projected classifiers, intuitively, a particular projection may be 'bad' in the sense that it may project the training data so that it is not as separable in the projected space as it would be had another projection been used. The mean separation between the class clusters may not be as large as could be and/or the clusters may not be as compact as could be. As is remarked in [10], the class structure of the data is frequently destroyed. This results in lower classification accuracy than could be potentially attained, and with one projection, the matter is essentially random. In fact, simulations show that a single random projection with LDA as a base classifier may perform worse than LDA alone, even in a small-sample regime for which we make use of the pseudo-inverse modified LDA. This prompts us to look into classifier ensembles comprised of multiple projections.

In [8], Durrant proposes an RP-LDA ensemble classifier in which M individual RP-LDA classifers each learned using a different random projection are combined by averaging their discriminants. The decision is then made based on this aggregated discriminant. The decision rule for this RP-LDA ensemble is

$$\hat{h}^{\text{ens}}_{\text{RP-LDA}}(\mathbf{x}_q) := \mathbb{1}\left\{ \frac{1}{M}\sum_{i=1}^{M}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbf{R}_i^T (\mathbf{R}_i \hat{\Sigma} \mathbf{R}_i^T)^{-1} \mathbf{R}_i \left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} > 0 \right\} \tag{2.6}$$

He goes on to derive an upper bound on the generalization error of this classifier. A related work is [10], in which Cannings and Samworth implement a more sophisticated randomly projected ensemble classifier. For this classifier, the distribution of the projections is chosen based on $\mathcal{T}$. In addition, the projections undergo a selection process, based on testing error, to be included in the ensemble. Finally, the combining method is an averaging that is subjected to a variable threshold. The analysis is done with LDA, QDA, and $k$nn as base classifiers. Again, the authors provide bounds on the generalization errors of these classifiers. Between the two of these, we choose to further investigate Durrant's formulation [7] because of the relative tractability of its decision rule (2.6).

As an extension of the work in [7], we conduct an asymptotic study of the RP-LDA ensemble in the regime where n is on the order of p (and on the order of d), in order to represent the small-sample regime. This regime requires the use of tools from random matrix theory. To pursue the analysis, we first formulate an RP-LDA ensemble, with the addition of a parameter $\gamma \in [0, \infty)$, having the decision rule

$$\hat{h}^{\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) := \mathbb{1}\left\{ \frac{1}{M} \sum_{i=1}^{M} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbf{R}_i^T (\mathbf{R}_i \hat{\Sigma} \mathbf{R}_i^T + \gamma \mathbf{I})^{-1} \mathbf{R}_i \left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} > 0 \right\} \tag{2.7}$$

This is equivalent to a randomly projected ensemble with RLDA as a base classifier. The addition of the regularization term is a necessary technicality in order to be able to apply random matrix theory. The main outcome of the asymptotic analysis is a deterministic function of the true class statistics $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$ and dimensions n, p, and d that serves as an estimate of the generalization error in the finite scenario, when n, p, and d are large enough. In the terminology of random matrix theory, this is called a deterministic equivalent. To obtain the deterministic equivalent for the generalization error of the RP-LDA ensemble classifier, we must obtain a deterministic equivalent of the generalization error of the RP-RLDA ensemble classifier and take the limit as $\gamma \to 0$. We are in fact carrying out an asymptotic analysis of the RP-RLDA ensemble classifier. However, this works out in our favor, as the γ in (2.7) allows us an extra degree of freedom compared to (2.6) which we may optimize to further enhance performance. We therefore consider the RP-RLDA ensemble classifier for asymptotic analysis in what remains of this thesis, of which the RP-LDA ensemble is a special case.

Chapter 3

Asymptotic Performance Analysis of the RP-RLDA Ensemble

In this chapter, as a first step towards the asymptotic analysis of the RP-RLDA ensemble classifier, we construct the RP-RLDA infinite ensemble and derive an expression for its generalization error. This approximates the RP-RLDA ensemble classifier generalization error when M is large. We then formulate the problem of deriving its deterministic equivalent and finally present the results.

3.1 The RP-RLDA Infinite Ensemble

In Chapter 2, we introduced the RP-RLDA ensemble. In this chapter, we pursue an asymptotic analysis of its performance, by obtaining an expression for its generalization error as n, p, and d grow to infinity. To make the discriminant more tractable, we let M, the number of random projections in the ensemble, grow as well (also done in [7]). We call this mathematical object the *RP-RLDA infinite ensemble.* Although it cannot be physically realized, we can come very close when M is large enough (typically around 100 to 200). In this section, the RP-RLDA infinite ensemble is constructed and its generalization error is derived.

3.1.1 Construction

Let $\hat{W}^{\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q)$ denote the discriminant of (2.7). In the limit as $M \to \infty$, for fixed d, p, and n, the discriminant becomes

$$\begin{aligned}
&\lim_{M\to\infty} \hat{W}^{\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) \\
&\quad = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbf{R}_i^T (\mathbf{R}_i \hat{\Sigma} \mathbf{R}_i^T + \gamma \mathbf{I})^{-1} \mathbf{R}_i \right) \left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \\
&\quad = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R} \hat{\Sigma} \mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right] \left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \qquad (3.1)
\end{aligned}$$

Since the sequence of random realizations $\{\mathbf{R}_i\}_{i=1}^{M}$ are i.i.d., the sequence $\{\mathbf{R}_i^T (\mathbf{R}_i \hat{\Sigma} \mathbf{R}_i^T + \gamma \mathbf{I})^{-1} \mathbf{R}_i\}_{i=1}^{M}$ is also i.i.d., and thus the second step follows from the law of large numbers. Let $\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q)$ denote this discriminant

$$\begin{aligned}
&\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) \\
&\quad = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right] \left(\mathbf{x}_q - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \qquad (3.2)
\end{aligned}$$

We define the RP-RLDA infinite ensemble as the classifier having the following decision rule

$$\hat{h}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) := \mathbb{1}\{\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) > 0\} \qquad (3.3)$$

3.1.2 Generalization Error

We now characterize the probability of misclassification of a query point $\mathbf{x}_q$ by the randomly projected RLDA infinite ensemble described in Section 3.1.1. This is also referred to as the generalization error of the classifier. Denote by $\varepsilon^{\mathcal{T}}$ the probability of misclassification given the training data $\mathcal{T}$. By the law of total probability

$$\varepsilon^{\mathcal{T}} = \pi_0 \varepsilon_0^{\mathcal{T}} + \pi_1 \varepsilon_1^{\mathcal{T}} \tag{3.4}$$

where $\varepsilon_0^{\mathcal{T}}$ is the probability of misclassification given the point belongs to $\mathcal{C}_0$ and conditioned on the training data $\mathcal{T}$. Similarly, $\varepsilon_1^{\mathcal{T}}$ is the probability of misclassification given the point belongs to $\mathcal{C}_1$ and conditioned on the training data $\mathcal{T}$. In terms of the discriminant $\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q)$,

$$\varepsilon_0^{\mathcal{T}} = \mathbb{P}[\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) > 0 | \mathbf{x}_q \in \mathcal{C}_0, \mathcal{T}] \tag{3.5}$$

$$\varepsilon_1^{\mathcal{T}} = \mathbb{P}[\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) < 0 | \mathbf{x}_q \in \mathcal{C}_1, \mathcal{T}] \tag{3.6}$$

Conditioned on the classes and the training data, the discriminant is a Gaussian random variable. More specifically,

$$\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) | \mathbf{x}_q \in \mathcal{C}_0, \mathcal{T} \sim \mathcal{N}(m_0, \sigma^2) \tag{3.7}$$

$$\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q) | \mathbf{x}_q \in \mathcal{C}_1, \mathcal{T} \sim \mathcal{N}(m_1, \sigma^2) \tag{3.8}$$

where

$$m_0 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right] \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \tag{3.9}$$

$$m_1 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right] \left(\boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \tag{3.10}$$

and

$$\sigma^2 =$$
$$(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right] \boldsymbol{\Sigma} \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right]^T (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) \tag{3.11}$$

From this it can be shown that

$$\varepsilon_0^{\mathcal{T}} = \mathbb{P}[\hat{W}_{\text{RP-RLDA}}^{\infty-\text{ens}}(\mathbf{x}_q) > 0 | \mathbf{x}_q \in \mathcal{C}_0, \mathcal{T}] = \Phi\left(\frac{m_0}{\sqrt{\sigma^2}}\right) \tag{3.12}$$

$$\varepsilon_1^{\mathcal{T}} = \mathbb{P}[\hat{W}_{\text{RP-RLDA}}^{\infty-\text{ens}}(\mathbf{x}_q) < 0 | \mathbf{x}_q \in \mathcal{C}_0, \mathcal{T}] = \Phi\left(\frac{-m_1}{\sqrt{\sigma^2}}\right) \tag{3.13}$$

Thus the exact probability of misclassification of the randomly projected RLDA infinite ensemble classifier conditioned on $\mathcal{T}$ is

$$\varepsilon^{\mathcal{T}} = \pi_0 \Phi\left(\frac{m_0}{\sqrt{\sigma^2}}\right) + \pi_1 \Phi\left(\frac{-m_1}{\sqrt{\sigma^2}}\right) \tag{3.14}$$

The unconditioned probability, which we denote by ε is obtained by simply taking the expectation of (3.14) over the training $\mathcal{T}$

$$\varepsilon = \mathbb{E}_{\mathcal{T}}\left[\pi_0 \Phi\left(\frac{m_0}{\sqrt{\sigma^2}}\right) + \pi_1 \Phi\left(\frac{-m_1}{\sqrt{\sigma^2}}\right)\right] \tag{3.15}$$

3.2 DE of the Generalization Error

As the main outcome of this thesis, we derive a deterministic sequence $\bar{\varepsilon}$ of n, p, and d (a deterministic equivalent) such that

$$\varepsilon - \bar{\varepsilon} \xrightarrow{\text{a.s.}} 0 \tag{3.16}$$

and $\bar{\varepsilon}$ is a function of the true statistics of the data. In this section, we show how to build the deterministic equivalent of ε given deterministic equivalents of m_0, m_1 and σ^2. We then present these deterministic equivalents. The details of their derivation are left for the appendix.

Given deterministic sequences of d, p, and n denoted $\bar{m}_0$, $\bar{m}_1$, and $\bar{\sigma}^2$ such that

$$\begin{aligned} m_0 - \bar{m}_0 &\xrightarrow{\text{a.s.}} 0 \\ m_1 - \bar{m}_1 &\xrightarrow{\text{a.s.}} 0 \\ \sigma^2 - \bar{\sigma}^2 &\xrightarrow{\text{a.s.}} 0 \end{aligned} \tag{3.17}$$

then by the continuous mapping theorem and other properties of almost sure convergence,

$$\varepsilon^{\mathcal{T}} - \left(\pi_0 \Phi\left(\frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2}}\right) + \pi_1 \Phi\left(\frac{-\bar{m}_1}{\sqrt{\bar{\sigma}^2}}\right)\right) \xrightarrow{\text{a.s.}} 0 \tag{3.18}$$

To find the unconditioned misclassification probability, simply take the expectation over the training data

$$\varepsilon - \mathbb{E}_{\mathcal{T}}\left[\left(\pi_0 \Phi\left(\frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2}}\right) + \pi_1 \Phi\left(\frac{-\bar{m}_1}{\sqrt{\bar{\sigma}^2}}\right)\right)\right] \xrightarrow{\text{a.s.}} 0 \tag{3.19}$$

Generally, the limit of the expectation of a random variable is not simply the expectation of its limit, but in this case it is. This is justified by the bounded convergence theorem, since ε is a probability and so $\varepsilon < 1$. Since $\pi_0 \Phi\left(\frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2}}\right) + \pi_1 \Phi\left(\frac{-\bar{m}_1}{\sqrt{\bar{\sigma}^2}}\right)$ is deterministic, we have the result

$$\varepsilon - \left(\pi_0 \Phi\left(\frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2}}\right) + \pi_1 \Phi\left(\frac{-\bar{m}_1}{\sqrt{\bar{\sigma}^2}}\right)\right) \xrightarrow{\text{a.s.}} 0 \tag{3.20}$$

The deterministic equivalent $\bar{\varepsilon} = \pi_0 \Phi\left(\frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2}}\right) + \pi_1 \Phi\left(\frac{-\bar{m}_1}{\sqrt{\bar{\sigma}^2}}\right)$ serves as a large-scale approximation of the misclassification probability at finite d, p, n, and M. To compute it, we must have at hand $\bar{m}_0$, $\bar{m}_1$, and $\bar{\sigma}^2$. These are presented in what follows.

Firstly, let us define some quantities which appear in the deterministic equivalents. In the following, the matrices $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ are eigendecomposed as $\mathbf{\Sigma} = \mathbf{V}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{V}^T$ and $\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ respectively. The quantity $\bar{m}(-\gamma)$ is the deterministic equivalent of the normalized trace of the resolvent of the matrix $\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T$, and is defined as

$$\bar{m}(z) = \frac{1}{d}\text{tr}\left\{-\frac{1}{z\left(1+\frac{p}{d}\tilde{\delta}(z)\right)}\mathbf{I}_d\right\} \tag{3.21}$$

The term $\tilde{\delta}(z)$ appearing in (3.21) is a standard result that can be computed in an iterative fashion from a system of equations, however this system depends on the sample statistics. As we wish $\bar{\varepsilon}$ to characterize the generalization error in terms of the true statistics, we resort to an alternative method of computation, which is fully outlined in Chapter 4. The quantity $\mathbf{T}(z)$ is defined as

$$\mathbf{T}(z) = -\frac{1}{z}(\mathbf{I}_p + \tilde{e}(z)\mathbf{D}_{\mathbf{\Sigma}})^{-1} \tag{3.22}$$

The quantity $\tilde{e}(z)$ appearing in (3.22) can be computed iteratively from the system of equations

$$e(z) = \frac{1}{n}\text{tr}\left\{-\frac{1}{z}\mathbf{D}_{\mathbf{\Sigma}}\left(\mathbf{I}_p + \tilde{e}(z)\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\} \tag{3.23}$$

$$\tilde{e}(z) = -\frac{1}{z(1+e(z))} \tag{3.24}$$

In addition to these definitions, we must state the assumptions on the growth of n, p, and d on which the analysis is based. We define by the following conditions our growth regime as n, p, $d \to \infty$

(a) $\frac{n}{p} \to c \in (0, \infty)$

(b) $\frac{d}{p} \to c' \in (0, \infty)$

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

(d) $\limsup ||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2 < \infty$

(e) $\limsup ||\boldsymbol{\Sigma}||_2 < \infty$

Assumptions (a) to (d) are made in order to ensure non-trivial asymptotic classification, while assumptions (d) and (e) are technicalities stemming from the use of random matrix theory tools. Assuming this growth regime, the deterministic equivalents, $\bar{m}_0$, $\bar{m}_1$, and $\bar{\sigma}^2$, as a function of the true statistics, are given by

$$\begin{aligned}\bar{m}_0 &= \\ &\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{V}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right) \mathbf{V}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\ &\quad + \frac{1}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right) \mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\right\} + \ln\frac{\pi_1}{\pi_0} \end{aligned} \tag{3.25}$$

$$\begin{aligned}\bar{m}_1 &= \\ &\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{V}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right) \mathbf{V}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ &\quad + \frac{1}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right) \mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\right\} + \ln\frac{\pi_1}{\pi_0} \end{aligned} \tag{3.26}$$

and

$$\begin{aligned}\bar{\sigma}^2 &= \\ &\kappa(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{V}\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right) \mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right) \mathbf{V}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ &\quad + \kappa\left(\frac{1}{n_0} + \frac{1}{n_1}\right) \mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right) \mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\right\} \end{aligned} \tag{3.27}$$

respectively, where $\kappa = \frac{(n/p)^2}{(n/p)^2 - (n/p)\frac{1}{(1+e(z))^2}\Omega}$, with $\Omega = \frac{1}{p}\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{T}(z)\, \mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}(z)\right\}$ and $z = -\frac{1}{\bar{m}(-\gamma)}$. The appendix includes the derivation of these quantities in full as well as simulations confirming the almost-sure convergence claimed in (3.17) and (3.20).

That outlines the basic form of the deterministic equivalents $\bar{m}_0$, $\bar{m}_1$, $\bar{\sigma}^2$, and $\bar{\varepsilon}$. Chapter 4 will describe in detail how to compute these deterministic equivalents and present simulations to motivate the use of $\bar{\varepsilon}$ as an approximation of the generalization error ε.

Chapter 4

Simulations

In this chapter, we present simulations comparing the RP-RLDA ensemble classifier's empirical error over a separate testing set with the deterministic equivalent of its generalization error presented in Chapter 3. Additionally, we present simulations to suggest how to use this DE to optimize the projected dimension d in the special case of the RP-LDA ensemble. Prior to all of this, we lay out the procedure to compute the DEs for the RP-RLDA ensemble and RP-LDA ensemble classifiers respectively.

4.1 Computing the DEs in Practice

To compute (3.25), (3.26), and (3.27), we must first compute $\bar{m}(-\gamma)$. As mentioned in Chapter 3, this depends on the computation of $\tilde{\delta}(-\gamma)$. The direct method of computing this quantity is undesirable as it depends on the sample statistics. We describe an alternative method of computation in Section 4.1.1. Once $\bar{m}(-\gamma)$ has been computed, it can be used to compute $\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)$. The special case of $\gamma \to 0$ which recovers the RP-LDA ensemble formulated in [7] is obtained by computing $\lim_{\gamma \to 0} \bar{m}(-\gamma)$. This is detailed in Section 4.1.2. This chapter outlines the procedures; details of derivation are left for the appendix.

4.1.1 The RP-RLDA Ensemble

In this section, we state the procedures to compute $\bar{m}(-\gamma)$ in each of two cases of the true covariance $\mathbf{\Sigma}$ of the class distributions: general covariance and isotropic covariance.

General Covariance

To compute $\bar{m}(-\gamma)$ under no assumption on the structure of $\mathbf{\Sigma}$, we compute the root of the function

$$h(x) = x - \frac{1}{1 - \frac{n}{d}\frac{e(-\gamma x)}{1+e(-\gamma x)}} \tag{4.1}$$

over $x > 0$. As this function has a discontinuity, after which the root occurs, we must restrict the search to the interval beyond the discontinuity, within which it is monotonically increasing. We find the discontinuity by computing the root of the monotonically decreasing function

$$f(x) = \frac{1}{n}\mathrm{tr}\left\{\mathbf{D}_{\mathbf{\Sigma}}\left(\frac{1}{1+e(-\gamma x)} + \gamma x\mathbf{I}_p\right)^{-1}\right\} - \frac{d}{n-d} \tag{4.2}$$

over $x > 0$. This can be done numerically using, for example, the bisection method. Once the discontinuity is found, the root of $h(x)$ can be computed, also by the bisection method, which we denote as x^*. From this we can compute $\tilde{\delta}(-\gamma) = \frac{d}{p}(x^* - 1)$ which we substitute into (3.21) to obtain $\bar{m}(-\gamma)$.

Isotropic Covariance

When $\mathbf{\Sigma}$ is assumed to have the isotropic structure $\mathbf{\Sigma} = s\mathbf{I}$ for some $s \in (0, \infty)$, the computations involved can be expressed in closed form. Note that the general procedure works for this case as well.

To compute $\bar{m}(-\gamma)$, we first compute the root, denoted x^*, over $x > 0$ of the cubic equation

$$g(x) = -\gamma s^2 x^3 + \left(-2\gamma s + s^2 + \frac{p^2}{nd}s^2 - \frac{p}{d}s^2 - \frac{p}{n}s^2\right)x^2 + \left(-\gamma + 2s - \frac{p}{n}s - \frac{p}{d}s\right)x + 1 \tag{4.3}$$

satisfying the properties of a Stieltjes transform (this happens to be the third root in MATLAB). We can then compute $\tilde{\delta}(-\gamma)$ as

$$\tilde{\delta}(-\gamma) = \frac{d}{p}\left(\frac{1+sx^*}{1+\left(1-\frac{p}{d}\right)sx^*} - 1\right) \tag{4.4}$$

which we substitute into (3.21) to obtain $\bar{m}(-\gamma)$.

4.1.2 The RP-LDA Ensemble

To compute the DEs for the RP-RLDA ensemble, we substitute $\lim_{\gamma\to 0}\bar{m}(-\gamma)$ for $\bar{m}(-\gamma)$ in all expressions. To compute $\lim_{\gamma\to 0}\bar{m}(-\gamma)$, first solve for the root of the polynomial

$$q(x) = 1 - \frac{p}{d} + \frac{1}{d}\text{tr}\left\{(\mathbf{I}_p + x\mathbf{D}_{\boldsymbol{\Sigma}})^{-1}\right\} \tag{4.5}$$

As this is a decreasing function of x, its root can be found using the bisection method over $x > 0$. Denote the root by x^* and solve for $\lim_{\gamma\to 0}\bar{m}(-\gamma)$ using the relation

$$\lim_{\gamma\to 0}\bar{m}(-\gamma) = \frac{x^*}{1 - x^*\frac{1}{n}\text{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\left(\mathbf{I}_p + x^*\mathbf{D}_{\boldsymbol{\Sigma}}\right)^{-1}\right\}} \tag{4.6}$$

4.2 Using the DE as an Approximation

In this section, we provide a simulation that motivates using the derived deterministic equivalent for the generalization error of the RP-RLDA ensemble classifier as an approximation of its true generalization error. We approximate the true generalization error by the empirical error of the RP-RLDA ensemble classifier with $M = 100$ and $\gamma = 1$ on a separate testing set consisting of 800 data points. This is averaged over 1800 training sets. All data is generated synthetically and under the assumption of stratified sampling, meaning that the two classes are sampled independently of each other and so n_0 and n_1 cannot be used to estimate the prior probabilities [4]. We therefore assume the prior probabilities are known and use them directly. The data is then generated such that $\frac{n_0}{n} \approx \pi_0$ and $\frac{n_1}{n} \approx \pi_1$. Here we consider a balanced set of sample points from each class, i.e. $\pi_0 = \pi_1 = 0.5$. We generate the samples according to the statistics $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}_{ij} = 0.6^{|i-j|}$, $\forall i, j$, $\mu_0 = \frac{1}{p^{1/4}} \left[\mathbf{1}^T_{\lceil \sqrt{p} \rceil} \ \mathbf{0}^T_{p - \lceil \sqrt{p} \rceil - 2} \ 2 \ 2 \right]^T$, and $\mu_1 = \mathbf{0}_p$. The DE is plotted against the empirical error for d, p, and n increasing at a constant rate in Figure 4.1. More specifically, $p = \{80, 120, 240, 360, 480, 600\}$, $n = \lceil 0.75p \rceil$, and $d = \lceil 0.5p \rceil$, or in the notation of the assumptions in Chapter 3, $c = 0.75$ and $c' = 0.5$. It can be observed from this plot that the DE approximates the empirical error well, as the difference between the two is on the order of 10^{-2}. As the DE is relatively quick to compute, it is a more efficient alternative to cross-

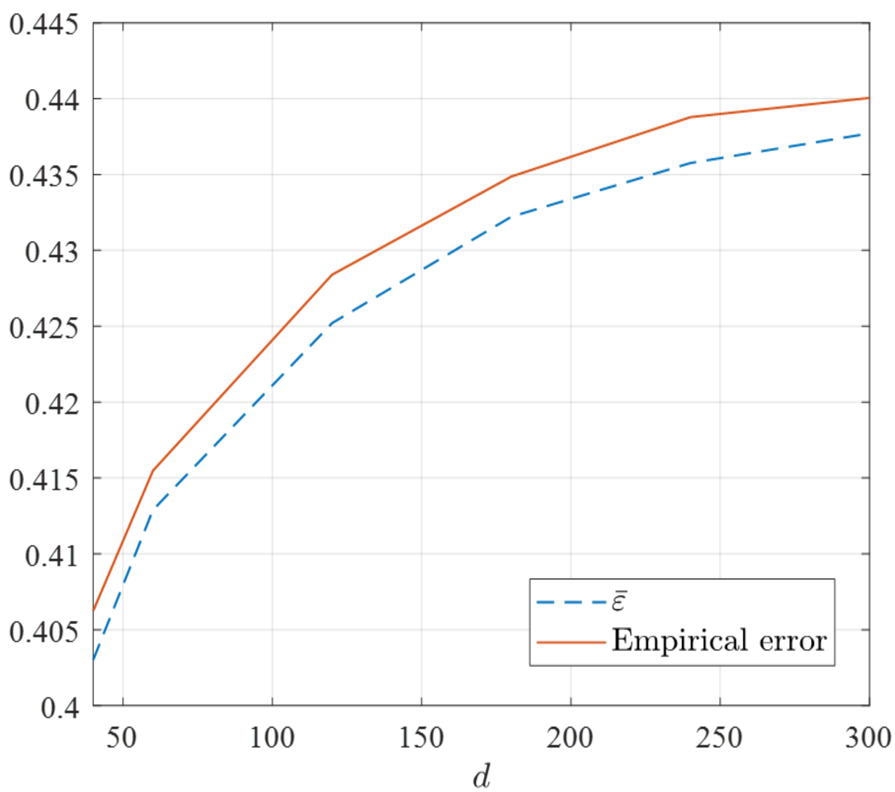

Figure 4.1: A plot of the generalization error DE $\bar{\varepsilon}$ against the empirical error of the RP-RLDA ensemble classifier with $M = 100$ and $\gamma = 1$ for n, p, and d increasing at constant rates $c = 0.75$ and $c' = 0.5$.

validation, which is the traditional method of error estimation in the machine learning community. The only drawback here is that the DE depends on knowledge of the true statistics of the classes from which the data is drawn. Regardless, the DE can be still be used as the basis for constructing a consistent estimator of the generalization error in terms of the sample statistics, but this is left for future work.

4.3 Parameter Optimization of the RP-LDA Ensemble

In this section, we provide a simulation to motivate using the derived deterministic equivalent of the generalization error of the RP-LDA ensemble classifier for tuning the projected dimension parameter d. Just as in Section 4.2, all data is generated synthetically under the assumption of stratified sampling and the true generalization error is approximated by the empirical error of the RP-LDA classifier with $M = 100$ projections on a separate test set consisting of 500 test points. This is averaged over 400 training sets. The statistics of the classes are: $\mu_0 = \frac{1}{p^{1/4}} \left[\mathbf{1}^T_{\lceil \sqrt{p} \rceil} \ \mathbf{0}^T_{p - \lceil \sqrt{p} \rceil - 2} \ 2 \ 2 \right]^T$, $\mu_1 = \mathbf{0}_p$, and $\mathbf{\Sigma}$ having 0.2 along the diagonal and 0.1 elsewhere. Figure 4.2 plots the error DE $\bar{\varepsilon}$ against the empirical error for $n = 100$, $p = 200$, and varying d. For the particular choice of statistics, there is an optimal d. The simulation shows that $\bar{\varepsilon}$ approximates the empirical error well in the region where it matters the most, as the difference between the two curves between $d = 15$ and $d = 70$ is mostly on the order of 10^{-2}. This simulation also shows, by the steep difference in error for different choices of d, that tuning d properly can be critical to the classifier's performance and therefore reliable estimation of the error is very important to facilitate optimal parameter selection. The derived DE $\bar{\varepsilon}$ is accurate and has the additional advantage of computational efficiency compared to the traditional method of cross-validation. Although in practice it cannot be used directly to tune the parameters, as it is a function of the true statistics which are usually unknown, as mentioned previously, the DE can be used as the basis of a consistent estimator with which the projection dimension d may be tuned in a similar manner.

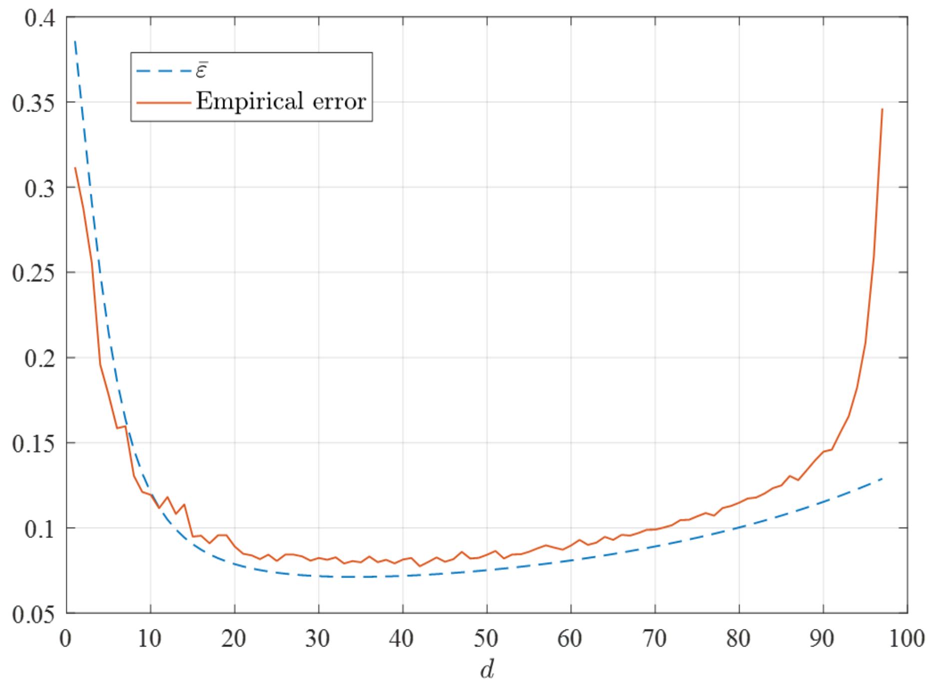

Figure 4.2: A plot of the generalization error DE $\bar{\varepsilon}$ against the empirical error of the RP-LDA ensemble classifier with $M = 100$ for $n = 100$, $p = 200$, and varying d.

Chapter 5

Conclusion and Future Work

In conclusion, we have conducted an asymptotic study of the generalization error of the RP-RLDA ensemble classifier under the small-sample regime. This regime necessitates the use of tools from random matrix theory. The main result of the study is a deterministic function of the true statistics of the data and the problem dimension that approximates the generalization error well for d, n, p, and M large enough. We show this by simulation. We also derive an analogous quantity for the RP-LDA ensemble classifier as a limiting case of the former. We motivate the use of this limit for tuning the projection dimension d of the RP-LDA ensemble classifier by simulation.

For future work, we plan to investigate the optimization of the parameters d and γ of RP-RLDA ensemble. We also wish to obtain a consistent estimator of the generalization error of both the RP-LDA ensemble and the RP-RLDA ensemble classifiers in terms of the sample statistics, based on the deterministic equivalents developed in this thesis. From there, we can apply the estimator to real data and not just synthetic data for which we know the true statistics.

REFERENCES

[1] E. R. Dougherty, "Small sample issues for microarray-based classification," *Comparative and Functional Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[2] P. M. Domingos, "A few useful things to know about machine learning." *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[3] E. R. Dougherty, C. Sima, B. Hanczar, U. M. Braga-Neto *et al.*, "Performance of error estimators for classification," *Current Bioinformatics*, vol. 5, no. 1, pp. 53–67, 2010.

[4] A. Zollanvari and E. R. Dougherty, "Generalized consistent error estimator of linear discriminant analysis," *IEEE transactions on signal processing*, vol. 63, no. 11, pp. 2804–2814, 2015.

[5] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis classifiers," *arXiv preprint arXiv:1711.00382*, 2017.

[6] X. Yang, K. Elkhalil, A. Kammoun, T. Y. Al-Naffouri, and M.-S. Alouini, "Regularized discriminant analysis: A large dimensional study," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 536–540.

[7] R. J. Durrant, "Learning in high dimensions with projected linear discriminants," Ph.D. dissertation, University of Birmingham, 2013.

[8] R. J. Durrant and A. Kabán, "Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions," *Machine Learning*, vol. 99, no. 2, pp. 257–286, 2015.

[9] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining*. Springer, 2017.

[10] T. I. Cannings and R. J. Samworth, "Random-projection ensemble classification," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 959–1035, 2017.

[11] A. Muller and M. Debbah, “Random matrix theory tutorial-Introduction to deterministic equivalents,” *Traitement du signal*, vol. 33, no. 2-3, pp. 223–248, 2016.

[12] W. Hachem, P. Loubaton, J. Najim, P. Vallet *et al.*, “On bilinear forms based on the resolvent of large random matrices,” in *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, vol. 49, no. 1. Institut Henri Poincaré, 2013, pp. 36–63.

[13] F. Benaych-Georges and R. Couillet, “Spectral analysis of the gram matrix of mixture models,” *ESAIM: Probability and Statistics*, vol. 20, pp. 217–237, 2016.

APPENDICES

Appendix A

Derivation of DEs

As was explained in Chapter 3, in order to derive the deterministic equivalent $\bar{\varepsilon}$ of the generalization error of the RP-RLDA ensemble classifier, we must first derive deterministic equivalents for the statistics of the discriminant, $\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q))$ (which, conditioned on the training data and the class of the query point $\mathbf{x}_q$, is Gaussian). Formally, what is required are deterministic sequences of d, p, and n, denoted $\bar{m}_0$, $\bar{m}_1$, and $\bar{\sigma}^2$, such that

$$\begin{aligned} m_0 - \bar{m}_0 &\xrightarrow{\text{a.s.}} 0 \\ m_1 - \bar{m}_1 &\xrightarrow{\text{a.s.}} 0 \\ \sigma^2 - \bar{\sigma}^2 &\xrightarrow{\text{a.s.}} 0 \end{aligned} \tag{A.1}$$

as d, p, and n grow to infinity according to the growth assumptions stated in Chapter 3. In the following sections, each of the deterministic equivalents $\bar{m}_0$, $\bar{m}_1$, and $\bar{\sigma}^2$ is derived in turn. Throughout, the random projection matrix $\mathbf{R} \in \mathbb{R}^{d\times p}$ is normalized by $\sqrt{d}$ so that $R_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$.

A.1 Derivation of the DE of m_0

The derivations here begin with an initial development of terms. Once the expressions are in a certain form, we can apply existing results from random matrix theory to obtain deterministic equivalents. The challenge in this particular problem, is that there are two sources of randomness: the projection $\mathbf{R}$ and the training data $\mathcal{T}$. To deal with this, we first condition on $\mathcal{T}$, develop the terms and obtain a deterministic equivalent with respect to $\mathbf{R}$ that is still random from $\mathcal{T}$. We then develop the resulting terms and derive the final deterministic equivalent with respect to both $\mathbf{R}$ and $\mathcal{T}$.

Firstly, rewrite (3.9) as

$$\begin{aligned} m_0 &= (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}} \left[\mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \right] \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \\ &= \mathbb{E}_{\mathbf{R}} \left[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbf{R}^T (\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \right] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \end{aligned}$$

where the expectation is with respect to $\mathbf{R}$, given the training data $\mathcal{T}$. Note that the asymptotic limit of the bias term $\ln\frac{\hat{\pi}_1}{\hat{\pi}_0}$ requires no special treatment as it involves scalar quantities. It converges to $\ln\frac{\pi_1}{\pi_0}$. Now, recall that we denote the eigendecomposition of $\hat{\boldsymbol{\Sigma}}$ by $\hat{\boldsymbol{\Sigma}} = \mathbf{UDU}$. For now, let us develop the inner part of the expectation $(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbf{R}^T (\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$. We will reintroduce the $\mathbb{E}_{\mathbf{R}}[\cdot]$ as well as the bias term later. Now, Let $\mathbf{a} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ and $\mathbf{b} = \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}$, then

$$\begin{aligned} &(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R} \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \\ &\quad = \mathbf{a}^T \mathbf{R}^T (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma \mathbf{I})^{-1} \mathbf{R}\mathbf{b} \\ &\quad = \mathbf{a}^T \mathbf{U}\mathbf{U}^T \mathbf{R}^T (\mathbf{R}\mathbf{U}\mathbf{D}\mathbf{U}^T \mathbf{R}^T + \gamma \mathbf{I}) \mathbf{R}\mathbf{U}\mathbf{U}^T \mathbf{b} \end{aligned} \tag{A.2}$$

Let $\tilde{\mathbf{a}}^T = \mathbf{a}^T\mathbf{U}$ and $\tilde{\mathbf{b}} = \mathbf{U}^T\mathbf{b}$. Also let $\tilde{\mathbf{R}} = \mathbf{RU}$. Each column $\{\mathbf{r}_i\}_{i=1}^p$ of $\mathbf{R}$ is independently distributed as $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$. Upon the transformation $\tilde{\mathbf{R}} = \mathbf{R}\mathbf{U}$, the columns are independently distributed as $\mathbf{r}_i\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{U}^T\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$, i.e. this transformation preserves the distribution of $\mathbf{R}$. Continuing from (A.2), we have

$$\begin{aligned}
&= \tilde{\mathbf{a}}^T\tilde{\mathbf{R}}^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{b}} \\
&= \sum_{i,j} \tilde{a}_i \left[\tilde{\mathbf{R}}^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{R}}\right]_{ij} \tilde{b}_j \\
&= \sum_{i,j} \tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j
\end{aligned} \tag{A.3}$$

where the last line follows from expressing $\tilde{\mathbf{R}}$ as $\tilde{\mathbf{R}} = [\tilde{\mathbf{r}}_1 \cdots \tilde{\mathbf{r}}_p]$. We can split this summation into a summation over indices i and j such that $i = j$ and a summation over i and j such that $i \neq j$, and consider each summation separately for further analysis. Doing this and reintroducing the expectation and bias term yields the exact expression

$$m_0 = \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_i \tilde{a}_i\tilde{b}_i\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_i\right] + \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j} \tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j\right] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \tag{A.4}$$

The next two subsections derive deterministic equivalents for each of the first two terms in (A.4). The second term is shown to converge almost surely to zero and so what remains of m_0 asymptotically is the deterministic equivalent of the first term.

A.1.1 DE of $\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_i \tilde{a}_i\tilde{b}_i\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_i\right]$

First consider the inner part of the expectation. By expressing $\tilde{\mathbf{R}}$ in terms of its columns we can write

$$\sum_i \tilde{a}_i\tilde{b}_i\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_i = \sum_i \tilde{a}_i\tilde{b}_i\tilde{\mathbf{r}}_i^T\left(\sum_{k=1}^{p} d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T + \gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_i \tag{A.5}$$

where d_k denotes the kth entry of $\mathbf{D}$. If the dependence between the middle term and the side vectors in this quadratic form is removed, i.e. $\tilde{\mathbf{r}}_i$ does not appear in the middle term, we will be able to apply the trace lemma (see page 10 in [11]). To remove the dependence, we can apply one of the matrix inversion lemmas (see page 11 in [11]) as follows:

$$\begin{aligned}\tilde{\mathbf{r}}_i^T \left(\sum_{k=1}^{p} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i &= \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} + d_i \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i{}^T\right)^{-1} \tilde{\mathbf{r}}_i \\ &= \frac{\tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i}{1 + d_i \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i}\end{aligned}$$

so we have

$$\sum_i \tilde{a}_i \tilde{b}_i \tilde{\mathbf{r}}_i^T (\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma \mathbf{I})^{-1} \tilde{\mathbf{r}}_i = \sum_i \tilde{a}_i \tilde{b}_i \frac{\tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i}{1 + d_i \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i} \tag{A.6}$$

We now reintroduce the $\mathbb{E}_{\mathbf{R}}[\cdot]$ into the expression. Using the result of (A.6) and letting $\alpha_i = \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i$, then

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_i \tilde{a}_i \tilde{b}_i \tilde{\mathbf{r}}_i^T (\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma \mathbf{I})^{-1} \tilde{\mathbf{r}}_i\right] &= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_i \tilde{a}_i \tilde{b}_i \frac{\tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i}{1 + d_i \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i}\right] \\ &= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_i \frac{\tilde{a}_i \tilde{b}_i \alpha_i}{1 + d_i \alpha_i}\right] \\ &= \sum_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\tilde{a}_i \tilde{b}_i \alpha_i}{1 + d_i \alpha_i}\right] \\ &= \sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1 + d_i \alpha_i}\right] \end{aligned} \tag{A.7}$$

Up to this point, no asymptotics are involved. Equation (A.7) is an exact expression. It contains the quadratic form α_i in both its numerator and denominator. In the following, we express it as the sum of a term with α_i in the denominator replaced by its expectation with respect to $\tilde{\mathbf{r}}_i$ and a term ϵ denoting the error in doing so. We will show that the error $\epsilon \xrightarrow{\text{a.s.}} 0$.

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_i \tilde{a}_i\tilde{b}_i\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_i\right] &= \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\alpha_i}\right] \\
&= \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\alpha_i}\right] - \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] \\
&\quad + \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] \\
&= \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] + \epsilon \qquad \text{(A.8)}
\end{aligned}$$

where

$$\begin{aligned}
\epsilon &= \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\alpha_i}\right] - \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] \\
&= \sum_i \tilde{a}_i\tilde{b}_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{d_i\alpha_i(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]-\alpha_i)}{(1+d_i\alpha_i)(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right]
\end{aligned}$$

We can show that ϵ converges almost surely to zero by bounding it by a decaying function of d. This would mean that every realization of the sequence $\epsilon_1, \epsilon_2, \ldots$ converges to zero so that $P\left[\lim_{n\to\infty}\epsilon_n = 0\right] = 1$, which is almost-sure convergence by definition. First bound ϵ as follows

$$|\epsilon| \le \sum_i |\tilde{a}_i||\tilde{b}_i|\left|\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{d_i\alpha_i(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]-\alpha_i)}{(1+d_i\alpha_i)(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right]\right|$$

Notice that the quadratic form $\alpha_i = \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i$ is positive. This follows from the fact that $\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}$ is positive definite (since $\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T \succeq 0$ and γ is chosen to make the matrix invertible, i.e. positive definite). In conjunction with the fact that $d_i \geq 0$ (since they are the eigenvalues of the positive semi-definite covariance matrix $\hat{\mathbf{\Sigma}}$), this leads to

$$\begin{aligned}
\alpha_i > 0 &\implies (1 + d_i\alpha_i)(1 + d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]) > 1 \\
&\implies |\epsilon| \leq \sum_i |\tilde{a}_i||\tilde{b}_i| \left|\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{d_i\alpha_i(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)}{(1 + d_i\alpha_i)(1 + d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right]\right| < \sum_i |\tilde{a}_i||\tilde{b}_i| \left|\mathbb{E}_{\tilde{\mathbf{R}}}\left[d_i\alpha_i(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)\right]\right|
\end{aligned}$$

Furthermore, by the triangle inequality

$$\begin{aligned}
|\epsilon| &< \sum_i |\tilde{a}_i||\tilde{b}_i| \left|\mathbb{E}_{\tilde{\mathbf{R}}}\left[d_i\alpha_i(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)\right]\right| \\
&\leq \sum_i |\tilde{a}_i||\tilde{b}_i| \mathbb{E}_{\tilde{\mathbf{R}}}\left[|d_i\alpha_i||\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i|\right]
\end{aligned}$$

We can expand $\mathbb{E}_{\tilde{\mathbf{R}}}\left[|d_i\alpha_i||\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i|\right]$ into $\mathbb{E}_{\tilde{\mathbf{R}}}\left[\mathbb{E}_{\tilde{\mathbf{r}}_i}[|d_i\alpha_i||\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i|]\right]$ by iterated expectations. By applying the Cauchy-Schwarz inequality, $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$, to the inner expectation we obtain

$$\mathbb{E}_{\tilde{\mathbf{R}}}\left[\mathbb{E}_{\tilde{\mathbf{r}}_i}[|d_i\alpha_i||\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i|]\right] \leq \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(d_i\alpha_i)^2]}\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)^2]}\right]$$

We now have

$$|\epsilon| < \sum_i |\tilde{a}_i||\tilde{b}_i| \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(d_i\alpha_i)^2]}\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)^2]}\right] \tag{A.9}$$

Consider $\mathbb{E}_{\tilde{\mathbf{r}}_i}[(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)^2] = \mathbb{E}_{\tilde{\mathbf{r}}_i}[(\alpha_i - \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2]$ first. Recall $\alpha_i = \tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i$, then

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] &= \mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i\right] \\
&= \mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\operatorname{tr}\left\{\tilde{\mathbf{r}}_i^T \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i\right\}\right] \\
&= \mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\operatorname{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T\right\}\right] \\
&= \operatorname{tr}\left\{\mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T\right]\right\} \\
&= \operatorname{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1} \mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T\right]\right\} \\
&= \frac{1}{d}\operatorname{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\}
\end{aligned} \tag{A.10}$$

where the last line follows from the fact that $\mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T\right] = \frac{1}{d}\mathbf{I}$. Now let $\mathbf{A} = \left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}$ and $\bar{\mathbf{r}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We are working with

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathbf{r}}_i}(\alpha_i - \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2] &= \mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\left(\tilde{\mathbf{r}}_i^T \mathbf{A} \tilde{\mathbf{r}}_i - \frac{1}{d}\operatorname{tr}\{\mathbf{A}\}\right)^2\right] \\
&= \mathbb{E}_{\bar{\mathbf{r}}_i}\left[\left(\frac{1}{d}\bar{\mathbf{r}}_i^T \mathbf{A} \bar{\mathbf{r}}_i - \frac{1}{d}\operatorname{tr}\{\mathbf{A}\}\right)^2\right] \\
&= \frac{1}{d^2}\mathbb{E}_{\bar{\mathbf{r}}_i}\left[\left(\bar{\mathbf{r}}_i^T \mathbf{A} \bar{\mathbf{r}}_i - \operatorname{tr}\{\mathbf{A}\}\right)^2\right]
\end{aligned}$$

The form $\mathbb{E}_{\bar{\mathbf{r}}_i}\left[\left(\bar{\mathbf{r}}_i^T\mathbf{A}\bar{\mathbf{r}}_i - \operatorname{tr}\{\mathbf{A}\}\right)^2\right]$ satisfies the conditions for applying the preliminary trace lemma result (see page 9 in [11]). Doing so yields

$$\begin{aligned}\frac{1}{d^2}\mathbb{E}_{\bar{\mathbf{r}}_i}\left[\left(\bar{\mathbf{r}}_i^T\mathbf{A}\bar{\mathbf{r}}_i - \operatorname{tr}\{\mathbf{A}\}\right)^2\right] &\leq \frac{C_1}{d^2}\operatorname{tr}\{\mathbf{A}\mathbf{A}^T\} \\ &\leq \frac{C_1}{\gamma^2 d} \\ &= \frac{C_1'}{d}\end{aligned}$$

where C_1 is a constant and the last lines follow from $\operatorname{tr}\{\mathbf{A}\mathbf{A}^T\} \leq d\lambda_{\max}\{\mathbf{A}\mathbf{A}^T\} = d\lambda_{\max}\{\mathbf{A}^2\} \leq \frac{d}{\gamma^2}$ $\left(\text{since } \lambda_{\max}\{\mathbf{A}\} = \frac{1}{\lambda_{\min}\{\sum_{k\neq i} d_k\tilde{\mathbf{r}_k}\tilde{\mathbf{r}_k}^T\}+\gamma} \leq \frac{1}{\gamma}\right)$ and combining C_1 and γ into the single constant C_1'. Notice this bound depends only on d. So now we have the result that

$$\mathbb{E}_{\tilde{\mathbf{r}}_i}[(\alpha_i - \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2] \leq \frac{C_1'}{d} \tag{A.11}$$

The bound on ϵ can be developed as

$$\begin{aligned}|\epsilon| &< \sum_i |\tilde{a}_i||\tilde{b}_i|\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(d_i\alpha_i)^2]}\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]-\alpha_i)^2]}\right] \\ &\leq \sqrt{\frac{C_1'}{d}}\sum_i |\tilde{a}_i||\tilde{b}_i|\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[(d_i\alpha_i)^2]}\right] \\ &\leq \sqrt{\frac{C_1'}{d}}\sum_i |\tilde{a}_i||\tilde{b}_i|d_i\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i^2]}\right] \\ &\leq \frac{K_1}{\sqrt{d}}\sum_i |\tilde{a}_i||\tilde{b}_i|\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i^2]}\right]\end{aligned}$$

where the last line follows from incorporating the largest d_i into the constant, assuming all d_i are finite, which follows from the assumed growth regime. We need to find a bound for the remaining expectation term to be able to check that the bound is overall decaying with increasing d. By Jensen's inequality,

$$\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i^2]}\right] \leq \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}\left[\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i^2]\right]}$$

Using the fact that $\sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}\left[\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i^2]\right]} = \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i^2]}$,

$$|\epsilon| < \frac{K_1}{\sqrt{d}} \sum_i |\tilde{a}_i||\tilde{b}_i| \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i^2]}$$

The expectation term can be bounded as follows

$$\begin{aligned}
\sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i^2]} &= \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[(\tilde{\mathbf{r}}_i^T \mathbf{A} \tilde{\mathbf{r}}_i)^2]} \\
&\leq \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[(||\tilde{\mathbf{r}}_i||_2 ||\mathbf{A}||_2 ||\tilde{\mathbf{r}}_i||_2)^2]} \\
&= \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[||\mathbf{A}||_2^2 ||\tilde{\mathbf{r}}_i||_2^4]} \\
&= \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[||\mathbf{A}||_2^2] \mathbb{E}_{\tilde{\mathbf{R}}}[||\tilde{\mathbf{r}}_i||_2^4]}
\end{aligned}$$

where the last line makes use of the fact that $\mathbf{A}$ and $\tilde{\mathbf{r}}_i$ are independent. Because $||\mathbf{A}||_2^2 = \lambda_{\max}\{\mathbf{A}\mathbf{A}^T\} \leq \frac{1}{\gamma^2}$ is bounded and because all of the moments of the Gaussian random vector $\tilde{\mathbf{r}}_i$ are bounded, we have $\sqrt{\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i^2]} \sim \mathcal{O}(1)$ which yields

$$|\varepsilon| < \frac{K_1'}{\sqrt{d}} \sum_i |\tilde{a}_i||\tilde{b}_i| \leq \frac{K_1'}{\sqrt{d}} ||\tilde{\mathbf{a}}||_2 ||\tilde{\mathbf{b}}||_2$$

where the second line follows by the Cauchy-Schwarz inequality. We still need to check the asymptotic behavior of $||\tilde{\mathbf{a}}||_2||\tilde{\mathbf{b}}||_2$ before we can claim that the overall behavior of the bound is decaying. Both $||\tilde{\mathbf{a}}||_2$ and $||\tilde{\mathbf{b}}||_2$ can be shown to be $\mathcal{O}(1)$, assuming that $||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2 < \infty$. We add this as an assumption of the growth regime. So overall, ϵ is bounded by a term that decays with $\sqrt{d}$, which means that $\epsilon \xrightarrow[d \longrightarrow \infty]{\text{a.s.}} 0$. Looking back at (A.8), we now need to analyze the remaining term

$$\sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right]$$

By using iterated expectations and (A.10),

$$\begin{aligned}
\sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\alpha_i}{1+d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] &= \sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\frac{\alpha_i}{1+d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right]\right] \\
&= \sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}{1+d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] \\
&= \sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\}}{1+d_i \frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\}}\right]
\end{aligned}$$

If we can find an almost-sure limit for $\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\}$, then since the function within $\mathbb{E}_{\tilde{\mathbf{R}}}[\cdot]$ is continuous over its domain, then the limit can be substituted inside by the continuous-mapping theorem to obtain the limit of the function. The expectation term is then equivalent to the expectation of this limit if this can be justified using the dominated convergence theorem or bounded convergence theorem. The limit for $\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\}$ is obtained by applying the rank-one perturbation lemma (see page 12 in [11]) yielding

$$\frac{1}{d}\left|\mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\} - \mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + d_i \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T + \gamma \mathbf{I}\right)^{-1}\right\}\right| \leq \frac{1}{\gamma d}$$

which means that

$$\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k\neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\} - \frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k=1}^{p} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I}\right)^{-1}\right\} \xrightarrow[d\longrightarrow\infty]{\text{a.s.}} 0$$

Note that $d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] \geq 0 \implies \frac{\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}{1+d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]} \leq \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] \leq \frac{1}{\gamma}$ since $\alpha_i = \tilde{\mathbf{r}}_i^T \mathbf{A} \tilde{\mathbf{r}}_i \leq \lambda_{\max}\{\mathbf{A}\} ||\tilde{\mathbf{r}}_i||_2^2 \leq \frac{1}{\gamma} \tilde{\mathbf{r}}_i^T \tilde{\mathbf{r}}_i = \frac{1}{\gamma d} \leq \frac{1}{\gamma}$. This satisfies the conditions for the bounded convergence theorem. Finally, by the continuous mapping theorem and the bounded convergence theorem we have

$$\begin{aligned} &\sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}} \left[\frac{\frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k \neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\}}{1 + d_i \frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k \neq i} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\}} \right] \\ &\quad - \sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}} \left[\frac{\frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\}}{1 + d_i \frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\}} \right] \xrightarrow[d \longrightarrow \infty]{\text{a.s.}} 0 \end{aligned} \tag{A.12}$$

Using a similar approach to that in Section A.1.2, we can show that

$$\begin{aligned} &\sum_i \tilde{a}_i \tilde{b}_i \mathbb{E}_{\tilde{\mathbf{R}}} \left[\frac{\frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\}}{1 + d_i \frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\}} \right] \\ &\quad - \sum_i \tilde{a}_i \tilde{b}_i \frac{\mathbb{E}_{\tilde{\mathbf{R}}} \left[\frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\} \right]}{1 + d_i \mathbb{E}_{\tilde{\mathbf{R}}} \left[\frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\} \right]} \xrightarrow[d \longrightarrow \infty]{\text{a.s.}} 0 \end{aligned} \tag{A.13}$$

The aim of all preceding derivations is to obtain an asymptotic expression of m_0 that involves quantities for which we have existing random matrix theory results from which we can directly obtain deterministic equivalents. Let

$$m(-\gamma) = \frac{1}{d}\mathrm{tr}\left\{ \left(\sum_{k=1}^p d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k{}^T + \gamma \mathbf{I} \right)^{-1} \right\} = \frac{1}{d}\mathrm{tr}\left\{ \left(\tilde{\mathbf{R}} \mathbf{D} \tilde{\mathbf{R}}^T + \gamma \mathbf{I} \right)^{-1} \right\}$$

which is the normalized trace of the resolvent of $\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T$. This is such a quantity. This quantity is the Stieltjes transform of the empirical spectral measure of $\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T$

evaluated at $-\gamma$ (see page 6 of [11]). We now have

$$\sum_i \tilde{a}_i \tilde{b}_i \frac{\mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k=1}^{p} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k^T + \gamma \mathbf{I}\right)^{-1}\right\}\right]}{1 + d_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k=1}^{p} d_k \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k^T + \gamma \mathbf{I}\right)^{-1}\right\}\right]} = \sum_i \tilde{a}_i \tilde{b}_i \frac{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}{1 + d_i \mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]} \tag{A.14}$$

This can be written in matrix form as

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right] \tilde{\mathbf{a}}^T \left(\mathbf{I} + \mathbf{D}\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]\right)^{-1} \tilde{\mathbf{b}} \\
&\quad = \mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right] (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbf{U} \left(\mathbf{I} + \mathbf{D}\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]\right)^{-1} \mathbf{U}^T \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) \\
&\quad = \mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right] (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\mathbf{I} + \hat{\boldsymbol{\Sigma}}\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]\right)^{-1} \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) \\
&\quad = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1} \left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right)
\end{aligned} \tag{A.15}$$

At this point we could obtain the deterministic equivalent for (A.15), but this will be random from the training data. To overcome this, we introduce the true statistics into the expressions. In the following, we first remove the dependence on the sample means.

It can be shown that the sample means, $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$, are independent of the pooled covariance matrix $\hat{\boldsymbol{\Sigma}}$ (and therefore of its components $\mathbf{U}$ and $\mathbf{D}$). We can take advantage of this by explicitly expressing the random part of $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ in (A.15) and taking the expectation over it so that any cross-terms with $\hat{\boldsymbol{\Sigma}}$, $\mathbf{U}$, or $\mathbf{D}$ simplify to zero. We expect m_0 to converge asymptotically to its mean anyway.

The randomness in $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ can be expressed through the random matrices $\mathbf{Z}_0$ and $\mathbf{Z}_1$ respectively, which each have i.i.d. Gaussian zero-mean and unit variance entries. Doing this, we obtain

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

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

Substituting these into (A.15) and taking the expectation over $\mathbf{Z}_0\mathbf{1}$ and $\mathbf{Z}_1\mathbf{1}$ yields

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}_0\mathbf{1},\mathbf{Z}_1\mathbf{1}}\left[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}\left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right)\right] \\
&= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}\mathbb{E}_{\mathbf{Z}_0\mathbf{1},\mathbf{Z}_1\mathbf{1}}\left[\left(\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{Z}_0\mathbf{1}}{n_0}\right)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}\left(\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{Z}_0\mathbf{1}}{n_0}\right)\right] \\
&\quad - \frac{1}{2}\mathbb{E}_{\mathbf{Z}_0\mathbf{1},\mathbf{Z}_1\mathbf{1}}\left[\left(\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{Z}_1\mathbf{1}}{n_1}\right)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}\left(\frac{\boldsymbol{\Sigma}^{1/2}\mathbf{Z}_1\mathbf{1}}{n_1}\right)\right] \\
&= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}\frac{1}{n_0}\mathrm{tr}\left\{\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}\right\} \\
&\quad - \frac{1}{2}\frac{1}{n_1}\mathrm{tr}\left\{\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}}\right)^{-1}\right\}
\end{aligned} \tag{A.16}$$

The first term is a quadratic form and the last two involve traces of the resolvent of $\hat{\boldsymbol{\Sigma}}$ evaluated at $-\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}[m(-\gamma)]}$. This expression can be further developed to remove the dependence on the sample covariance matrix $\hat{\Sigma}$. This is continued later. For now, we will move on to obtain the deterministic equivalent for $\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]$ which we will definitely need.

Now we will find a deterministic equivalent for $\mathbb{E}_{\tilde{\mathbf{R}}}[m(-\gamma)]$. First note that since $m(-\gamma) \leq \frac{1}{\gamma}$, then by the bounded convergence theorem, if we have

$$m(-\gamma) - \bar{m}(-\gamma) \xrightarrow{\text{a.s}} 0$$

then

$$\mathbb{E}_{\tilde{\mathbf{R}}}[m(-\gamma)] - \mathbb{E}_{\tilde{\mathbf{R}}}[\bar{m}(-\gamma)] \xrightarrow{\text{a.s}} 0$$

with $\mathbb{E}_{\tilde{\mathbf{R}}}[\bar{m}(-\gamma)] = \bar{m}(-\gamma)$. It therefore suffices to find the deterministic equivalent of $m(-\gamma)$, denoted $\bar{m}(-\gamma)$. Recall that $m(-\gamma) = \frac{1}{d}\text{tr}\left\{\left(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T + \gamma\mathbf{I}\right)^{-1}\right\}$, the normalized trace of the resolvent of the matrix $\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T$. This matrix has a separable variance profile [12] as it can be expressed as

$$\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T = \tilde{\mathbf{R}}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\tilde{\mathbf{R}}^T = (\tilde{\mathbf{R}}\mathbf{D}^{1/2})(\tilde{\mathbf{R}}\mathbf{D}^{1/2})^T \tag{A.17}$$

Reference [12] describes how to find deterministic equivalents for normalized traces and bilinear forms of resolvents of matrices with separable variance profiles. Equation (A.17) fits their model with $n = p$, $N = d$, $D_n = \frac{p}{d}\mathbf{I}_d$, $\tilde{D}_n = \mathbf{D}$, and $A_n = \mathbf{0}$. (Note that this result requires a normalization by $n = p$, whereas our matrix $\tilde{\mathbf{R}}$ is normalized by d. Rather than changing this normalization on which all the previous results are based, we introduced the factor $\frac{p}{d}$ which appears in D_n and is a result of multiplication and division by p.) In addition, we have $||\mathbf{D}||_2 < \infty$ since $||\mathbf{\Sigma}|| < \infty$ by the assumptions of the growth regime. This satisfies the conditions indicated in [12]. According to this reference and based on our model, we need to solve the system of equations given by

$$\delta(z) = -\frac{1}{z\left(1 + \frac{p}{d}\tilde{\delta}(z)\right)} \tag{A.18}$$

$$\tilde{\delta}(z) = \frac{1}{p}\text{tr}\left\{\mathbf{D}\left(-z\mathbf{I}_p - z\delta(z)\mathbf{D}\right)^{-1}\right\} \tag{A.19}$$

for $\tilde{\delta}(z)$, from which we can compute $\bar{m}(z)$ defined by (3.21). The system of equations through which we solve for $\tilde{\delta}(z)$ to obtain the deterministic equivalent $\bar{m}(z)$ of $m(z)$ depends on $\mathbf{D}$, which is derived from the sample covariance matrix. We now introduce the true statistics into the expressions to remove the dependence on the training data. To begin with, we write the system of equations (A.18) and (A.19) as a fixed point equation obtained by substituting (A.18) into (A.19)

$$\tilde{\delta}(z) = \frac{1}{p}\mathrm{tr}\left\{\mathbf{D}\left(\frac{d}{d+p\tilde{\delta}(z)}\mathbf{D} - z\mathbf{I}_p\right)^{-1}\right\} \tag{A.20}$$

We then introduce $\hat{\mathbf{\Sigma}}$, which we can later write in terms of $\mathbf{\Sigma}$. Starting from the fixed point equation (A.20), we have

$$\begin{aligned}
\tilde{\delta}(z) &= \frac{1}{p}\mathrm{tr}\left\{\mathbf{D}\left(\frac{d}{d+p\tilde{\delta}(z)}\mathbf{D} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{1}{p}\mathrm{tr}\left\{\mathbf{D}\mathbf{U}^T\mathbf{U}\left(\frac{d}{d+p\tilde{\delta}(z)}\mathbf{D} - z\mathbf{I}_p\right)^{-1}\mathbf{U}^T\mathbf{U}\right\} \\
&= \frac{1}{p}\mathrm{tr}\left\{\mathbf{U}\mathbf{D}\mathbf{U}^T\mathbf{U}\left(\frac{d}{d+p\tilde{\delta}(z)}\mathbf{D} - z\mathbf{I}_p\right)^{-1}\mathbf{U}^T\right\} \\
&= \frac{1}{p}\mathrm{tr}\left\{\hat{\mathbf{\Sigma}}\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\}
\end{aligned}$$

To manipulate this into the form of a normalized trace of a resolvent (which we can obtain an additional DE for), we do the following:

$$\begin{aligned}
\tilde{\delta}(z) &= \frac{1}{p}\mathrm{tr}\left\{\hat{\mathbf{\Sigma}}\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{1}{p}\mathrm{tr}\left\{\frac{d}{d}\frac{d+p\tilde{\delta}(z)}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}}\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{d+p\tilde{\delta}(z)}{d}\frac{1}{p}\mathrm{tr}\left\{\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}}\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{d+p\tilde{\delta}(z)}{d}\frac{1}{p}\mathrm{tr}\left\{\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p + z\mathbf{I}_p\right)\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{d+p\tilde{\delta}(z)}{d}\frac{1}{p}\mathrm{tr}\left\{\mathbf{I}_p\right\} + \frac{d+p\tilde{\delta}(z)}{d}\frac{1}{p}\mathrm{tr}\left\{z\mathbf{I}_p\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{d+p\tilde{\delta}(z)}{d}\frac{1}{p}p + \frac{z(d+p\tilde{\delta}(z))}{d}\frac{1}{p}\mathrm{tr}\left\{\left(\frac{d}{d+p\tilde{\delta}(z)}\hat{\mathbf{\Sigma}} - z\mathbf{I}_p\right)^{-1}\right\} \\
&= \frac{d+p\tilde{\delta}(z)}{d} + z\left(\frac{d+p\tilde{\delta}(z)}{d}\right)^2\frac{1}{p}\mathrm{tr}\left\{\left(\hat{\mathbf{\Sigma}} - z\frac{d+p\tilde{\delta}(z)}{d}\mathbf{I}_p\right)^{-1}\right\} \qquad \text{(A.21)}
\end{aligned}$$

We now express the pooled sample covariance in terms of the true covariance. Recall that the pooled sample covariance matrix $\hat{\mathbf{\Sigma}}$ is defined as

$$\hat{\mathbf{\Sigma}} = \frac{(n_0-1)\hat{\mathbf{\Sigma}}_0 + (n_1-1)\hat{\mathbf{\Sigma}}_1}{n-2}$$

where $\hat{\mathbf{\Sigma}}_0 = \frac{1}{n_0-1}\sum_{i\in\mathcal{C}_0}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0)^T$ and similarly $\hat{\mathbf{\Sigma}}_1 = \frac{1}{n_1-1}\sum_{i\in\mathcal{C}_1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^T$. We can develop $\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0$ as follows:

$$\begin{aligned}
\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0 &= \boldsymbol{\mu}_\mathbf{0} + \mathbf{\Sigma}^{1/2}\mathbf{z}_i - \boldsymbol{\mu}_\mathbf{0} - \frac{1}{n_0}\sum_{j\in\mathcal{C}_0}\mathbf{\Sigma}^{1/2}\mathbf{z}_j \\
&= \left(1 - \frac{1}{n_0}\right)\mathbf{\Sigma}^{1/2}\mathbf{z}_i - \sum_{j\neq i}\frac{1}{n_0}\mathbf{\Sigma}^{1/2}\mathbf{z}_j
\end{aligned}$$

We have $\left(1-\frac{1}{n_0}\right)\boldsymbol{\Sigma}^{1/2}\mathbf{z}_i \sim \mathcal{N}\left(\mathbf{0},\left(1-\frac{1}{n_0}\right)^2\boldsymbol{\Sigma}\right)$ and each of the n_0-1 terms in the summation has $\frac{1}{n_0}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_j \sim \mathcal{N}(\mathbf{0},\frac{1}{n_0^2}\boldsymbol{\Sigma})$ therefore $\mathbf{x}_i-\hat{\boldsymbol{\mu}}_0 \sim \mathcal{N}\left(\mathbf{0},\frac{n_0-1}{n_0}\boldsymbol{\Sigma}\right)$, $i \in \mathcal{C}_0$, where the covariance is the summation $\left(\left(1-\frac{1}{n_0}\right)^2+\frac{n_0-1}{n_0^2}\right)\boldsymbol{\Sigma}$ of the individual covariance matrices, which follows from the vectors being independent. Similarly, $\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1 \sim \mathcal{N}\left(\mathbf{0},\frac{n_1-1}{n_1}\boldsymbol{\Sigma}\right)$, $i \in \mathcal{C}_1$. These can be represented as

$$\mathbf{x}_i-\hat{\boldsymbol{\mu}}_0=\sqrt{\frac{n_0-1}{n_0}}\boldsymbol{\Sigma}^{1/2}\mathbf{y}_{0i},\ i \in \mathcal{C}_0$$

and

$$\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1=\sqrt{\frac{n_1-1}{n_1}}\boldsymbol{\Sigma}^{1/2}\mathbf{y}_{1i},\ i \in \mathcal{C}_1$$

respectively, where $\mathbf{y}_{0i},\mathbf{y}_{1i} \in \mathbb{R}^p \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. By letting $\mathbf{Y}_0 \in \mathbb{R}^{p\times n_0}=[\mathbf{y}_{01},\ldots,\mathbf{y}_{0n_0}]$ and $\mathbf{Y}_1 \in \mathbb{R}^{p\times n_1}=[\mathbf{y}_{11},\ldots,\mathbf{y}_{1n_1}]$, we can write

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_0 &= \frac{1}{n_0-1}\sum_{i\in\mathcal{C}_0}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_0)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_0)^T \\ &= \frac{1}{n_0-1}\frac{n_0-1}{n_0}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{Y}_0\right)\left(\boldsymbol{\Sigma}^{1/2}\mathbf{Y}_0\right)^T \\ &= \frac{1}{n_0}\boldsymbol{\Sigma}^{1/2}\mathbf{Y}_0\mathbf{Y}_0^T\boldsymbol{\Sigma}^{1/2}\end{aligned}$$

Similarly

$$\hat{\boldsymbol{\Sigma}}_1=\frac{1}{n_1}\boldsymbol{\Sigma}^{1/2}\mathbf{Y}_1\mathbf{Y}_1^T\boldsymbol{\Sigma}^{1/2} \tag{A.22}$$

Finally, the pooled covariance matrix can be expressed as

$$\begin{aligned}\hat{\mathbf{\Sigma}} &= \frac{(n_0-1)\hat{\mathbf{\Sigma}}_0 + (n_1-1)\hat{\mathbf{\Sigma}}_1}{n-2} \\ &= \frac{1}{n-2}\left(\frac{n_0-1}{n_0}\mathbf{\Sigma}^{1/2}\mathbf{Y}_0\mathbf{Y}_0^T\mathbf{\Sigma}^{1/2} + \frac{n_1-1}{n_1}\mathbf{\Sigma}^{1/2}\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{\Sigma}^{1/2}\right) \\ &= \frac{1}{n-2}\mathbf{\Sigma}^{1/2}\left(\frac{n_0-1}{n_0}\mathbf{Y}_0\mathbf{Y}_0^T + \frac{n_1-1}{n_1}\mathbf{Y}_1\mathbf{Y}_1^T\right)\mathbf{\Sigma}^{1/2}\end{aligned}$$

Since $\lim_{n\to\infty}\frac{n_0-1}{n_0} = \lim_{n\to\infty}\frac{n_1-1}{n_1} = 1$, we have that $\frac{n_0-1}{n_0}\mathbf{Y}_0\mathbf{Y}_0^T + \frac{n_1-1}{n_1}\mathbf{Y}_1\mathbf{Y}_1^T$ can be treated as $\mathbf{Y}_0\mathbf{Y}_0^T + \mathbf{Y}_1\mathbf{Y}_1^T = \mathbf{Y}\mathbf{Y}^T$ since they behave the same asymptotically, where $\mathbf{Y} \in \mathbb{R}^{p\times n}$ has i.i.d. columns distributed as $\mathcal{N}(\mathbf{0},\mathbf{I})$. Similarly, the normalization $\frac{1}{n-2}$ can be treated as $\frac{1}{n}$. Recall that we eigendecompose $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \mathbf{V}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{V}^T$ so that $\mathbf{\Sigma}^{1/2} = \mathbf{V}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{V}^T$. By substituting the simplified expression into (A.21) and making use of the cyclic property of the trace, we obtain

$$\begin{aligned}\tilde{\delta}(z) &= \frac{d+p\tilde{\delta}(z)}{d} \\ &\quad + z\left(\frac{d+p\tilde{\delta}(z)}{d}\right)^2 \frac{1}{p}\mathrm{tr}\left\{\left(\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{V}^T\mathbf{Y}\mathbf{Y}^T\mathbf{V}\mathbf{D}_{\mathbf{\Sigma}}^{1/2} - z\frac{d+p\tilde{\delta}(z)}{d}\mathbf{I}_p\right)^{-1}\right\} \\ &\quad + \mathcal{O}\left(\frac{1}{n}\right)\end{aligned} \tag{A.23}$$

Define $\mathbf{W} = \mathbf{V}^T\mathbf{Y}$. As multiplication by an orthogonal matrix preserves a Gaussian distribution, the columns of $\mathbf{W} \in \mathbb{R}^{p\times n}$ are also i.i.d and distributed as $\mathcal{N}(\mathbf{0},\mathbf{I})$. Finally, we have

$$\begin{aligned}\tilde{\delta}(z) &= \frac{d+p\tilde{\delta}(z)}{d} \\ &\quad + z\left(\frac{d+p\tilde{\delta}(z)}{d}\right)^2 \frac{1}{p}\mathrm{tr}\left\{\left(\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2} - z\frac{d+p\tilde{\delta}(z)}{d}\mathbf{I}_p\right)^{-1}\right\} \\ &\quad + \mathcal{O}\left(\frac{1}{n}\right)\end{aligned} \tag{A.24}$$

Now consider the normalized trace of the resolvent of $\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}$ which occurs in (A.24). This is $\frac{1}{p}\text{tr}\left\{\left(\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2} - z'\mathbf{I}_p\right)^{-1}\right\}$, where $z' = z\frac{d+p\tilde{\delta}(z)}{d}$. We will apply the result in [12] again to find the deterministic equivalent of this quantity. The matrix $\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}$ has a separable variance profile [12], as it can be expressed as

$$\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2} = \left(\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\frac{\mathbf{W}}{\sqrt{n}}\right)\left(\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\frac{\mathbf{W}}{\sqrt{n}}\right)^T \tag{A.25}$$

Equation (3.22) fits the model in [12] with $n = n$, $N = p$, $D_n = \mathbf{D}_{\mathbf{\Sigma}}$, $\tilde{D}_n = \mathbf{I}_n$, $A_n = \mathbf{0}$, and with the required normalization by n. In addition, we have $||\mathbf{D}_{\mathbf{\Sigma}}||_2 < \infty$, since $||\mathbf{\Sigma}||_2 < \infty$ as required by the growth regime. Accordingly, we solve the system of equations given by (3.23) and (3.24) for $\tilde{e}(z')$, from which we can compute the matrix $\mathbf{T}(z')$ defined by (3.22) which serves as an approximation of the resolvent $\left(\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2} - z'\mathbf{I}_p\right)^{-1}$ in the sense that

$$\frac{1}{p}\text{tr}\left\{\left(\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2} - z'\mathbf{I}_p\right)^{-1}\right\} - \frac{1}{p}\text{tr}\left\{\mathbf{T}(z')\right\} \xrightarrow[n,p\longrightarrow\infty]{\text{a.s}} 0 \tag{A.26}$$

After substituting the deterministic equivalent of (3.22) into (A.24), we have

$$\begin{aligned}\tilde{\delta}(z) = {} & \frac{d+p\tilde{\delta}(z)}{d} \\ & + z\left(\frac{d+p\tilde{\delta}(z)}{d}\right)^2\frac{1}{p}\text{tr}\left\{-\frac{1}{z'}(\mathbf{I}_p + \tilde{e}(z')\mathbf{D}_{\mathbf{\Sigma}})^{-1}\right\} \\ & + o(1)\end{aligned} \tag{A.27}$$

We now have a relationship involving $\tilde{\delta}(z)$ that depends only on the true statistics. The system of equations (3.23) and (3.24) from which we can solve for $\tilde{e}(z')$ also involves only the true statistics, so there is no need for extra work on that part. Now we need to find a way to solve for $\tilde{\delta}(z)$ because once we have $\tilde{\delta}(z)$, we have the deterministic equivalent for $\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]$ which we denoted by $\bar{m}(-\gamma)$. The solution must coincide with that of the systems (A.18), (A.19) and (3.23), (3.24) respectively. Since the quantity $\frac{d+p\tilde{\delta}(z)}{d}$ makes an appearance several times in (A.27), it is only natural to do a change of variables. Let $\nu(z) = \frac{d+p\tilde{\delta}(z)}{d}$ and notice that $z' = z\frac{d+p\tilde{\delta}(z)}{d} = z\nu(z)$, then (A.27) can be rewritten as

$$\frac{d}{p}\nu(z) - \frac{d}{p} = \nu(z) + z\nu^2(z)\frac{1}{p}\mathrm{tr}\left\{-\frac{1}{z\nu(z)}(\mathbf{I}_p + \tilde{e}(z\nu(z))\mathbf{D}_{\mathbf{\Sigma}})^{-1}\right\} + o(1) \tag{A.28}$$

Before moving on to how to find $\bar{m}(-\gamma)$, we go back to (A.16) and remove the dependence on the sample covariance matrix:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}_0\mathbf{1},\mathbf{Z}_1\mathbf{1}}\left[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\mathbf{\Sigma}}\right)^{-1}\left(\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right)\right] \\
&= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\mathbf{\Sigma}}\right)^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right)\mathrm{tr}\left\{\mathbf{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\mathbf{\Sigma}}\right)^{-1}\right\} \\
&= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\mathbf{V}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\right)^{-1}\mathbf{V}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right)\mathrm{tr}\left\{\mathbf{D}_{\mathbf{\Sigma}}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\right)^{-1}\right\}
\end{aligned} \tag{A.29}$$

where $\mathbf{W} \in \mathbb{R}^{p\times n}$ is a matrix with i.i.d. standard Gaussian entries. Each of the terms in (A.67) involves the resolvent of the matrix $\frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}$, which we have seen before, evaluated at $-\frac{1}{\bar{m}(-\gamma)}$. From this, and since the second term of (A.4) is shown to converge almost-surely to zero in Section A.1.2, we have the following result [12]:

The deterministic quantity

$$
\begin{aligned}
&\bar{m}_0 = \\
&\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{V}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\mathbf{V}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right)\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\right\} + \ln\frac{\pi_1}{\pi_0}
\end{aligned}
\tag{A.30}
$$

satisfies

$$
m_0 - \bar{m}_0 \xrightarrow{\text{a.s.}} 0
$$

This result states the deterministic equivalent $\bar{m}_0$ for m_0 as a function of the true statistics. We just need to find $\bar{m}(-\gamma)$ by which $\mathbf{T}\left(\frac{1}{\bar{m}(-\gamma)}\right)$ can be computed from (3.23) and (3.24). This is also true of the expressions for $\bar{m}_1$ and $\bar{\sigma}^2$. We derive the computations for three cases: the case of general covariance, the case of isotropic covariance, and finally, the case when $\gamma \to 0$, which recovers the RP-LDA ensemble of [7].

$\bar{m}(-\gamma)$ for General Covariance

In this case, we do not impose any structure on $\boldsymbol{\Sigma}$. We can manipulate (A.28) as follows:

$$
\begin{aligned}
\frac{d}{p}\nu(z) - \frac{d}{p} &= \nu(z) + z\nu^2(z)\frac{1}{p}\mathrm{tr}\left\{-\frac{1}{z\nu(z)}(\mathbf{I}_p + \tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}})^{-1}\right\} + o(1) \\
&= \nu(z) - \nu(z)\frac{1}{p}\mathrm{tr}\left\{(\mathbf{I}_p + \tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}})^{-1}\right\} + o(1) \\
&= \nu(z) - \nu(z)\frac{1}{p}\mathrm{tr}\left\{(\mathbf{I}_p + \tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}} - \tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}})(\mathbf{I}_p + \tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}})^{-1}\right\} + o(1) \\
&= \nu(z) - \nu(z)\frac{1}{p}\mathrm{tr}\{\mathbf{I}_p\} + \nu(z)\frac{1}{p}\mathrm{tr}\left\{\tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}}(\mathbf{I}_p + \tilde{e}(z\nu(z))\mathbf{D}_{\boldsymbol{\Sigma}})^{-1}\right\} + o(1) \\
&= -\frac{n}{p}z\nu^2(z)\tilde{e}((z\nu(z))e(z\nu(z)) + o(1) \\
&= \frac{n}{p}\nu(z)\frac{e(z\nu(z))}{1 + e(z\nu(z))} + o(1)
\end{aligned}
$$

Putting $\nu(z)$ to one side, we obtain

$$\left(1 - \frac{n}{d}\frac{e(z\nu(z))}{1+e(z\nu(z))}\right)\nu(z) = 1 + o(1) \tag{A.31}$$

We then have

$$\nu(z) = \frac{1}{1 - \frac{n}{d}\frac{e(z\nu(z))}{1+e(z\nu(z))}} + o(1) \tag{A.32}$$

provided that $1 - \frac{n}{d}\frac{e(z\nu(z))}{1+e(z\nu(z))}$ is bounded away from zero. This is not always true, and the function is actually discontinuous. The discontinuity of (A.33) occurs at x such that $\frac{n}{d}\frac{e(zx)}{1+e(zx)} = 0 \Leftrightarrow e(zx) = \frac{d}{n-d}$. Inspired by (A.32), we define the function

$$h(x) = x - \frac{1}{1 - \frac{n}{d}\frac{e(zx)}{1+e(zx)}} \tag{A.33}$$

which we expect $\nu(z)$ to satisfy asymptotically. One property of Stieltjes transforms is that they are increasing functions. It can thus be deduced that $e(zx) = e(-\gamma x)$ is a decreasing function of x. As a result, $h(x)$ is increasing before and after the discontinuity. At some point, $h(x)$ should be negative for $x > 0$ so that it has a root. Its root will coincide with $\nu(z)$ obtained from the system of equations (A.18) and (A.19). The root must be positive as $\nu(-\gamma) = 1 + \frac{p}{d}\tilde{\delta}(-\gamma)$ must be positive since $\tilde{\delta}(-\gamma)$ must be positive by definition. The quantity $e(zx)$ is also positive. The root seems to occur after the discontinuity (but this has not been analytically verified). The strategy to find the root, is to first solve for the discontinuity, then locate the root in the interval beyond the discontinuity.

By substituting (3.24) into (3.23), we obtain

$$e(zx) = \frac{1}{n}\mathrm{tr}\left\{\mathbf{D}_{\mathbf{\Sigma}}\left(\frac{1}{1+e(zx)} - zx\mathbf{I}_p\right)^{-1}\right\} \tag{A.34}$$

so that to solve for the discontinuity we must find the root of the function $f(x)$ defined as follows

$$f(x) = \frac{1}{n}\mathrm{tr}\left\{\mathbf{D}_{\mathbf{\Sigma}}\left(\frac{1}{1+e(zx)} - zx\mathbf{I}_p\right)^{-1}\right\} - \frac{d}{n-d} \tag{A.35}$$

$f(x)$ is a decreasing function of x. The bisection method can be used to numerically solve for x at which the discontinuity of (A.33) occurs. This can be used to determine the search interval for the root of (A.33) which occurs after the discontinuity. This root can be similarly solved for using the bisection method. Note that x should be positive as it represents $\nu(z)$ which is a Stieltjes transform. Once $\nu(-\gamma)$ is computed, it can be used to compute $\tilde{\delta}(z) = \frac{d}{p}(\nu(z) - 1)$. This can then be used to compute the desired quantity $\bar{m}(-\gamma)$.

$\bar{m}(-\gamma)$ for Isotropic Covariance

Now consider the case of isotropic population covariance matrix of the form $\mathbf{\Sigma} = s\mathbf{I}_p$. The resulting system of equations reduces to a cubic equation as is shown in the following.

When $\mathbf{\Sigma} = s\mathbf{I}_p$, we have $\mathbf{D}_{\mathbf{\Sigma}} = s\mathbf{I}_p$ and (A.28) simplifies to

$$\begin{aligned}\frac{d}{p}\nu(z) - \frac{d}{p} &= \nu(z) + z\nu^2(z)\frac{1}{p}\mathrm{tr}\left\{-\frac{1}{z\nu(z)}(\mathbf{I}_p + s\tilde{e}(z\nu(z))\mathbf{I}_p)^{-1}\right\} + o(1) \\ &= \nu(z) - \frac{\nu(z)}{1+s\tilde{e}(z\nu(z))} + o(1) \end{aligned} \tag{A.36}$$

Solving this equation for $\nu(z)$, we obtain

$$\nu(z) = \frac{d(1+s\tilde{e}(z\nu(z)))}{d(1+s\tilde{e}(z\nu(z))) - ps\tilde{e}(z\nu(z))} + o(1) \tag{A.37}$$

which is valid for $d(1+s\tilde{e}(z\nu(z))) - p\sigma^2\tilde{e}(z\nu(z))$ bounded away from 0. Based on this, we have that the following relation holds **asymptotically**

$$\nu(z) \approx \frac{d(1+s\tilde{e}(z\nu(z)))}{d(1+s\tilde{e}(z\nu(z))) - ps\tilde{e}(z\nu(z))} = \frac{1+s\tilde{e}(z\nu(z))}{1+\left(1-\frac{p}{d}\right)s\tilde{e}(z\nu(z))} \tag{A.38}$$

We can obtain a second equation by substituting (3.23) into (3.24):

$$zns\nu(z)\tilde{e}^2(z\nu(z)) + (zn\nu(z) + ns - ps)\tilde{e}(z\nu(z)) + n = 0 \tag{A.39}$$

This is quadratic in $\tilde{e}(z\nu(z))$. It is therefore easier to make the substitution of $\nu(z)$ into this expression and solve for $\tilde{e}(z\nu(z))$ and then from that for $\nu(z)$ rather than the other way around. Hence by substituting (A.37) into (A.39) and dividing both sides by nd we obtain the following equation in $\tilde{e}(z\nu(z))$:

$$\begin{aligned} zs^2\tilde{e}^3(z\nu(z)) &+ \left(2zs + s^2 + \frac{p^2}{nd}s^2 - \frac{p}{d}s^2 - \frac{p}{n}s^2\right)\tilde{e}^2(z\nu(z)) \\ &+ \left(z + 2s - \frac{p}{n}s - \frac{p}{d}s\right)\tilde{e}(z\nu(z)) + 1 + o(1) = 0 \end{aligned} \tag{A.40}$$

Inspired by (A.40), we define a polynomial $g(x)$ as

$$g(x) = zs^2x^3 + \left(2zs + s^2 + \frac{p^2}{nd}s^2 - \frac{p}{d}s^2 - \frac{p}{n}s^2\right)x^2 + \left(z + 2s - \frac{p}{n}s - \frac{p}{d}s\right)x + 1 \tag{A.41}$$

Since from (A.40), we know that $\tilde{e}(z\nu(z))$ satisfies (A.41) asymptotically, it must be that one of the roots of this polynomial coincides with that obtained from solving the system of equations (3.23) and (3.24) numerically. (Note that although it is fine to obtain $\tilde{e}(z\nu(z))$ directly from this system of equations as it depends on the true statistics, it is not possible to do so as we do not know $\nu(z)$.) Without referring to the solution of the system, this root can be identified by the fact that it satisfies the properties of a Stieltjes transform. After solving for this root, we can then obtain $\nu(z)$ from the asymptotic relation (A.38) and then use it to compute $\tilde{\delta}(z) = \frac{d}{p}(\nu(z) - 1)$. This can then be used to compute the desired quantity $\bar{m}(-\gamma)$.

$\bar{m}(-\gamma)$ for the RP-LDA Ensemble

To obtain the DE for the generalization error of the RP-LDA ensemble, we compute the DE of the RP-RLDA ensemble in the limit $\gamma \to 0$. Since γ is contained in $\bar{m}(-\gamma)$, the key is to obtain the limit for this quantity, which depends on $\delta(-\gamma)$ and $\tilde{\delta}(-\gamma)$. Start by substituting (A.19) into (A.18), with $z = -\gamma$, to obtain

$$\delta(-\gamma) = \frac{1}{\gamma + \frac{1}{d}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{I}_p + \delta(-\gamma)\mathbf{D}\right)^{-1}\right\}} \tag{A.42}$$

Taking $\lim\limits_{\gamma \to 0}$ on each side of (A.42) results in

$$\lim_{\gamma \to 0} \delta(-\gamma) = \frac{1}{\frac{1}{d}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{I}_p + \lim\limits_{\gamma \to 0} \delta(-\gamma)\mathbf{D}\right)^{-1}\right\}} \tag{A.43}$$

Denote the limit by the symbol $\zeta = \lim\limits_{\gamma \to 0} \delta(-\gamma)$ so that we have

$$\zeta = \frac{1}{\frac{1}{d}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{I}_p + \zeta\mathbf{D}\right)^{-1}\right\}} \tag{A.44}$$

Similar to the preceding derivations, we can express

$$\begin{aligned}
\frac{1}{d}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{I}_p + \zeta\mathbf{D}\right)^{-1}\right\} &= \frac{1}{d}\mathrm{tr}\left\{\hat{\mathbf{\Sigma}}\left(\mathbf{I}_p + \zeta\hat{\mathbf{\Sigma}}\right)^{-1}\right\} \\
&= \frac{p}{d\zeta} - \frac{1}{d\zeta^2}\mathrm{tr}\left\{\left(\frac{1}{\zeta}\mathbf{I}_p + \hat{\mathbf{\Sigma}}\right)^{-1}\right\} \\
&= \frac{p}{d\zeta} - \frac{1}{d\zeta^2}\mathrm{tr}\left\{\left(\frac{1}{\zeta}\mathbf{I}_p + \frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\right)^{-1}\right\} + \mathcal{O}\left(\frac{1}{n}\right)
\end{aligned} \tag{A.45}$$

We now have

$$\frac{1}{\zeta} = \frac{p}{d\zeta} - \frac{1}{d\zeta^2}\mathrm{tr}\left\{\left(\frac{1}{\zeta}\mathbf{I}_p + \frac{1}{n}\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\mathbf{\Sigma}}^{1/2}\right)^{-1}\right\} + \mathcal{O}\left(\frac{1}{n}\right) \tag{A.46}$$

We solve the system of equations (3.23) and (3.24) with $z' = -\frac{1}{\zeta}$ to obtain the DE $\mathbf{T}\left(-\frac{1}{\zeta}\right) = \zeta\left(\mathbf{I}_p + \tilde{e}\left(-\frac{1}{\zeta}\right)\right)$, which we substitute in (A.45) resulting in the equation

$$\frac{1}{\zeta} = \frac{p}{d\zeta} - \frac{1}{d\zeta}\mathrm{tr}\left\{\left(\mathbf{I}_p + \tilde{e}\left(-\frac{1}{\zeta}\right)\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\} + o\left(1\right) \tag{A.47}$$

This suggests that asymptotically, the solution $\tilde{e}\left(-\frac{1}{\zeta}\right)$ should be the root of the polynomial

$$q(x) = 1 - \frac{p}{d} + \frac{1}{d}\mathrm{tr}\left\{\left(\mathbf{I}_p + x\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\} \tag{A.48}$$

This is a decreasing function of x and can be solved using the bisection method over $x > 0$. The root is the desired $\tilde{e}\left(-\frac{1}{\zeta}\right)$. We can then use this to solve for the desired limit ζ. Substitute (3.23) into (3.24) with $z' = -\frac{1}{\zeta}$ and solve for ζ to obtain

$$\zeta = \frac{\tilde{e}\left(-\frac{1}{\zeta}\right)}{1 - \tilde{e}\left(-\frac{1}{\zeta}\right)\frac{1}{n}\mathrm{tr}\left\{\mathbf{D}_{\mathbf{\Sigma}}\left(\mathbf{I}_p + \tilde{e}\left(-\frac{1}{\zeta}\right)\mathbf{D}_{\mathbf{\Sigma}}\right)^{-1}\right\}} \tag{A.49}$$

Now substitute (A.19) into (3.21) with $z = -\gamma$ and take $\lim_{\gamma\to 0}$ on both sides to obtain the relationship

$$\lim_{\gamma\to 0}\bar{m}(-\gamma) = \frac{1}{d}\mathrm{tr}\left\{\frac{1}{\frac{1}{d}\mathrm{tr}\left\{\mathbf{D}\left(\mathbf{I}_p + \zeta\mathbf{D}\right)^{-1}\right\}}\mathbf{I}_d\right\} \tag{A.50}$$

Substituting (A.44) into (A.50), we finally obtain

$$\begin{aligned}\lim_{\gamma\to 0}\bar{m}(-\gamma) &= \frac{1}{d}\mathrm{tr}\{\zeta\mathbf{I}_d\} \\ &= \zeta \end{aligned} \tag{A.51}$$

So to compute the DEs for the RP-RLDA ensemble, substitute $\lim_{\gamma\to 0}\bar{m}(-\gamma)$ for $\bar{m}(-\gamma)$ in all expressions.

A.1.2 DE of $\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j\right]$

In a similar fashion to the previous term we express

$$\sum_{i\neq j}\tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j = \sum_{i\neq j}\tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T\left(\sum_{k=1}^{p}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_j$$

Note that this summation can be expressed in matrix form as $\tilde{\mathbf{a}}^T\left(\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}}-\mathrm{diag}(\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}})\right)\tilde{\mathbf{b}}$, where $\mathbf{A}'=(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I})^{-1}$. This is important for later derivations.

First, remove $\tilde{\mathbf{r}}_i$ from the middle term of the quadratic form by applying the matrix inversion lemma (see page 11 in [11]).

$$\begin{aligned}\tilde{\mathbf{r}}_i^T\left(\sum_{k=1}^{p}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_j &= \tilde{\mathbf{r}}_i^T\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}+d_i\tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^T\right)^{-1}\tilde{\mathbf{r}}_j\\ &= \frac{\tilde{\mathbf{r}}_i^T\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_j}{1+d_i\tilde{\mathbf{r}}_i^T\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_i}\end{aligned}$$

Reintroducing the $\mathbb{E}_{\tilde{\mathbf{R}}}[\cdot]$ we obtain

$$\begin{aligned}&\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j\right]\\ &\quad=\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\tilde{a}_i\tilde{b}_j\frac{\tilde{\mathbf{r}}_i^T\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_j}{1+d_i\tilde{\mathbf{r}}_i^T\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_i}\right]\\ &\quad=\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\alpha_i}\right] \qquad (A.52)\end{aligned}$$

where $\zeta_{ij}=\tilde{\mathbf{r}}_i^T(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j$ and $\alpha_i=\tilde{\mathbf{r}}_i^T(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_i$. We can then proceed in a similar manner to the derivation in Section A.1.1 by expressing (A.52) as the sum of a term with α_i in the denominator replaced by its expectation with respect to $\tilde{\mathbf{r}}_i$ and a term ϵ denoting the error in doing so. We will show that

$\epsilon \xrightarrow{\text{a.s.}} 0$.

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\alpha_i}\right] &= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\alpha_i}\right] \\
&\quad + \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] \\
&\quad - \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] \\
&= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]}\right] + \epsilon
\end{aligned}$$

where

$$\epsilon = \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j d_i\zeta_{ij}(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]-\alpha_i)}{(1+d_i\alpha_i)(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right] \tag{A.53}$$

If we proceed as before, we end up with the bound

$$|\epsilon| \leq \frac{K}{\sqrt{d}}\sum_{i\neq j}|\tilde{a}_i||\tilde{b}_j| \tag{A.54}$$

which is overall $\mathcal{O}(\sqrt{d})$, since d and p are of the same order and

$$\begin{aligned}
\sum_{i\neq j}|\tilde{a}_i||\tilde{b}_j| &= \sum_{i,j}|\tilde{a}_i||\tilde{b}_j| - \sum_i|\tilde{a}_i||\tilde{b}_i| \\
&= \sum_i|\tilde{a}_i|\sum_i|\tilde{b}_i| - \sum_i|\tilde{a}_i\tilde{b}_i| \\
&\sim \mathcal{O}(p) - \mathcal{O}(\sqrt{p}) \\
&\sim \mathcal{O}(p)
\end{aligned}$$

where the second to last line uses the l_1, l_2-norm equivalence relation along with the assumption that $\sum_i \tilde{a}_i^2 < \infty$ as follows:

$$\sum_i \tilde{a}_i \leq \sum_i |\tilde{a}_i| = ||\tilde{\mathbf{a}}||_1 \leq \sqrt{p}\ ||\tilde{\mathbf{a}}||_2 = \sqrt{p}\sqrt{\sum_i \tilde{a}_i^2} \sim \mathcal{O}(\sqrt{p})$$

and similarly for $\sum_i |\tilde{b}_i|$ and $\sum_i |\tilde{a}_i\tilde{b}_i|$. This result is misleading and suggests that the error is growing asymptotically. In reality, there is a compensation effect in the interaction of the terms. We can see this if we express everything in terms of matrices,

$$\begin{aligned}
\epsilon &= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j} \frac{\tilde{a}_i\tilde{b}_j d_i \zeta_{ij}(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)}{(1+d_i\alpha_i)(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right] \\
&= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j} \frac{\tilde{a}_i\tilde{b}_j d_i \beta_{ij}(1+d_i\alpha_i)(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)}{(1+d_i\alpha_i)(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right] \\
&= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j} \frac{\tilde{a}_i\tilde{b}_j d_i \tilde{\mathbf{r}}_i^T\mathbf{A}'\tilde{\mathbf{r}}_j(1+d_i\alpha_i)(\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] - \alpha_i)}{(1+d_i\alpha_i)(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])}\right] \\
&= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\tilde{\mathbf{a}}^T\mathbf{BCDFG}\left(\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}} - \mathrm{diag}(\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}})\right)\tilde{\mathbf{b}}\right]
\end{aligned} \tag{A.55}$$

where $\mathbf{A}' = (\sum_{k=1}^p d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T + \gamma\mathbf{I})^{-1}$, $\beta_{ij} = \tilde{\mathbf{r}}_i^T\mathbf{A}'\tilde{\mathbf{r}}_j$, $\alpha_i = \tilde{\mathbf{r}}_i^T(\sum_{k\neq i} d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T + \gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_i$, $\mathbf{B} = \mathrm{diag}\left(\frac{1}{1+d_1\alpha_1}, \ldots, \frac{1}{1+d_p\alpha_p}\right)$, $\mathbf{C} = \mathrm{diag}\left(\frac{1}{1+d_1\mathbb{E}_{\tilde{\mathbf{r}}_1}[\alpha_1]}, \ldots, \frac{1}{1+d_p\mathbb{E}_{\tilde{\mathbf{r}}_p}[\alpha_p]}\right)$, $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$, $\mathbf{F} = \mathrm{diag}(\mathbb{E}_{\tilde{\mathbf{r}}_1}[\alpha_1] - \alpha_1, \ldots, \mathbb{E}_{\tilde{\mathbf{r}}_p}[\alpha_p] - \alpha_p)$, and $\mathbf{G} = \mathrm{diag}(1+d_1\alpha_1, \ldots, 1+d_p\alpha_p)$. The quantity ζ_{ij} is expressed in terms of β_{ij} in order to remove the dependence on i in the original $\mathbf{A}$ in ζ_{ij} so that the expression can be written out in matrices. We then have

$$\begin{aligned}
|\epsilon| &\leq \mathbb{E}_{\tilde{\mathbf{R}}}\left[\left|\tilde{\mathbf{a}}^T\mathbf{BCDFG}\left(\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}} - \mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})\right)\tilde{\mathbf{b}}\right|\right] \\
&\leq \mathbb{E}_{\tilde{\mathbf{R}}}\left[||\mathbf{GFDCB}\tilde{\mathbf{a}}||_2||\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}} - \mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})||_2||\tilde{\mathbf{b}}||_2\right] \\
&= \mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\sum_i g_i^2 f_i^2 d_i^2 c_i^2 b_i^2 \tilde{a}_i^2}\ ||\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}} - \mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})||_2||\tilde{\mathbf{b}}||_2\right]
\end{aligned}$$

where the second line follows from applying the Cauchy-Schwarz inequality and the subordinance property of matrix norms. The lower-case letters g_i, f_i, d_i, c_i, b_i and

a_i denote the ith diagonal entry of the matrix with the corresponding upper-case letter. As noted before, both $||\tilde{\mathbf{a}}||_2$ and $||\tilde{\mathbf{b}}||_2$ can be shown to be $\mathcal{O}(1)$, assuming that $||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2 < \infty$, which is an assumption of the growth regime. In addition,

$$b_i^2 = \frac{1}{(1 + d_i\alpha_i)^2} \leq 1$$

since $d_i\alpha_i \geq 0$ so that $\frac{1}{(1+d_i\alpha_i)^2} \leq 1, \ \forall i$. Similarly,

$$c_i^2 = \frac{1}{(1 + d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2} \leq 1$$

since $d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i] \geq 0$ so that $\frac{1}{(1+d_i\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2} \leq 1, \ \forall i$. Also,

$$g_i^2 = (1 + d_i\alpha_i)^2 \leq \frac{1}{\gamma}$$

In addition, $d_i^2 < \infty$ follows from $||\boldsymbol{\Sigma}||_2 < \infty$, an assumption of the growth regime. We can group all of these terms into a single constant so that we have

$$|\epsilon| \leq K\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sqrt{\sum_i f_i^2\tilde{a}_i^2}\ ||\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}} - \mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})||_2\right]$$

The quantity $||\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}} - \mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})||_2 \leq ||\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}}||_2 + ||\mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})||_2$ can also be shown to be $\mathcal{O}(1)$. First of all, it is an established fact that terms of the form $||\tilde{\mathbf{R}}^T\mathbf{A}'\tilde{\mathbf{R}}||_2$ are bounded. The remaining term $||\mathrm{diag}(\mathbf{R}^T\mathbf{A}'\tilde{\mathbf{R}})||_2$ is also bounded by the following result relating the spectral norm of any matrix $\boldsymbol{\Omega}$ and its individual elements:

$$\begin{aligned}
||\mathbf{\Omega}||_2 &= \sqrt{\sup \frac{\mathbf{x}^T \mathbf{\Omega}\mathbf{\Omega}^T \mathbf{x}}{||\mathbf{x}||_2^2}} \\
&= \sup_{||\mathbf{y}||_2=1} \sup_{||\mathbf{x}||_2=1} \mathbf{x}^T \mathbf{\Omega} \mathbf{y} \\
&\geq \mathbf{e}_i^T \mathbf{\Omega} \mathbf{e}_j \\
&= \mathbf{\Omega}_{ij}
\end{aligned}$$

where $\mathbf{e}_i$ is a zero vector with a single 1 in the ith entry. It follows that $||\text{diag}(\mathbf{\Omega})||_2 = \max_{i,j} |\mathbf{\Omega}_{ij}| \leq ||\mathbf{\Omega}||_2$. By combining this $\mathcal{O}(1)$ term with the constant K, we have

$$\begin{aligned}
|\epsilon| &\leq K' \mathbb{E}_{\tilde{\mathbf{R}}} \left[\sqrt{\sum_i f_i^2 \tilde{a}_i^2} \right] \\
&\leq K' \sqrt{\mathbb{E}_{\tilde{\mathbf{R}}} \left[\sum_i f_i^2 \tilde{a}_i^2 \right]} \\
&= K' \sqrt{\sum_i \tilde{a}_i^2 \mathbb{E}_{\tilde{\mathbf{R}}} \left[f_i^2 \right]} \\
&= K' \sqrt{\sum_i \tilde{a}_i^2 \mathbb{E}_{\tilde{\mathbf{R}}} \left[(\alpha_i - \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2 \right]} \\
&\leq K' \sqrt{\sum_i \tilde{a}_i^2 \mathbb{E}_{\tilde{\mathbf{R}}} \left[(\alpha_i - \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2 \right]}
\end{aligned}$$

$\mathbb{E}_{\tilde{\mathbf{R}}} \left[(\alpha_i - \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i])^2 \right]$ is $\mathcal{O}\left(\frac{1}{d}\right)$ while $\sum_i \tilde{a}_i^2 = ||\tilde{\mathbf{a}}||_2$ is $\mathcal{O}(1)$. Thus the overall bound is $\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$ and we have $\epsilon \xrightarrow{\text{a.s.}} 0$. We are now left with the term

$$\mathbb{E}_{\tilde{\mathbf{R}}} \left[\sum_{i \neq j} \frac{\tilde{a}_i \tilde{b}_j \zeta_{ij}}{1 + d_i \mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]} \right] \tag{A.56}$$

We can make use of the result (the trace lemma)

$$\mathrm{Var}\left[\frac{1}{p}\mathrm{tr}\left\{\left(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I}\right)^{-1}\right\}\right]=\mathcal{O}\left(\frac{1}{p^2}\right) \tag{A.57}$$

combined with

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\alpha_i\right]&=\mathbb{E}_{\tilde{\mathbf{r}}_i}\left[\tilde{\mathbf{r}}_i^T\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\tilde{\mathbf{r}}_i\right]\\&=\frac{1}{d}\mathrm{tr}\left\{\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\right\}\\&=\frac{p}{d}\frac{1}{p}\mathrm{tr}\left\{\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\right\}\end{aligned} \tag{A.58}$$

and the rank-one perturbation lemma

$$\frac{1}{p}\mathrm{tr}\left\{\left(\sum_{k\neq i}d_k\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^T+\gamma\mathbf{I}\right)^{-1}\right\}-\frac{1}{p}\mathrm{tr}\left\{\left(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I}\right)^{-1}\right\}\xrightarrow{\text{a.s.}}0 \tag{A.59}$$

to justify substituting $\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]$ by its expectation $\mathbb{E}_{\tilde{\mathbf{R}}}[\mathbb{E}_{\tilde{\mathbf{r}}_i}[\alpha_i]]=\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i]$ in (A.56) (by (A.59), the expression in (A.58) converges almost surely to the expression involved in the result of (A.57) but multiplied by the factor $\frac{p}{d}$). The fact that the variance is vanishing means that we can expect the error in performing this substitution to converge almost surely to zero. This can be proven by expressing the error in terms of matrices in a similar manner to that shown previously. We are then left with

$$\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\zeta_{ij}}{1+d_i\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i]}\right]=\sum_{i\neq j}\frac{\tilde{a}_i\tilde{b}_j\mathbb{E}_{\tilde{\mathbf{R}}}\left[\zeta_{ij}\right]}{1+d_i\mathbb{E}_{\tilde{\mathbf{R}}}[\alpha_i]}=0 \tag{A.60}$$

since $\mathbb{E}_{\tilde{\mathbf{R}}}\left[\zeta_{ij}\right]=0$.

The conclusion here is that, asymptotically, the term $\mathbb{E}_{\tilde{\mathbf{R}}}\left[\sum_{i\neq j}\tilde{a}_i\tilde{b}_j\tilde{\mathbf{r}}_i^T(\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T+\gamma\mathbf{I})^{-1}\tilde{\mathbf{r}}_j\right]$ tends almost surely to 0. The DE for m_0 is wholly dependent on the first term of

(A.4) whose DE is derived in Section A.1.1 of the appendix.

A.2 Derivation of the DE for m_1

By symmetry, we have that the deterministic sequence

$$\begin{aligned}
&\bar{m}_1 = \\
&\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{V}\ \mathbf{T}\left(\frac{1}{\bar{m}(-\gamma)}\right)\mathbf{V}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\
&\quad + \frac{1}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right)\operatorname{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(\frac{1}{\bar{m}(-\gamma)}\right)\right\} + \ln\frac{\pi_1}{\pi_0}
\end{aligned} \tag{A.61}$$

satisfies

$$m_1 - \bar{m}_1 \xrightarrow{\text{a.s.}} 0$$

A.3 Derivation of the DE for σ^2

Recall that the variance of the discriminant $\hat{W}^{\infty-\text{ens}}_{\text{RP-RLDA}}(\mathbf{x}_q)$ conditioned on the training data and class of $\mathbf{x}_q$ is given by

$$\sigma^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \mathbb{E}_{\mathbf{R}}\left[\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R}\right]\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{R}}\left[\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R}\right]^T(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) \tag{A.62}$$

Define $\mathbf{t} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ and $\mathbf{v} = \boldsymbol{\Sigma}\mathbb{E}_{\mathbf{R}}\left[\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R}\right]^T(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$, then we can rewrite σ^2 as

$$\sigma^2 = \mathbb{E}_{\mathbf{R}}\left[\mathbf{t}^T\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R}\mathbf{v}\right] \tag{A.63}$$

With the variance in this form, we can reuse much of the computation involved for the mean, provided that $\tilde{\mathbf{t}} = \mathbf{U}^T\mathbf{t} \sim \mathcal{O}(1)$ and $\tilde{\mathbf{v}} = \mathbf{U}^T\mathbf{v} \sim \mathcal{O}(1)$, where $\mathbf{U}$ comes from the diagonalization of $\hat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$. This can be shown to hold. We then have the result, directly from the mean derivation,

$$\sigma^2 - \mathbf{t}^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \mathbf{v} \xrightarrow{\text{a.s.}} 0 \tag{A.64}$$

We can further develop the second term of (A.64) as

$$\begin{aligned}
&\mathbf{t}^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \mathbf{v} \\
&\quad = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{\Sigma}\mathbb{E}_{\mathbf{R}}\left[\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R}\right]^T (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) \\
&\quad = \mathbb{E}_{\mathbf{R}}\left[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{\Sigma}(\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R})^T (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) \right] \\
&\quad = \mathbb{E}_{\mathbf{R}}\left[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{\Sigma}\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \gamma\mathbf{I})^{-1}\mathbf{R}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) \right]
\end{aligned}$$

Now again let $\bar{\mathbf{t}}^T = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}[m(-\gamma)]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{\Sigma}$ and $\bar{\mathbf{v}} = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$. Since $\mathbf{U}^T\bar{\mathbf{t}} \sim \mathcal{O}(1)$ and $\mathbf{U}^T\bar{\mathbf{v}} \sim \mathcal{O}(1)$, we can apply the result from the mean derivation once again and obtain

$$\mathbf{t}^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \mathbf{v} - \bar{\mathbf{t}}^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \bar{\mathbf{v}} \xrightarrow{\text{a.s.}} 0 \tag{A.65}$$

The second term expands to

$$\begin{aligned}
&\bar{\mathbf{t}}^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \bar{\mathbf{v}} \\
&\quad = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{\Sigma} \left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I} + \hat{\boldsymbol{\Sigma}} \right)^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)
\end{aligned} \tag{A.66}$$

Just as in the mean derivation, we substitute

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

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

Substituting these into (A.66) and then taking the expectation over $\mathbf{Z}_0\mathbf{1}$ and $\mathbf{Z}_1\mathbf{1}$ yields

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}_0\mathbf{1},\mathbf{Z}_1\mathbf{1}}\left[(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0)\right]\\
&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\\
&\quad+\frac{1}{n_0}\mathrm{tr}\left\{\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}\right\}\\
&\quad+\frac{1}{n_1}\mathrm{tr}\left\{\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}\boldsymbol{\Sigma}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\hat{\boldsymbol{\Sigma}}\right)^{-1}\right\}\\
&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\mathbf{V}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\frac{1}{n}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\right)^{-1}\mathbf{D}_{\boldsymbol{\Sigma}}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\frac{1}{n}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\right)^{-1}\mathbf{V}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\\
&\quad+\frac{1}{n_0}\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m_{\tilde{\mathbf{R}}\mathbf{D}\tilde{\mathbf{R}}^T}(-\gamma)\right]}\mathbf{I}+\frac{1}{n}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\right)^{-1}\mathbf{D}_{\boldsymbol{\Sigma}}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\frac{1}{n}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\right)^{-1}\right\}\\
&\quad+\frac{1}{n_1}\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\frac{1}{n}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\right)^{-1}\mathbf{D}_{\boldsymbol{\Sigma}}\left(\frac{1}{\mathbb{E}_{\tilde{\mathbf{R}}}\left[m(-\gamma)\right]}\mathbf{I}+\frac{1}{n}\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\mathbf{W}\mathbf{W}^T\mathbf{D}_{\boldsymbol{\Sigma}}^{1/2}\right)^{-1}\right\}
\end{aligned}\tag{A.67}$$

Based on this and some results in [13], the deterministic sequence

$$\begin{aligned}
&\bar{\sigma}^2 =\\
&\kappa(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\mathbf{V}\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\mathbf{V}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\\
&\quad+\kappa\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(-\frac{1}{\bar{m}(-\gamma)}\right)\right\}
\end{aligned}\tag{A.68}$$

satisfies

$$\sigma^2-\bar{\sigma}^2\xrightarrow{\text{a.s.}}0$$

where $\kappa=\frac{(n/p)^2}{(n/p)^2-(n/p)\frac{1}{(1+e(z))^2}\Omega}$, with $\Omega=\frac{1}{p}\mathrm{tr}\left\{\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{T}\left(z\right)\mathbf{D}_{\boldsymbol{\Sigma}}\ \mathbf{T}\left(z\right)\right\}$ and $z=-\frac{1}{\bar{m}(-\gamma)}$.

A.4 Verifying Convergence by Simulation

This section presents simulations to verify the claims of almost-sure convergence in (3.17) and (3.16). We provide simulations for both the general covariance and isotropic covariance assumptions whose DE computation procedures are outlined in Section 4.1.1 as well as for the RP-LDA ensemble classifier whose DE computation procedure is outlined in Section 4.1.2.

For each of these cases, simulations are presented in the case of balanced ($n_0 = n_1$) and unbalanced ($n_0 \neq n_1$) number of samples from each class. The data is generated synthetically and under the assumption of stratified sampling, meaning that two classes are sampled independently of each other and so n_0 and n_1 cannot be used to estimate the prior probabilities [4]. We therefore assume the prior probabilities are known and use them directly. The data is then generated such that $\frac{n_0}{n} \approx \pi_0$ and $\frac{n_1}{n} \approx \pi_1$. For the balanced case, $\pi_0 = \pi_1 = 0.5$ and for the unbalanced case we used $\pi_0 = 0.7$ and $\pi_1 = 0.3$.

To check the validity of the deterministic equivalents (3.25), (3.26), and (3.27) by simulation, we compute them and compare them to their respective counterparts (3.9), (3.10), and (3.11) which are computed by averaging over many realizations of $\mathbf{R}$ for fixed training data, as d, p, and n grow at a constant rate to each other. We then compute the variance between the pairs over multiple realizations of the training set. The logarithm of the variance of the deviation between each pair is then plotted against the logarithm of the dimension d. A slope of -1 indicates that the variance decays with order $\frac{1}{d}$.

In the following simulations, the class means are set to $\mu_0 = \frac{1}{p^{1/4}} \left[\mathbf{1}^T_{\lceil \sqrt{p} \rceil} \ \mathbf{0}^T_{p - \lceil \sqrt{p} \rceil - 2} \ 2 \ 2 \right]^T$ and $\mu_1 = \mathbf{0}_p$ and the number of projections is set to $M = 200$. The dimensions are $p = \{80, 120, 240, 360, 480, 600, 720, 840\}$, $n = \lceil 0.75p \rceil$, and $d = \lceil 0.5p \rceil$, or in other words, $c = 0.75$ and $c' = 0.5$.

To check the general covariance DE computation procedure of the RP-RLDA ensemble classifier detailed in Section 4.1.1, we use $\mathbf{\Sigma}$ such that $\mathbf{\Sigma}_{ij} = 0.6^{|i-j|},\ \forall i, j$ and set $\gamma = 1$. The corresponding figures in the balanced and unbalanced case are Figure A.1 and Figure A.2 respectively.

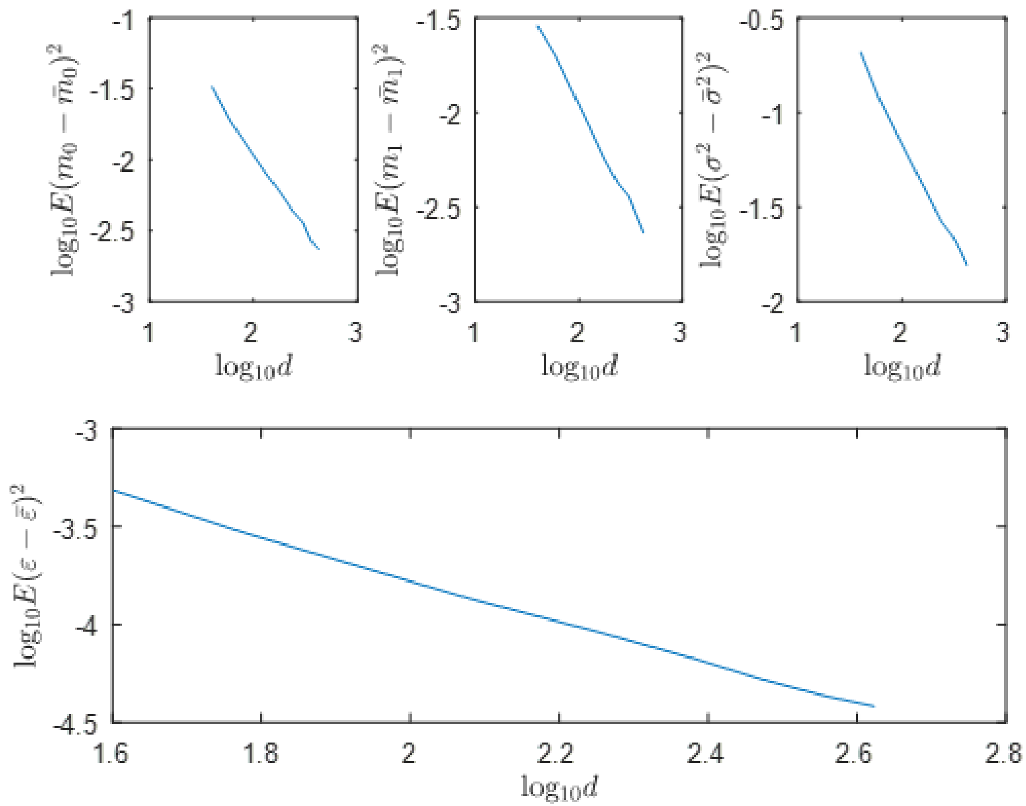

Figure A.1: Convergence check for the balanced case of the RP-RLDA ensemble classifier for general covariance with $M = 200$ and $\gamma = 1$.

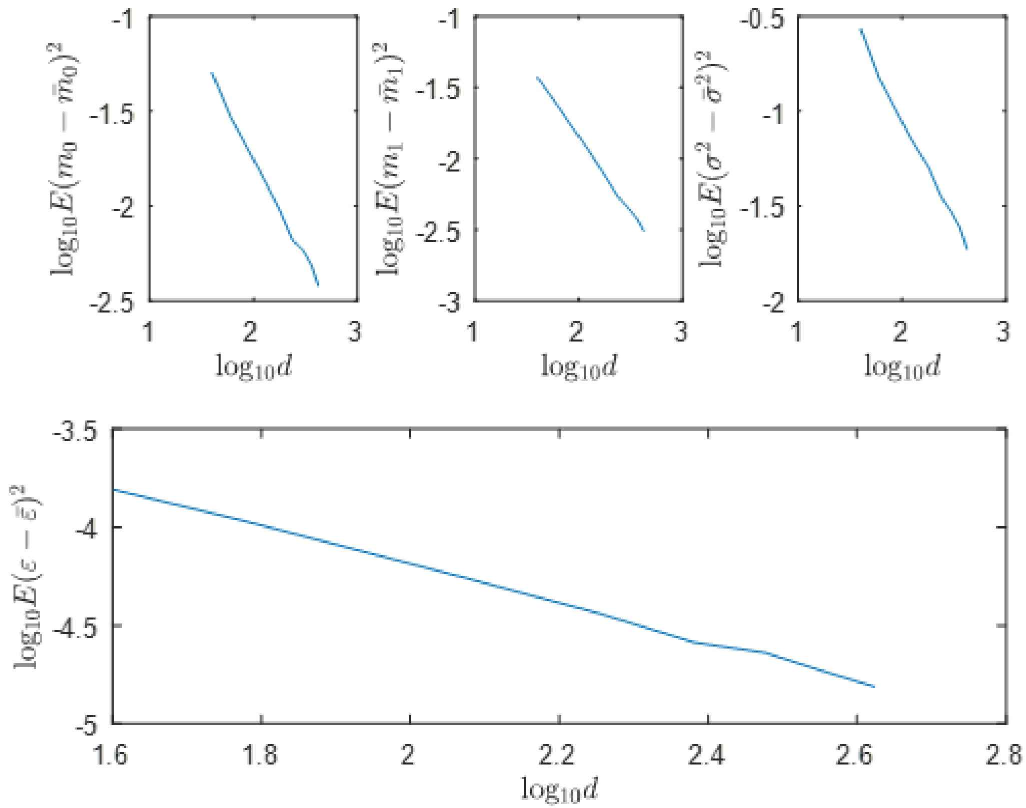

Figure A.2: Convergence check for the unbalanced case of the RP-RLDA ensemble classifier for general covariance with $M = 200$ and $\gamma = 1$.

To check the isotropic covariance DE computation procedure of the RP-RLDA ensemble classifier detailed in Section 4.1.1, we use $\mathbf{\Sigma} = \mathbf{I}_p$ and set $\gamma = 1$. The corresponding figures in the balanced and unbalanced case are Figure **??** and Figure **??** respectively.

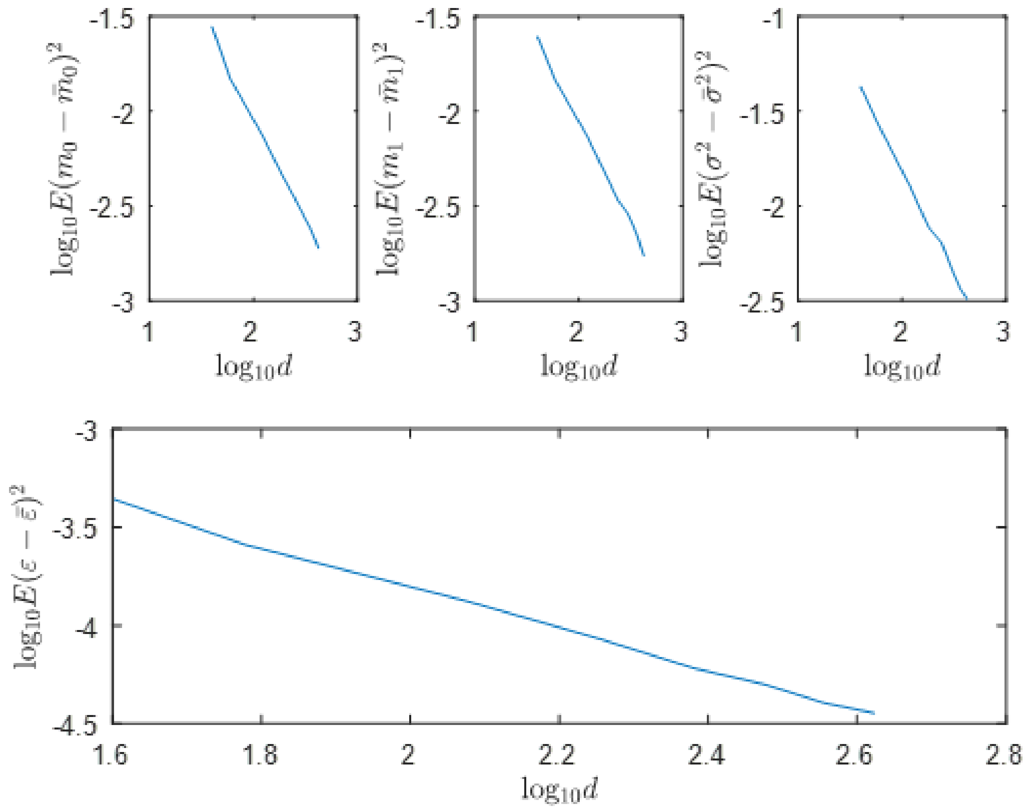

Figure A.3: Convergence check for the balanced case of the RP-RLDA ensemble classifier for isotropic covariance with $M = 200$ and $\gamma = 1$.

To check the DE computation procedure of the RP-LDA ensemble classifier detailed in Section 4.1.2, we use $\mathbf{\Sigma}$ such that $\mathbf{\Sigma}_{ij} = 0.6^{|i-j|}$, $\forall i, j$ and set $\gamma = 0$. The corresponding figures in the balanced and unbalanced case are Figure **??** and Figure **??** respectively.

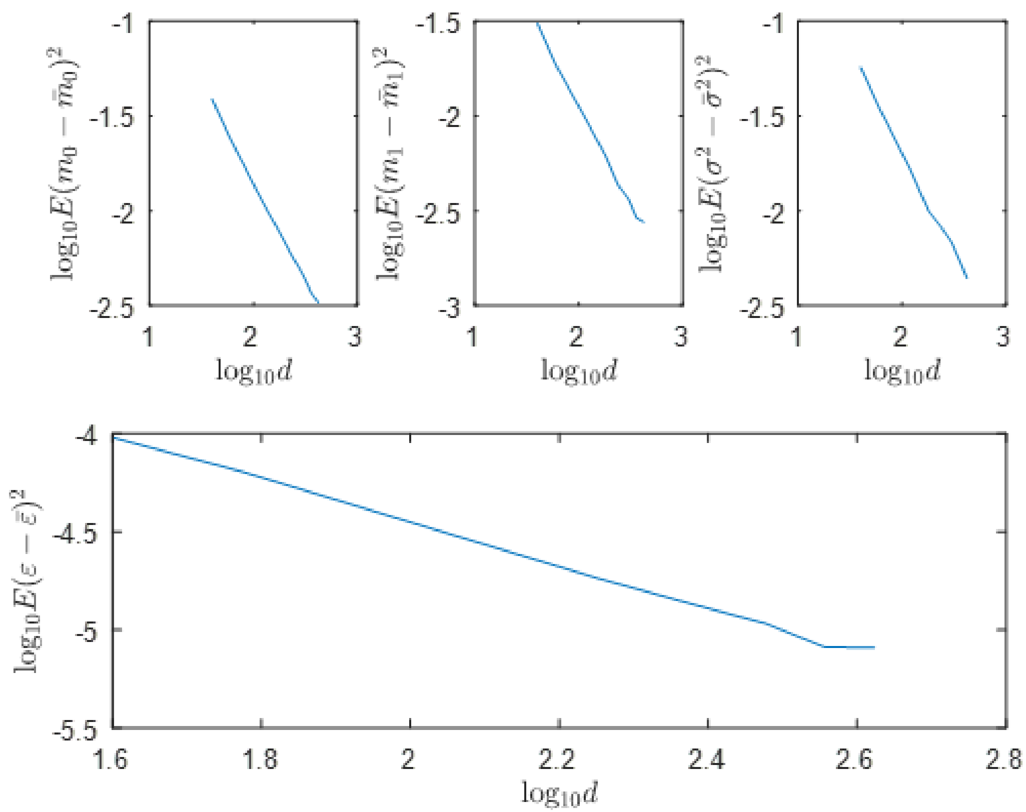

Figure A.4: Convergence check for the unbalanced case of the RP-RLDA ensemble classifier for isotropic covariance with $M = 200$ and $\gamma = 1$.

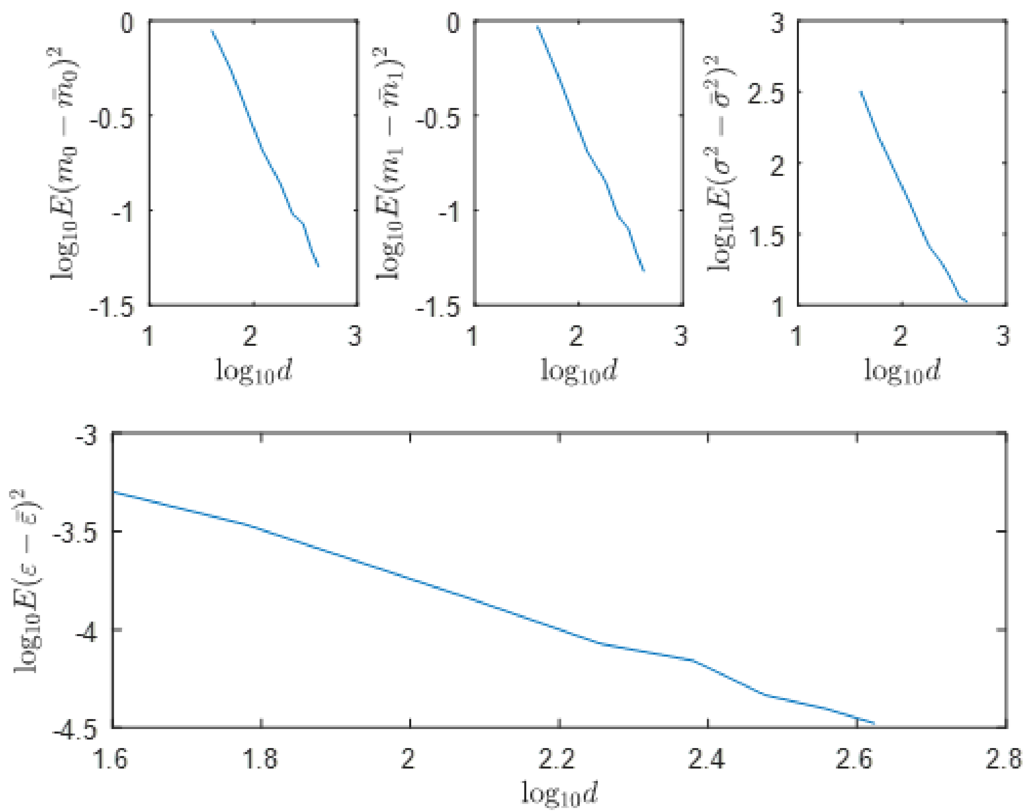

Figure A.5: Convergence check for the balanced case of the RP-LDA ensemble classifier with $M = 200$ and $\gamma = 0$.

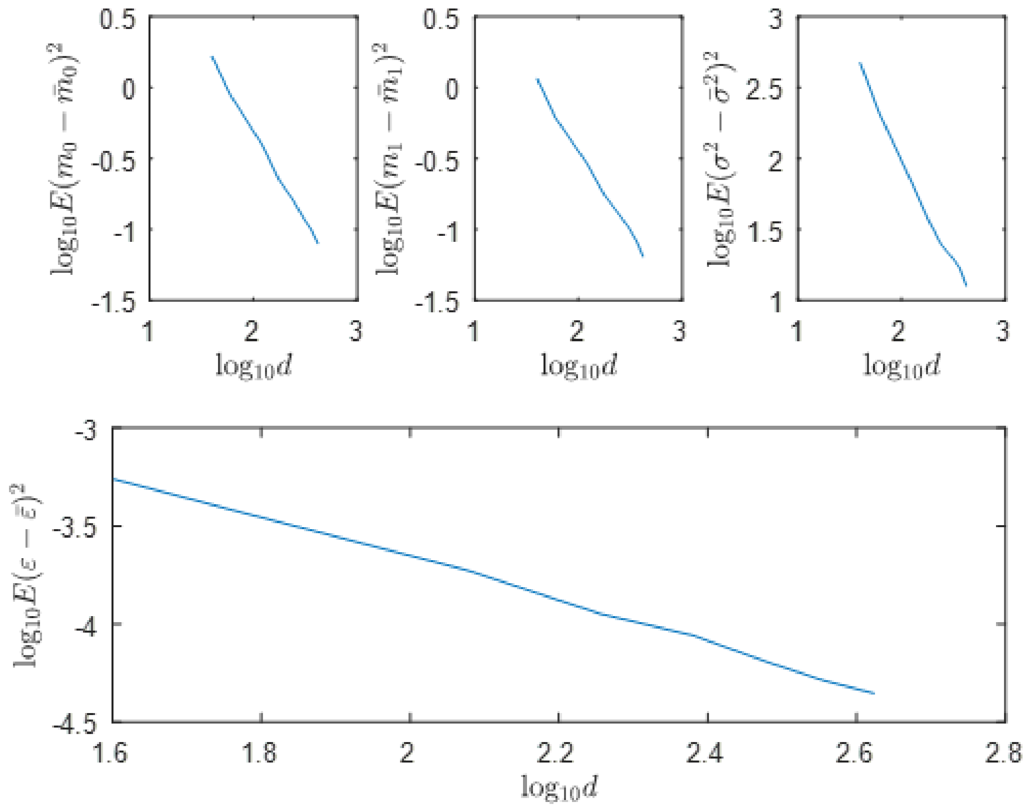

Figure A.6: Convergence check for the unbalanced case of the RP-LDA ensemble classifier with $M = 200$ and $\gamma = 0$.