Learning in High Dimensions with Projected Linear Discriminants

By

Robert John Durrant

A thesis submitted to the
University of Birmingham
for the degree of
Doctor of Philosophy

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
January 2013

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Robert John Durrant

Abstract

The enormous power of modern computers has made possible the statistical modelling of data with dimensionality that would have made this task inconceivable only decades ago. However, experience in such modelling has made researchers aware of many issues associated with working in high-dimensional domains, collectively known as ‘the curse of dimensionality’, which can confound practitioners’ desires to build good models of the world from these data. When the dimensionality is very large, low-dimensional methods and geometric intuition both break down in these high-dimensional spaces.

To mitigate the dimensionality curse we can use low-dimensional representations of the original data that capture most of the information it contained. However, little is currently known about the effect of such dimensionality reduction on classifier performance. In this thesis we develop theory quantifying the effect of random projection – a recent, very promising, non-adaptive dimensionality reduction technique – on the classification performance of Fisher’s Linear Discriminant (FLD), a successful and widely-used linear classifier. We tackle the issues associated with small sample size and high-dimensionality by using randomly projected FLD ensembles, and we develop theory explaining why our new approach performs well. Finally, we quantify the generalization error of Kernel FLD, a related non-linear projected classifier.

Dedication

To my family.

Acknowledgments

It is traditional to begin one's thesis by thanking those individuals without whom it would not have been completed, and since it is a good tradition I don't intend to break with it.

My wife, Sam, has been an unfailing source of support for as long as we have known each other, and her love, patience and good humour have kept me going much more than I think she knows. Thank you Sam. On the other hand my daughters Amy (8) and Lucy (2) have been an unfailing source of distraction; however in my experience it sometimes pays to be distracted for a while, and besides that is how things are supposed to be. Thank you girls. Without my family, this would have been a lonelier and harder road, and I dedicate this thesis to them.

I would also like to especially thank my supervisor, Dr. Ata Kabán, for her encouragement, patience, and support – and especially for starting the ball rolling by convincing me to attempt some theory during my Masters' degree. We have had a productive partnership, and the contents of this thesis are the fruit of our collaborative effort. I would also like to thank Ata for reading my thesis drafts and for finding in them more bugs than I imagined were possible. The usual caveat that any remaining errors are my own is, of course, still true.

I would also like to thank the other researchers I have been privileged to meet during the course of this research, too many to name, for interesting and often enlightening discussions and ideas. In particular, thanks are due to Peter Tiňo, Jakramate Bootkrajang, and Hector Basevi for asking hard questions in the measure concentration group, and to my thesis group members Steve Vickers and Iain Styles for keeping me on track and encouraging me to keep at it when sometimes things weren't going as well as planned.

Thanks are also due to my examiners Professor John Shawe-Taylor and Peter Tiňo, as well as chair Iain Styles, for taking time out of their busy schedules to examine my thesis.

Finally I would like to thank the School of Computer Science at the University of Birmingham for their generous financial support which enabled me to spend my time in research, and also the School Research Committee, Royal Academy of Engineering, and KDD 2010 and ECML 2012 programme committees, for financial support towards attending conferences, schools and workshops.

Bob Durrant,
Halesowen, January 2013

Publications

During the course of this project, the following peer-reviewed papers have been published in journals and conference proceedings, and the following technical reports were written:

Journal Papers

R. Durrant and A. Kabán. *A tight bound on the performance of Fisher's linear discriminant in randomly projected data spaces.* Pattern Recognition Letters (2011).

R. Durrant and A. Kabán. *Random Projections as Regularizers: Learning a Linear Discriminant Ensemble from Fewer Observations than Dimensions.* (Submitted).

Refereed Conference Papers

R. Durrant and A. Kabán. *Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data.* In *Proceedings16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)* (2010).

R. Durrant and A. Kabán. *A bound on the performance of LDA in randomly projected data spaces.* In *Proceedings 20th International Conference on Pattern Recognition (ICPR 2010)*, pp. 4044–4047 (2010). **IBM Best Student Paper Award**– Machine Learning and Pattern Recognition Track.

R. Durrant and A. Kabán. *Error bounds for Kernel Fisher Linear Discriminant in Gaussian Hilbert space.* In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AIStats 2012)* (2012).

Technical Reports

R. Durrant and A. Kabán. *Flip probabilities for random projections of θ-separated vectors.* University of Birmingham, School of Computer Science Technical Report No. CSR-10-10.

R. Durrant and A. Kabán. *A comparison of the moments of a quadratic form involving orthonormalised and normalised random projection matrices.* University of Birmingham, School of Computer Science Technical Report No. CSR-11-04.

Fare forward, travellers! not escaping from the past
Into different lives, or into any future;
You are not the same people who left that station
Or who will arrive at any terminus,
While the narrowing rails slide together behind you;
And on the deck of the drumming liner
Watching the furrow that widens behind you,
You shall not think 'the past is finished'
Or 'the future is before us'.
At nightfall, in the rigging and the aerial,
Is a voice descanting (though not to the ear,
The murmuring shell of time, and not in any language)
'Fare forward, you who think that you are voyaging;
You are not those who saw the harbour
Receding, or those who will disembark.'

- T.S.Eliot - 'The Dry Salvages'

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Statement

Machine learning (‘learning’) is the computer analogue of ‘learning from experience’. If we take $\mathcal{H}$ to be a class of learning functions then such learning comprises finding the (parameters of a) function $\hat{h} \in \mathcal{H}$ which minimises the expected error of the learned function given an observation $x_q \sim \mathcal{D}_{x|y}$. Usually the data generating distribution $\mathcal{D}_{x,y}$ is unknown, and so it must be estimated from a collection of examples (‘training set’) $\mathcal{T} \overset{i.i.d}{\sim} \mathcal{D}_{x,y}$. Typically the function evaluation $h(x_q)$ is either a class label or a real number depending on whether the problem to be solved is a classification problem or a regression problem respectively. Learning algorithms may also be further classified by the learning approach employed ‘supervised’, ‘unsupervised’, ‘semi-supervised’, ‘reinforcement’ and so on depending on the information that is available to the learner.

Now consider the case where the data observed have very high dimensionality, for example they are real valued vectors with arity in the thousands. Such high dimensional (HD) datasets arise in many areas of practical application of machine learning approaches, from gene array datasets to dictionaries for spam filters, and these HD data exhibit characteristics leading to a range of problems collectively known as ‘the curse of dimensionality’ (Bellman, 1970) that can make learning from the HD data difficult.

Two common situations in practice are:

When data are high dimensional and cheap to collect (a common situation in domains such as retail and web-mining) then both the dimensionality, d, and the number of observations, N, is large. In principle there should be no problem with statistical inference in this setting, however in practice when $N \gg d$ and d is large there are time- and space-complexity issues arising which can mean that efficient algorithms, i.e. those polynomial in d, run too slowly to be of practical use. In particular fitting the data in

memory can be a problem, and the speed of data processing is then bottle-necked by the speed of disk access.
On the other hand when data are expensive to collect then typically $N \ll d$ and the main issues in this setting concern bogus interactions in the training data and the quality of parameter estimates. This is an increasingly common situation in scientific domains such as medical imaging and bioinformatics.
Some further aspects of the curse include that, as data dimensionality grows:

- The number of data examples required to estimate the distribution of the data, to some fixed level of confidence, grows exponentially with the dimensionality (Hastie et al., 2001).

- If the relative variance in the pairwise distances between points, as measured in *any* metric, goes to zero then the distance between the closest together and farthest apart points becomes almost the same (and vice-versa); hence the concept of near neighbours becomes meaningless (Beyer et al., 1999; Durrant & Kabán, 2009). It can be shown that in high dimensions this happens under quite general conditions.

- Our intuition regarding the geometric structure of data lets us down in high dimensions. A simple, but powerful, example is the fact that the Euclidean norm of a point drawn from the univariate standard normal distribution $\mathcal{N}(0,1)$ is typically not very far from zero (e.g. with probability about 1/2 its norm is no more than 2/3), whereas a point drawn from the d-dimensional standard normal distribution $N(0, I_d)$ has Euclidean norm close to $\sqrt{d}$ – with overwhelming probability when d is large (Ledoux, 2001).

This list is not exhaustive, but even from just these few examples it is apparent that HD data bring with them their own particular problems that are not observed when working in lower dimensional domains. Furthermore, the first of these problems essentially guarantees that for all HD problems one never has as much data with which to train a learner as one would like, while the computational issues mean that even when sufficient data is available processing it efficiently may be difficult. Clearly, if one could work with a lower dimensional representation of the data then these problems would be mitigated.

There are many dimensionality reduction techniques that aim to achieve low dimensional representations of HD data (see, for example, Fodor (2002) for a survey) but, as is frequently the case when working on complex problems, there is no 'silver bullet' approach guaranteed to work well on every data set.

A recent research direction in signal processing, the field of 'compressed sensing' (CS), has raised hopes however that for *sparse* data (that is, data where the observations can be represented in some basis by vectors with mostly zero components) the situation is better. We can apply a very simple, non-adaptive, and computationally cheap method of dimensionality reduction called *random projection* to such signals as follows: If the original data were d-dimensional and the number of non-zero components in some linear basis (for example a wavelet or Fourier basis) is m then, provided that the projected dimension is at least $k = \mathcal{O}(m \log d)$, with high probability the signal can be perfectly (and efficiently) reconstructed from such a projection (Candes et al., 2008; Donoho, 2006).

This exciting result implies that, with high probability, no information is lost by randomly projecting the sparse data in this way and that therefore we could learn as good a model from the k-dimensional randomly-projected data as we could from the initial high dimensional data. On the other hand, data are very frequently not sparse in a linear basis. For example, this is the case when the data are noisy or the underlying classes are not linearly separable (Eltoft & deFigueiredo, 1996; Khor et al., 2005).

Although the sparsity condition of CS has recently been shown to imply a learning guarantee for randomly-projected soft-margin SVM (Calderbank et al., 2009), it is not clear if this condition is actually required for a learning task such as classification. Indeed, earlier empirical and theoretical results concerning classification in randomly-projected domains (e.g. Arriaga & Vempala, 1999; Bingham & Mannila, 2001; Fradkin & Madigan, 2003a; Indyk & Motwani, 1998) suggest that sparsity of the data may not be a requirement for successful classification. This thesis will study this issue by determining what dimensionality k is required for classification from random projections of data to succeed without assuming any sparse structure on the data, and which quantities k depends on.
The starting point for this investigation is the observation that classification is an intrinsically simpler task than the perfect reconstruction of signals. This is not a novel observation, indeed it is the fundamental idea underlying Support Vector Machine and other discriminative classifiers, but it leads us to conjecture that it may be possible to relax some of the constraints that apply to compressed sensing whilst still maintaining learning performance working with random projections of the data.

Our main focus in this thesis is on random projections and learning in randomly-projected domains. This may seem counterintuitive since it is easy to believe that the best possible dimensionality reduction scheme in any particular situation would be one which took account of properties of the data relevant to the problem at hand, and therefore random projection must be a sub-optimal approach to dimensionality reduction. However, a key motivation in studying random projections is that we are not aware of any non-adaptive deterministic dimensionality reduction approach for which one can derive the kind of strong guarantees that we will be presenting later in this thesis. Furthermore randomness is now a standard tool in algorithm design; for instance, it is the basis of the only known polynomial-time algorithm for primality testing (Dasgupta et al., 2008) and of the only known algorithms for reaching consensus in distributed systems in finite time (Dubhashi & Panconesi, 2012; Panconesi, 2012).

1.2 Research Questions

We are interested in the effect of random projection on the generalization performance of a classifier. Examples of questions we would like to know the answers to include:

- What qualitative and quantitative guarantees can we give for classifiers working with randomly-projected data?

  - Under what conditions does random projection degrade generalization performance, and why?
  - Under what conditions does random projection improve generalization performance, and why?
  - Can we give sharp estimates of the level of degradation or improvement?
  - Are there intuitive ways to understand the effect of random projections on classifier performance?

- To what extent do the data dimensionality and projection dimensionality affect the generalization performance of a randomly projected classifier? What is the interplay between them?

- Are Johnson-Lindenstrauss- and compressed sensing-type results the only ways in which to quantify the effect of random projections? In particular, can the effect of random projection on classification be quantified without resorting to uniform geometry preservation, or without requiring data to be sparse?

1.3 Contributions of this Thesis

The main contributions of this thesis are theoretical results focusing on randomly-projected Fisher linear discriminants (RP-FLD), an instance of which is kernel Fisher discriminant (KFLD), and on voting ensembles of randomly-projected linear discriminants. More specifically: In chapter 4 we develop some general tools bounding the generalization error of FLD in the dataspace: These we use in the subsequent chapters 5, 7 and 8 to quantify the generalization error of randomly-projected FLD classifiers (RP-FLD), ensembles of RP-FLD classifiers, and kernel Fisher discriminant classifiers (KFLD).

In chapter 5 we analyse Fisher's Linear Discriminant (FLD) when the classifier is both learned from and employed on randomly-projected data. Two reasons why we find this analysis interesting are that (1) Presently such data arise in practice when random projections have been used as a preprocessing step to reduce the dimensionality of high dimensional observations, and (2) The technological promise of compressed sensing theory is that in the future (compressible) data might be collected, stored and processed in just such a compressed form.
Unlike previous analyses of other classifiers in this setting, we integrate the random projections which will reduce the dimensionality of the data and the specific classifier (FLD) to be employed on the randomly-projected data into a single random system to be analyzed, and by doing this we avoid the unnatural effects that arise when one insists that all pairwise distances are approximately preserved under projection. This enables us to take advantage of the class structure inherent in the classification problem – in particular our analysis requires no sparsity or underlying low-dimensional structure restrictions on the data.
We obtain upper bounds on the estimated misclassification error on average over the random choice of the projection, which are always non-trivial (less than 1) and which

are tight up to a small constant scaling factor. In contrast to early distance preserving approaches, our bounds tighten in a natural way as the number of training examples increases and we are able to show that, for good generalization of FLD, the required projection dimension only grows logarithmically with the number of classes. We also show that the error contribution of a covariance misspecification is always no worse in the low-dimensional projected space than in the initial high-dimensional data space. Our analysis also reveals and quantifies the effect of class 'flipping' – a potential issue when randomly projecting a finite sample.
A preliminary version of part of the work in chapter 5 received an IBM Best Student Paper Award at the 20th International Conference on Pattern Recognition.

In chapter 6 we formally derive the 'flip probability' discussed in chapter 5 and give probabilistic and geometric characterizations of this quantity. To the best of our knowledge this probability has not been exactly quantified before for any projection dimension greater than 1. Using straightforward tools we also derive a tight upper bound on the flip probability that applies in a more general setting than was analyzed earlier in the chapter. This upper bound has completely intuitive behaviour (given the findings that precede it) and reveals more about the general class-flipping problem.

In chapter 7 we examine the performance of a voting ensemble of randomly projected Fisher Linear Discriminant classifiers, focusing on the case when there are fewer training observations than data dimensions. The specific form and simplicity of this ensemble permits a direct and much more detailed analysis than existing generic tools in previous works. In particular, we are able to derive the exact form of the generalization error of our ensemble, conditional on the training set, and based on this we give theoretical guarantees which directly link the performance of the ensemble to that of the corresponding linear discriminant learned in the full data space.
Furthermore we show that the randomly projected ensemble implements a sophisticated regularization scheme to the linear discriminant learned in the original data space and this prevents overfitting in conditions of small sample size where pseudo-inverse FLD learned in the data space is provably poor.
To the best of our knowledge these are the first theoretical results to prove such an explicit link for any classifier and classifier ensemble pair.
We confirm our theoretical findings, and demonstrate the utility of our approach, with experiments on several gene array datasets from the bioinformatics domain where fewer observations than dimensions are the norm.

In chapter 8 we derive a bound on the generalization error of KFLD which, under mild assumptions, holds with high probability for any training set of a given size. KFLD can be viewed as a particular kind of instance of a randomly-projected FLD classifier, but now the training observations are replaced by their feature-mapped counterparts and the projection is onto the (random w.r.t the random training set) subspace spanned by these feature vectors. Our bound is always non-trivial (less than 1), and is given in terms of quantities in the full Hilbert space in which the feature-mapped data lie. A key term in our bound turns out to be the distance between the class mean functions scaled by the largest eigenvalue of the covariance operator; this implies that (with a suitable kernel choice) classes can always be separated in the feature space except when the densities of the two classes coincide in the input (data) space, and therefore good generalization can be achieved by KFLD as long as the original data have different class-conditional densities.

1.4 Organization of this Thesis

This thesis contains nine chapters, including this one. In the next chapter 2 we introduce the mathematical notation and conventions that we follow in the majority of this thesis (for chapter 8 we need some additional notation which we introduce there) and we survey the mathematical tools and results that we will use frequently: The majority of these are standard, and we indicate any novel lemmas of our own where they appear. In chapter 3 we briefly summarize the state of the art in learning from randomly projected data. Chapters 4, 5, 6, 7 and 8 are described above. In chapter 9 we summarize our findings, and highlight some remaining open problems and potential research directions that we believe are interesting.

2
Mathematical Background and Tools

Summary In this chapter, we review some background and tools from linear algebra, random matrix theory and statistics and applied probability that we will use later in this thesis. For standard results we give sources, and for tools which were derived as part of this research we give full proofs.

2.1 Notation and Conventions

We denote by lower-case Roman characters a column vector, e.g. v, x, and by subscripted lower-case Roman characters entries of the same vector, e.g. v_i, x_j are the i-th entry of v and the j-th entry of x respectively. We denote scalars by lower-case Greek characters, e.g. α, β, λ. We denote by upper-case Roman and Greek characters a matrix, e.g. R, Σ, Λ and we denote the entry in the i-th row and the j-th column of the matrix A by a_{ij} or $(A)_{ij}$. Transposition is indicated by a superscripted 'T', e.g. x^T is a row vector and A^T is the transpose of the matrix A. We denote the set of $k \times d$ matrices, that is matrices with k rows and d columns, by $\mathcal{M}_{k \times d}$.
We denote by μ, σ^2, and Σ the mean, variance and covariance matrix of a probability distribution and denote estimated quantities by adding a hat: $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\Sigma}$.
For a square matrix A we denote by $\lambda_{\min}(A), \lambda_{\max}(A)$ its least and greatest eigenvalues respectively, and by $\lambda_{\min \neq 0}(A)$ the least non-zero eigenvalue of A.
Throughout this thesis R is always a *random projection matrix* but it may have entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ or $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1/d)$ (i.e. with variance chosen to normalize the rows) or it may be that $R = (\tilde{R}\tilde{R}^T)^{-\frac{1}{2}}\tilde{R}$ and $\tilde{r}_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ (i.e. R has orthonormalized rows). We will make clear which particular flavour of R is being considered in the different sections of this thesis.[1]
In chapter 8 the part of R is played by an orthogonal projection onto the span of the training data, and we introduce the appropriate notation for that setting there.

2.2 Linear Algebra

Linear algebra is the study of vector spaces and linear operators (matrices). We review briefly some key facts relating to real-valued vector spaces that we will use without further reference in the following chapters - a thorough treatment of the basics can be found in Artin (2010). We also review some key properties of real positive semi-definite (p.s.d) matrices for which an exhaustive treatment can be found in Horn & Johnson (1985). Although these results typically hold for complex-valued vector spaces and p.s.d matrices too, e.g. by taking real and imaginary parts separately or replacing transposition with Hermitian transposition, here and from now on we will restrict our attention to the real-valued setting.

2.2.1 Normed Vector Spaces

Recall that a *vector space* $(V, +, \cdot)$ over a field $\mathbb{F}$ is a set of elements (vectors) that is closed under vector addition $(+)$ between elements of V and scalar multiplication $(\cdot)$ of elements of V by elements of $\mathbb{F}$. Here we will always take $\mathbb{F} = \mathbb{R}$ and we will drop the references to vector addition and scalar multiplication and simply write V for a real valued vector space. Furthermore we will usually denote scalar multiplication of the vector v by the scalar α using αv rather than $\alpha \cdot v$. Examples of vector spaces include the

[1] We shall see later that the choice of the form of random projection matrix from any of these alternatives, or from the further alternatives given in Achlioptas (2003), is not crucial to the conclusions we can draw about their effect.

real line $\mathbb{R}$ with the usual definitions of addition and multiplication, the d-dimensional real coordinate space $\mathbb{R}^d$ under element-wise addition and scalar multiplication, square $m \times m$ matrices under matrix addition and scalar multiplication, polynomials of degree no more than d with the usual definitions of addition and multiplication, and so on. A concrete example of a real vector space is $\mathbb{R}^3 = \{(x, y, z)^T : x, y, z \in \mathbb{R}\}$, i.e. the set of ordered triples of real numbers.

A vector *subspace* is a subset, $U \subseteq V$, that is a vector space in its own right when inheriting the notions of vector addition and scalar multiplication from V. Every vector space V has at least the two subspaces V and $\{0\}$; a concrete example of a non-trivial vector subspace is $\{(x, y, 0)^T : x, y \in \mathbb{R}\}$ which is a 2-dimensional subspace of $\mathbb{R}^3$.

A *normed vector space* is a vector space equipped with a positive real measure of the magnitude of a vector called a norm. The canonical example is Euclidean space, namely $\mathbb{R}^d$ equipped with the ℓ_2 (or Euclidean) norm $\|\cdot\|$ defined by:

$$\|x\| := \sqrt{\sum_{i=1}^{d} x_i^2}$$

Normed vector spaces are metric spaces if we take the distance between vectors to be the norm of their difference, since all norms then satisfy the requirements of a metric. For all $v_1, v_2, v_3 \in V$ we have:

1. $\|v_1 - v_2\| = 0$ if and only if $v_1 = v_2$.
2. $\|v_1 - v_2\| = \|v_2 - v_1\|$.
3. $\|v_1 - v_3\| \leqslant \|v_1 - v_2\| + \|v_2 - v_3\|$.

In this thesis we work in $\mathbb{R}^d$ (and its subspaces) and, unless otherwise noted, norms in this thesis are always the Euclidean norm. From time to time we will also make use of the ℓ_∞ norm:

$$\|v\|_\infty := \sup\{|v_i|\}$$

Furthermore although in Euclidean space a vector is an element that has both direction and magnitude, the term vector is frequently overloaded to mean the point described by a vector; throughout this thesis we will use the terms point and vector interchangeably and the meaning intended will be clear from context.

Closely related to the Euclidean norm is the standard *inner product* or *dot product* on $\mathbb{R}^d$. This is the element-wise product of two vectors v, w defined by:

$$v^T w := \sum_{i=1}^{d} v_i w_i$$

Note that therefore $v^T w = w^T v$ and also that $v^T v = \sum_{i=1}^{d} v_i^2 = \|v\|^2$. The dot product has an equivalent definition in terms of the principal angle, $\theta \in [-\pi/2, \pi/2]$, between two vectors which is:

$$v^T w := \cos(\theta)\|v\|\|w\|$$

Rearranging we see that therefore:

$$\theta = \arccos\left(\frac{v^T w}{\|v\|\|w\|}\right)$$

and, in particular, v is orthogonal to w if and only if $v^T w = 0$ (where, by convention, we take the zero vector to be orthogonal to any other).

2.2.2 Linear Independence and Bases

If V is a vector space, $v_1, v_2, \ldots, v_n$ are vectors in V and $\alpha_1, \alpha_2, \ldots, \alpha_n$ are scalars, then the vector $v = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n$ is called a *linear combination* of the vectors v_i. If the linear equation $\beta_1 v_1 + \beta_2 v_2 + \ldots + \beta_n v_n = 0$ is satisfied by some set of β_i which are not all zero, then we say the vectors v_i are *linearly dependent* (or just *dependent*). If this equation is only satisfied when all of the β_i are zero, then we say the v_i are *linearly independent* (or just *independent*). Note that any subset of an independent set must itself be independent and furthermore, since $\{0\}$ is a dependent set, any set containing the zero vector is linearly dependent. It can be that a set of vectors is linearly dependent, but that any proper subset of it is independent (e.g. the set $\{(1,0)^T, (0,1)^T, (1,1)^T\}$ has this property).

It is easy to check that the set of all linear combinations of the set of vectors $\{v_1, v_2, \ldots, v_n\}$ is a vector space. We call this the vector space *spanned* by the v_i and write $\langle v_i \rangle_{i=1}^n$ for this space. The *dimension* of this space is the size of the largest linearly independent subset from $\{v_1, v_2, \ldots, v_n\}$. A set of linearly independent vectors is called a *basis* for the space it spans. An example of a (dependent) set which spans $\mathbb{R}^2$ is $\{(1,0)^T, (0,1)^T, (1,1)^T\}$, so $\langle (1,0)^T, (0,1)^T, (1,1)^T \rangle = \mathbb{R}^2$. Furthermore, any two vectors chosen from $\{(1,0)^T, (0,1)^T, (1,1)^T\}$ are independent and span $\mathbb{R}^2$ and hence form a basis for $\mathbb{R}^2$; we therefore see that bases are not unique.

We define the *standard basis* for $\mathbb{R}^d$ to be the set of vectors $\{e_i\}_{i=1}^d$, in which e_i is the vector with zero entries everywhere except for a 1 in the i-th entry. That is:

$$(e_i)_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

For example, the standard basis for $\mathbb{R}^2$ is $\{(1,0)^T, (0,1)^T\}$. When the vectors in a basis are all of unit norm and orthogonal to one another we say that they form an *orthonormal basis.* The standard basis is a particular example of an orthonormal basis.

Of particular interest to us will be *eigenvector* bases of $\mathbb{R}^d$. Recall that a non-zero vector x is called an *eigenvector* of the matrix A with scalar *eigenvalue* λ if $Ax = \lambda x$. Eigenvectors are not uniquely defined since any scaling of x is also an eigenvector of A with the same eigenvalue. By convention, we always take x to be an eigenvector that has unit norm so that x is then unique up to its sign.

If $A \in \mathcal{M}_{d \times d}$ has d distinct eigenvalues then the corresponding eigenvectors are linearly independent and hence form a basis for $\mathbb{R}^d$ (likewise if there are k distinct eigenvalues then the eigenvectors of A span a subspace of dimension at least k). In general the eigenvectors of a matrix are not pairwise orthogonal, and it is *not* true in general that the eigenvectors of a matrix span $\mathbb{R}^d$ if it does not have d distinct eigenvalues. However, for the special case of *symmetric* matrices $A = A^T$ it *is* true that one can always find a basis for $\mathbb{R}^d$ of eigenvectors of A even when there are repeated eigenvalues; moreover for symmetric matrices the eigenvectors are always pairwise orthogonal and so one can always find an *orthonormal* basis of eigenvectors (although if there are repeated eigenvalues this basis is not uniquely determined). We will use these facts often when we work with positive semi-definite matrices.

2.2.3 Positive (Semi-)definite Matrices

We begin this section by defining what we mean by a positive definite or semi-definite matrix.

Definition 1 (Positive (Semi-)Definite Matrices, Horn & Johnson (1985)) *Let A be a square symmetric matrix. We say that A is positive semi-definite (p.s.d) if, for every vector $v \in \mathbb{R}^d$, A has the property $v^T A v \geqslant 0$, and we write $A \succcurlyeq 0$. Similarly if B is p.s.d and $A - B \succcurlyeq 0$ we write $A \succcurlyeq B$. If A also satisfies $v^T A v > 0$, $\forall v \in \mathbb{R}^d$ then we say that A is positive definite (p.d), and we write $A \succ 0$.*

Note in particular that every p.s.d matrix A is symmetric: There are non-symmetric real matrices B for which $x^T B x \geqslant 0 \ \forall x \in \mathbb{R}^d$ holds[2] but such matrices are not p.s.d according to our definition.
We will make frequent use of several standard results and properties in this thesis. Apart from lemma 7 (for which we give a proof) these are more or less well-known, but for convenience we state them here for later reference.

Lemma 1 (Properties of symmetric matrices)
Let $A, B \in \mathcal{M}_{d \times d}$ be symmetric matrices, i.e. such that $A = A^T$ and $B = B^T$, then:

1. *$A + B$ and $A - B$ are symmetric.*
2. *All eigenvalues of A are real.*
3. *A is diagonalizable as $A = U \Lambda U^T$ where $\Lambda = diag(\lambda_i)$ is a diagonal matrix of the eigenvalues of A and U is an orthogonal matrix whose columns are unit eigenvectors of A, i.e. such that $UU^T = I = U^T U$.*
4. *AB is symmetric if and only if $AB = BA$. Equivalently, if and only if A and B are both diagonalized by the same orthogonal matrix of eigenvectors.*
5. *Corollary to 3 and 4: For all non-negative integers, n, A^n is symmetric and if A is invertible then A^{-n} is symmetric. $A^n = U \Lambda^n U^T$ where $\Lambda^n = diag(\lambda_i^n)$ and, if A is invertible, $A^{-n} = U \Lambda^{-n} U^T$ where $\Lambda^{-n} = diag(\lambda_i^{-n})$.*

[2] For example, it is easy to verify that for $x \in \mathbb{R}^2$ the matrix $\left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right)$ has this property.

6. $x^T Ay = y^T Ax$, $\forall\ x, y \in \mathbb{R}^d$ *if and only if* A *is symmetric.*

Lemma 2 (Properties of p.s.d matrices)
Let $A, B \in \mathcal{M}_{d\times d}$ *be positive semi-definite matrices,* $Q \in \mathcal{M}_{d\times d}$ *be a positive definite matrix and* $R \in \mathcal{M}_{k\times d}$ *be any matrix then the following hold:*

1. *The eigenvalues of* A *are all non-negative.*

2. *The eigenvalues of* Q *are all strictly positive.*

3. Q *is invertible and* $Q^{-1} \succ 0$.

4. $RAR^T \succcurlyeq 0$. *If rank*$(R) = k$ *(i.e.* R *has full row rank) then* $RQR^T \succ 0$ *and, in particular, if rank*$(R) = k$ *then* $RR^T = RIR^T \succ 0$.

5. A *has a unique p.s.d square root* $A^{\frac{1}{2}} = \left(A^T\right)^{\frac{1}{2}} = \left(A^{\frac{1}{2}}\right)^T$ *and* Q *has a unique p.d square root* $Q^{\frac{1}{2}} = \left(Q^T\right)^{\frac{1}{2}} = \left(Q^{\frac{1}{2}}\right)^T$. *(Theorem 7.2.6, pg. 406 Horn & Johnson, 1985)*

6. $A + B \succcurlyeq 0$ *and* $A^{\frac{1}{2}}BA^{\frac{1}{2}} \succcurlyeq 0$.

Lemma 3 (Eigenvalue and Trace identities)

1. *Let* $A \in \mathcal{M}_{d\times k}$ *and* $B \in \mathcal{M}_{k\times d}$ *then the non-zero eigenvalues of* AB *and* BA *are the same and have the same multiplicity. If* x *is an eigenvector of* AB *associated with a non-zero eigenvalue* λ*, then* $y = Bx$ *is an eigenvector of* BA *with the same non-zero eigenvalue. (Theorem A.6.2 Pg 468 Mardia et al., 1979).*

2. *Corollary to 1:* $Tr(AB) = Tr(BA)$.

3. *Let* $C \in \mathcal{M}_{d\times d}$ *be an invertible matrix, then* $\lambda_{\max}(C) = 1/\lambda_{\min}(C^{-1})$, $\lambda_{\min}(C) = 1/\lambda_{\max}(C^{-1})$.

4. *Corollary to 3: The condition number of* C, $\kappa(C) = \frac{\lambda_{\max}(C)}{\lambda_{\min}(C)} = \kappa(C^{-1})$.

Lemma 4 (Weyl's inequality. Horn & Johnson (1985) Theorem 4.3.1 Pg 181)
Let $A, B \in \mathcal{M}_{d\times d}$ *be symmetric matrices and arrange the eigenvalues* $\lambda_i(A)$, $\lambda_i(B)$ *and* $\lambda_i(A + B)$ *in increasing order, i.e.* $\lambda_{\max} = \lambda_d \geqslant \lambda_{d-1} \geqslant \ldots \geqslant \lambda_1 = \lambda_{\min}$. *Then, for each* $j \in \{1, 2, \ldots, d\}$ *we have:*

$$\lambda_j(A) + \lambda_{\min}(B) \leqslant \lambda_j(A + B) \leqslant \lambda_j(A) + \lambda_{\max}(B)$$

In particular, $\lambda_{\min}(A)+\lambda_{\min}(B) \leqslant \lambda_{\min}(A+B)$ *and* $\lambda_{\max}(A+B) \leqslant \lambda_{\max}(A)+\lambda_{\max}(B)$. *Equality holds when* A *and* B *have the same eigenvectors and eigenvalue ordering.*

Lemma 5 (Rayleigh quotient. Horn & Johnson (1985), Theorem 4.2.2 Pg 176) *If Q is a real symmetric matrix then, for any non-zero vector v, its eigenvalues λ satisfy:*

$$\lambda_{\min}(Q) \leqslant \frac{v^T Q v}{v^T v} \leqslant \lambda_{\max}(Q)$$

and, in particular:

$$\lambda_{\min}(Q) = \min_{v \neq 0} \tfrac{v^T Q v}{v^T v} = \min_{v^T v = 1} v^T Q v \text{ and} \tag{2.2.1}$$

$$\lambda_{\max}(Q) = \max_{v \neq 0} \tfrac{v^T Q v}{v^T v} = \max_{v^T v = 1} v^T Q v \tag{2.2.2}$$

Lemma 6 (Poincaré Inequality. Horn & Johnson (1985), Cor 4.3.16 Pg 190) *Let A be a symmetric matrix $A \in \mathcal{M}_{d \times d}$, let k be an integer, $1 \leqslant k \leqslant d$, and let $r_1, \ldots, r_k \in \mathbb{R}^d$ be k orthonormal vectors. Let $R \in \mathcal{M}_{k \times d}$ be the matrix with r_i its i-th row, and let $T = R^T A R \in \mathcal{M}_{k \times k}$ (in our setting, R is instantiated as an orthonormalized random projection matrix). Arrange the eigenvalues λ_i of A and T in increasing magnitude, then:*

$$\lambda_i(A) \leqslant \lambda_i(T) \leqslant \lambda_{i+d-k}(A), \quad i \in \{1, \ldots, k\}$$

and, in particular:

$$\lambda_{\min}(A) \leqslant \lambda_{\min}(T) \text{ and } \lambda_{\max}(T) \leqslant \lambda_{\max}(A)$$

Lemma 7 (Corollary to lemmas 5 and 6. Durrant & Kabán (2010b)) *Let Q be positive definite, such that $\lambda_{\min}(Q) > 0$ and so Q is invertible. Let $u = Rv$, $v \in \mathbb{R}^d$, $u \neq 0 \in \mathbb{R}^k$, with R any $k \times d$ matrix with full row rank and orthonormal rows. Then:*

$$u^T \left[R Q R^T \right]^{-1} u \geqslant \lambda_{\min}(Q^{-1}) u^T u > 0$$

Proof: We use the eigenvalue identity $\lambda_{\min}(Q^{-1}) = 1/\lambda_{\max}(Q)$. Combining this identity with lemma 5 and lemma 6 we have:

$$\lambda_{\min}([RQR^T]^{-1}) = 1/\lambda_{\max}(RQR^T)$$

Since RQR^T is positive definite. Then by positive definiteness and lemma 6 it follows that:

$$0 < \lambda_{\max}(RQR^T) \leqslant \lambda_{\max}(Q) \tag{2.2.3}$$

$$\iff 1/\lambda_{\max}(RQR^T) \geqslant 1/\lambda_{\max}(Q) > 0 \tag{2.2.4}$$

$$\iff \lambda_{\min}([RQR^T]^{-1}) \geqslant \lambda_{\min}(Q^{-1}) > 0 \tag{2.2.5}$$

And so by lemma 5:

$$u^T \left[RQR^T \right]^{-1} u \geqslant \lambda_{\min}([RQR^T]^{-1}) u^T u \tag{2.2.6}$$

$$\geqslant \lambda_{\min}(Q^{-1}) u^T u \tag{2.2.7}$$

$$= u^T u / \lambda_{\max}(Q) > 0 \tag{2.2.8}$$

Lemma 8 (Kantorovich Inequality. Horn & Johnson (1985), Theorem 7.4.41 Pg 444) *Let Q be a positive definite matrix $Q \in \mathcal{M}_{d \times d}$ with eigenvalues $0 < \lambda_{\min} \leqslant \ldots \leqslant \lambda_{\max}$. Then, for all $v \in \mathbb{R}^d$:*

$$\frac{(v^T v)^2}{(v^T Q v)(v^T Q^{-1} v)} \geqslant \frac{4 \cdot \lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}$$

With equality holding for some unit vector v.
This can be rewritten:

$$\frac{(v^T v)^2}{(v^T Q v)} \geqslant (v^T Q^{-1} v) \cdot \frac{4 \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)}{\left(1 + \frac{\lambda_{\max}}{\lambda_{\min}}\right)^2}$$

Lemma 9 (De Bruijn. De Bruijn (1956), Theorem 14.2) *Let Q be a positive definite matrix $Q \in \mathcal{M}_{d \times d}$, let k be an integer, $1 \leqslant k \leqslant d$, and let R be an arbitrary $k \times d$ matrix then:*

$$\lambda_{\min}(RQR^T) \geqslant \lambda_{\min}(Q) \cdot \lambda_{\min}(RR^T) \tag{2.2.9}$$

and:

$$\lambda_{\max}(RQR^T) \leqslant \lambda_{\max}(Q) \cdot \lambda_{\max}(RR^T) \tag{2.2.10}$$

Lemma 10 (Corollary to lemma 9) *Let $Q \in \mathcal{M}_{d \times d}$ be a positive definite matrix and $A \in \mathcal{M}_{d \times d}$ be a positive (semi-)definite matrix, then:*

$$\lambda_{\min}(QA) \geqslant \lambda_{\min}(Q) \cdot \lambda_{\min}(A) \tag{2.2.11}$$

and:

$$\lambda_{\max}(QA) \leqslant \lambda_{\max}(Q) \cdot \lambda_{\max}(A) \tag{2.2.12}$$

Proof: Take $A^{\frac{1}{2}}$ for R in lemma 9.

2.3 Random Matrix Theory

2.3.1 Random Projections

Random projection (RP) is a simple method of non-adaptive dimensionality reduction. Given d-dimensional data, which is to be compressed to a k-dimensional representation, the procedure is to generate a $k \times d$ matrix, R, with entries drawn i.i.d from a zero-mean Gaussian or subgaussian distribution (Achlioptas, 2003; Arriaga & Vempala, 1999; Dasgupta & Gupta, 2002), and then left multiply the data with R. Note that, with respect to the Gaussian measure, the random matrix R almost surely (or for the sub-Gaussians given in (Achlioptas, 2003), with high probability) has rank k.

Theoretical treatments of RP frequently assume that the rows of R have been orthonormalized, but in practice if the original data dimensionality d is very high this may not be necessary (Bingham & Mannila, 2001; Dasgupta, 2000a; Fradkin & Madigan, 2003a) as the rows of R, treated as random vectors, with high probability will have nearly identical norms and be approximately orthogonal to each other. These facts are folklore in the data mining community, but we have not seen a formal proof of this very general phenomenon. For completeness we will address this point now, in the following lemma:

Lemma 11 (**Durrant & Kabán** (**2011**))
Let s and t be vectors in $\mathbb{R}^d$ with their components $s_i, t_i \overset{i.i.d}{\sim} \mathcal{D}$, a non-degenerate zero-mean distribution i.e. with finite non-zero variance $0 < \sigma^2 < \infty$. Let $\|\cdot\|$ denote the Euclidean norm of its argument and $\langle s, t\rangle$ denote the inner product of s and t. Then:

$$Pr_{s,t}\left\{\lim_{d\to\infty}\left\langle \frac{s}{\|s\|}, \frac{t}{\|t\|}\right\rangle = 0\right\} = 1 \tag{2.3.1}$$

and

$$Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\|s\|}{\|t\|} = 1\right\} = 1 \tag{2.3.2}$$

that is, as $d \to \infty$, s becomes orthogonal to t almost surely and the norms $\|s\|, \|t\|$ become the same almost surely.
Proof: First, we show that $\|s\|/\sqrt{d}$ converges almost surely to σ. We start by noting $E[s_i^2] = Var[s_i] = \sigma^2$. Then, since s_i^2, d and σ^2 are all positive and the s_i^2 are i.i.d, we have:

$$Pr_s\left\{\lim_{d\to\infty}\frac{\sum_{i=1}^d s_i^2}{d} = \sigma^2\right\} = Pr_s\left\{\lim_{d\to\infty}\sqrt{\frac{\sum_{i=1}^d s_i^2}{d}} = \sigma\right\} \tag{2.3.3}$$

and this probability is equal to 1 by applying the strong law of large numbers for i.i.d random variables (e.g. (Rosenthal, 2006) Thm. 5.4.4 Pg 62) to the LHS of (2.3.3). A similar argument shows that $\|t\|/\sqrt{d}$ also converges almost surely to σ.
Next, since s_i and t_i are independent and zero-mean we have $E[s_i t_i] = 0$ for all i, so applying the strong law of large numbers once more we see that:

$$Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\langle s, t\rangle}{d} = 0\right\} = Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\sum_{i=1}^d s_i t_i}{d} = 0\right\} = 1 \tag{2.3.4}$$

We may rewrite (2.3.4) as:

$$Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\sum_{i=1}^d s_i t_i}{d} = 0\right\} = Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\langle s, t\rangle}{\|s\|\|t\|}\cdot\frac{\|s\|\|t\|}{d} = 0\right\} \tag{2.3.5}$$

we will prove (2.3.1) by showing that $\frac{\|s\|\|t\|}{d} \xrightarrow{a.s.} \sigma^2 \in (0, \infty)$ and hence conclude that $\frac{\langle s,t\rangle}{\|s\|\|t\|} \xrightarrow{a.s.} 0$.
Utilising the independence of s and t we see, via the strong law and by applying the product rule for limits of continuous functions to (2.3.3), that:

$$Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\|s\|\|t\|}{d} = \sigma^2\right\} = 1 \tag{2.3.6}$$

Indeed, negating and combining (2.3.4) and (2.3.6), via the union bound we observe:

$$Pr_{s,t}\left\{\left(\lim_{d\to\infty}\frac{\langle s,t\rangle}{d}\neq 0\right)\vee\left(\lim_{d\to\infty}\frac{\|s\|\|t\|}{d}\neq\sigma^2\right)\right\}$$
$$\leqslant Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\langle s,t\rangle}{d}\neq 0\right\}+Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\|s\|\|t\|}{d}\neq\sigma^2\right\}=0+0=0 \tag{2.3.7}$$

and so:

$$Pr_{s,t}\left\{\left(\lim_{d\to\infty}\frac{\langle s,t\rangle}{d}=0\right)\wedge\left(\lim_{d\to\infty}\frac{\|s\|\|t\|}{d}=\sigma^2\right)\right\}$$
$$\geqslant 1-\left(Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\langle s,t\rangle}{d}\neq 0\right\}+Pr_{s,t}\lim_{d\to\infty}\left\{\frac{\|s\|\|t\|}{d}\neq\sigma^2\right\}\right)=1 \tag{2.3.8}$$

Finally, since $0<\sigma^2<\infty$ we conclude that:

$$Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\langle s,t\rangle}{\|s\|\|t\|}\cdot\sigma^2=0\right\}=Pr_{s,t}\left\{\lim_{d\to\infty}\left\langle\frac{s}{\|s\|},\frac{t}{\|t\|}\right\rangle=0\right\}=1 \tag{2.3.9}$$

as required.
To prove the almost sure convergence of norms (2.3.2) we again use equation (2.3.3) and the fact that $\|s\|/\sqrt{d}$ and $\|t\|/\sqrt{d}$ converge almost surely to σ. Then applying the quotient rule for limits, we have (since $\sigma\neq 0$):

$$Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\left(\frac{\sum_{i=1}^d s_i^2}{d}\right)^{1/2}}{\left(\frac{\sum_{i=1}^d t_i^2}{d}\right)^{1/2}}=1\right\}=Pr_{s,t}\left\{\lim_{d\to\infty}\frac{\|s\|}{\|t\|}=1\right\}=1 \tag{2.3.10}$$

as required. □

In what follows we will frequently be interested in the spectral properties of random projection matrices. In particular in the extreme singular values of $k\times d$ rectangular random matrices R, with $r_{ij}\sim\mathcal{N}(0,\sigma^2)$ and $k\ll d$, and the extreme eigenvalues of the related p.s.d matrices RR^T and R^TR. Recall that if $s_j(A)$ is the j-th non-zero singular value of the matrix $A\in\mathcal{M}_{k\times d}$ then $\lambda_j(AA^T)$ is the j-th non-zero eigenvalue of the p.s.d square matrix AA^T and $\lambda_j(AA^T)=(s_j(A))^2=\lambda_j(A^TA)$. The following result bounds the extreme (non-zero) singular values of an instance of the random projection matrix R with high probability.

Theorem 1 (Singular Values of Gaussian Random Matrices. Vershynin (2012)) *Let R be a $k\times d$ matrix with i.i.d $\mathcal{N}(0,1)$ entries. Then for all $\epsilon>0$ with probability at least $1-2\exp(-\epsilon^2/2)$ we have:*

$$\sqrt{d}-\sqrt{k}-\epsilon\leqslant s_{\min}(R)\leqslant s_{\max}(R)\leqslant\sqrt{d}+\sqrt{k}+\epsilon \tag{2.3.11}$$

In particular, the extreme singular values of $(1/\sqrt{d})R$ are approximately $1\pm\sqrt{k/d}$ with high probability, so that when d is large (compared to k and ϵ) these matrices act like approximate isometries on their range. Closely related (and earlier) results, which are the basis of earlier theory for randomly-projected learning, are the following rather surprising facts:

Theorem 2 (Johnson-Lindenstrauss Lemma (JLL)) *Let $\epsilon \in (0,1)$. Let $N, k \in \mathbb{N}$ such that $k \geqslant C\epsilon^{-2}\log N$, for a large enough absolute constant C. Let $V \subseteq \mathbb{R}^d$ be a set of N points. Then there exists a linear mapping $R : \mathbb{R}^d \to \mathbb{R}^k$, such that for all $u, v \in V$:*

$$(1-\epsilon)\|u-v\|^2 \leqslant \|Ru - Rv\|^2 \leqslant (1+\epsilon)\|u-v\|^2$$

There are several elementary proofs of the Johnson-Lindenstrauss lemma e.g. (Achlioptas, 2003; Dasgupta & Gupta, 2002; Matoušek, 2008) in which it is demonstrated that random projection is, with high probability, a suitable choice for the linear mapping R. The key step in each of these proofs is to show that, with high probability, the norm of a randomly projected unit vector is close to its expected value. The proof of the existence of a suitable linear mapping R then follows by applying the probabilistic method to the following obtained 'randomized' version of the JLL, and taking $\delta = 1/N$.

Theorem 3 (Randomized Johnson-Lindenstrauss Lemma (JLL)) *Let $\epsilon \in (0,1)$. Let $k \in \mathbb{N}$ such that $k \geqslant C\epsilon^{-2}\log\delta^{-1}$, for a large enough absolute constant C. Then there exists a* random *linear mapping $P : \mathbb{R}^d \to \mathbb{R}^k$, such that for any unit vector $x \in \mathbb{R}^d$:*

$$Pr_P\left\{(1-\epsilon) \leqslant \|Px\|^2 \leqslant (1+\epsilon)\right\} \geqslant 1-\delta$$

One suitable family of choices for P is the family of random matrices with zero-mean i.i.d subgaussian entries. Note that by linearity of norms, theorem 3 implies that the random matrix P approximately preserves norms for *any* fixed vector $z \in \mathbb{R}^d$ with high probability, i.e. by taking $x := z/\|z\|$ to be the unit vector in theorem 3.

2.4 Other Useful Inequalities

Lemma 12 (Markov's Inequality)
Let X be any (scalar) random variable and let $\epsilon > 0$. Then:

$$Pr\{|X| \geqslant \epsilon\} \leqslant \frac{E(|X|)}{\epsilon}$$

The following proof is folklore. Let E be an event, and let $\mathbf{1}_E$ be the indicator function which returns 1 if E occurs and 0 otherwise. Then:

$$\begin{aligned}
\epsilon\mathbf{1}_{|X|\geqslant\epsilon} &\leqslant |X| \\
\mathrm{E}[\epsilon\mathbf{1}_{|X|\geqslant\epsilon}] &\leqslant \mathrm{E}[|X|] \\
\epsilon\mathrm{E}[\mathbf{1}_{|X|\geqslant\epsilon}] &\leqslant \mathrm{E}[|X|] \\
\epsilon\mathrm{Pr}\{|X| \geqslant \epsilon\} &\leqslant \mathrm{E}[|X|]
\end{aligned}$$

There is also a matrix variant of Markov's Inequality:

Lemma 13 (Matrix Markov Inequality)
Let X be a random positive semi-definite matrix and A a fixed positive definite matrix. Then:

$$Pr\{X \npreceq A\} \leqslant Tr\left(E[X]A^{-1}\right)$$

The following proof is very similar to the one for the scalar case given above, and is due to Recht (2011). First note that if $X \npreceq A$, then $A^{-1/2}XA^{-1/2} \npreceq I$, and hence $\lambda_{\max}\left(A^{-1/2}XA^{-1/2}\right) > 1$. Then $\mathbf{1}_{X \npreceq A} \leqslant \text{Tr}\left(A^{-1/2}XA^{-1/2}\right)$ as the right hand side is always nonnegative, and, if the left hand side equals 1, the trace of the right hand side must exceed the largest eigenvalue of the right hand side which is greater than 1. Thus we have:

$$\Pr\{X \npreceq A\} = \mathrm{E}\left[\mathbf{1}_{X \npreceq A}\right] \leqslant \mathrm{E}\left[\text{Tr}\left(A^{-1/2}XA^{-1/2}\right)\right] = \text{Tr}\left(\mathrm{E}[X]A^{-1}\right)$$

where the last equality follows from the linearity and cyclic properties of the trace, lemma 3.

Lemma 14 (Jensen's Inequality (e.g. Anthony & Bartlett (1999), Sec A1.1 Pg 358))
Let V be a vector space and let $f : V \to \mathbb{R}$ be a convex function. *That is, f satisfies for all $x_1, x_2 \in V$ and all $\alpha \in (0, 1)$:*

$$f(\alpha x_1 + (1 - \alpha)x_2) \leqslant \alpha f(x_1) + (1 - \alpha)f(x_2)$$

Then if X is a random variable taking values on V:

$$E[f(X)] \geqslant f(E[X])$$

If $g : V \to \mathbb{R}$ is a concave function, *that is for all $x_1, x_2 \in V$ and all $\alpha \in (0, 1)$:*

$$g(\alpha x_1 + (1 - \alpha)x_2) \geqslant \alpha g(x_1) + (1 - \alpha)g(x_2)$$

and X is a random variable taking values on V, then the sense of the inequality is reversed:

$$E[g(X)] \leqslant g(E[X])$$

Lemma 15 (McDiarmid's Inequality. McDiarmid (1989))
Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A.

1. *Let $f : A^n \to \mathbb{R}$ be a real-valued function and assume there are $f_i : A^{n-1} \to \mathbb{R}$ functions such that, for all $1 \leqslant i \leqslant n$ the bounded differences condition:*

$$\sup_{x_1 \ldots x_n} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)| \leqslant c_i$$

holds almost surely. Then, $\forall t > 0$ we have:

$$Pr\{|f(X) - E[f(X)]| \geqslant t\} \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right) \tag{2.4.1}$$

2. *Let $g : A^n \to \mathbb{R}$ be a real valued function. Assume that, for all $1 \leqslant i \leqslant n$, the bounded differences condition:*

$$\sup_{x_1 \ldots x_n, x_i'} |g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leqslant c_i$$

holds almost surely. That is, if we exchange one instance of the random variable X_i for any other instance of the same random variable, then the difference in the evaluation of g is almost surely bounded. Then, $\forall t > 0$ we have:

$$Pr\{|g(X) - E[g(X)]| \geqslant t\} \leqslant 2 \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right) \tag{2.4.2}$$

3
State of the Art

Summary This chapter reviews the current literature, both empirical and theoretical, on random projections for dimensionality reduction for learning. We highlight some gaps in the current knowledge, and outline the contribution that this thesis makes.

Random projection is a recent and very promising non-adaptive dimensionality reduction procedure, for which a diverse collection of motivations and wide range of applications can be found in the literature. These different motivations for the use of random projections include:

- To trade some accuracy in order to reduce computational expense and/or storage overhead (e.g. KNN (Indyk & Naor, 2007), low-rank matrix approximation (Recht, 2011)).
- To bypass the collection of lots of data then throwing away most of it at preprocessing (compressed sensing (Donoho, 2006), compressed imaging (Mu et al., 2011)).
- To create a new theory of cognitive learning (RP perceptron (Arriaga & Vempala, 2006)).
- To replace a heuristic optimizer with a provably correct algorithm with performance guarantees (mixture learning (Dasgupta, 1999)).
- To obscure identifiable properties of data to allow third-party data processing (privacy-preserving data-mining (Liu & Liu, 2009; Zhou et al., 2009)).
- To improve the speed and accuracy of local search algorithms for identifying recurring substrings in DNA sequences ('motif' detection (Buhler & Tompa, 2002)).

while the growing list of applications includes:

- Dimensionality Reduction. e.g. (Bingham & Mannila, 2001)
- Classification. e.g. (Blum, 2006; Boyali & Kavakli, 2012; Calderbank et al., 2009; Chen et al., 2011; Durrant & Kabán, 2012b; Fradkin & Madigan, 2003a; Goel et al., 2005; Pillai et al., 2011; Rahimi & Recht, 2008a; Schclar & Rokach, 2009; Wright et al., 2009)
- Regression. e.g. (Boutsidis & Drineas, 2009; Hegde et al., 2007; Maillard & Munos, 2009; Zhou et al., 2009)
- Clustering and Density estimation. e.g. (Ailon & Chazelle, 2006; Avogadri & Valentini, 2009; Dasgupta, 1999; Fern & Brodley, 2003; Indyk & Motwani, 1998; Kalai et al., 2012)
- Other related applications: structure-adaptive kd-trees (Dasgupta & Freund, 2008), low-rank matrix approximation (Recht, 2011; Sarlos, 2006), sparse signal reconstruction (compressed sensing, compressed imaging) (Candes & Tao, 2006; Donoho, 2006; Mu et al., 2011), data stream computations (Alon et al., 1996), search (Buhler & Tompa, 2002).

In this thesis we study the effect of random projection on classification, and our motivation is to mitigate two common aspects of the dimensionality curse. More precisely: When data are high-dimensional and plentiful, we can view random projection as a form of lossy compression which allows us to reduce the algorithmic time and space complexity of learning a classifier and carrying out classification. In this setting we want to quantify the expected cost, in terms of classification accuracy, of working in the randomly projected domain. On the other hand when data are high-dimensional and scarce, we can view random projection as a regularization scheme which transforms an ill-posed parameter estimation problem into a well-posed one. A side-effect of this approach will be that it also allows reductions in algorithmic time and space complexity -vs- working in the data space. In this setting we want to understand this regularization effect and quantify the likely improvement, in terms of classification accuracy, from working in the randomly projected domain.

3.1 Summary of Existing Work

Before briefly summarizing the state-of-the-art, it is useful to highlight two theoretical results concerning randomly projected data. The first is the randomized version of the Johnson-Lindenstrauss Lemma (JLL) which we gave in Chapter 2 as theorem 3; namely that with high probability random projection approximately preserves the distances between a point set of N vectors, provided that the projection dimensionality k is taken to be at least $k \in \mathcal{O}(\epsilon^{-2} \log N)$. The second is the key result from the literature on Compressed Sensing which states that for an s-sparse vector $x \in \mathbb{R}^d$, i.e. x is d-dimensional but admits a sparse representation in some linear basis such that it has no more than $s \ll d$ non-zero entries, then the original vector x can be perfectly reconstructed from a k-dimensional representation of x, Rx, provided that R satisfies the so-called 'restricted isometry property' (RIP). One can show that, with high probability, if R is a (dense) random projection matrix then it satisfies the RIP provided that the projection dimensionality k is taken to be at least $k \in \mathcal{O}(s \log d)$. Armed with these results, we are ready to make a brief tour of the existing literature.
Empirical results on random projections in the literature were noted in (Bingham & Mannila, 2001) to be sparse and, within the machine learning and data-mining literature, that is still the case. Motivated by theoretical findings in the domain of compressed sensing (CS) the last few years have, however, seen a rapid growth in the number of experimental papers adopting random projection as a preprocessing step for classification of *sparse* data, especially in associated fields such as computer vision.
Within the machine learning and data-mining literature, experimental papers using random projection are typically method-specific and mainly focus on methods where the uniform approximate geometry preservation properties of random projection (via the JLL) are the key theoretical property motivating the use of RP. These include unsupervised methods such as density estimation (Dasgupta, 2000a), clustering (Kaski, 1998), clustering ensembles (Fern & Brodley, 2003), and SOM (Kurimo & IDIAP, 2000), as well as classification using decision trees, 1-NN, 5-NN, and linear SVM (Fradkin & Madigan, 2003a), AdaBoost (Paul et al., 2009), ensembles of IB1 (Aha et al., 1991) classifiers (a variant of 1-NN) (Schclar & Rokach, 2009), and also regression on

sparse data (Fard et al., 2012) - the latter motivated by CS-type results. Within the vision community, experimental papers are also method-specific and focus on applications using image data, which admit a sparse representation. Here signal reconstruction guarantees from compressed sensing are the key theoretical idea motivating the use of RP; in particular the central intuition applied in this setting is that when data are sparse then with high probability no information is lost by randomly projecting them, and so with the same probability there should be no degradation in the performance of a classifier or regressor following preprocessing by random projection. Examples of papers in the vision literature making empirical studies of classification of randomly projected data include: Target recognition and anomaly detection using linear discriminants (Chen et al., 2011), gesture recognition using a 'closest match' classifier (Boyali & Kavakli, 2012), face recognition using a 1NN variant employing the Hamming metric (Wright et al., 2009) and iris recognition using the same classifier (Pillai et al., 2011), face recognition using 'closest match', a majority-voting ensemble of 5 'closest match' classifiers, and a 'scoring' ensemble that weights the ensemble votes according to the number of training examples (Goel et al., 2005) - the last of these papers gives the JLL as motivation, the others are motivated by CS results.

For the experimental papers above, with the sole exception of (Dasgupta, 2000a) where the theory supporting the approach appears separately in (Dasgupta, 1999; 2000b), although theory clearly motivates the authors' use of random projections no serious attempt at theoretical analysis of the combination of random projection with the learning approaches employed is made. The experiments are carried out using different methods and on different data sets all with different characteristics and therefore it is hard to draw firm conclusions about which problem-specific properties are affected by random projection and which properties (if any) indicate that random projection will or will not work well. On the other hand, a common and particularly striking property of the experimental papers on learning a classifier from randomly-projected data is that, even when the projection dimension k is taken to be considerably smaller than one would expect for geometry preservation guarantees under the randomized version of the JLL, good classification performance can still be achieved. This mirrors the similar findings in (Bingham & Mannila, 2001) where the authors empirically verified the randomized JLL by measuring the amount of distortion random projection actually gives rise to in practice. On image data they found empirically that there was a low level of distortion on a sample of 100 interpoint distances even when taking $k = 50 \ll \epsilon^{-2} \log N$.[1]

Theoretical results concerning random projections applied to learning are somewhat more common. However, although there is a sizeable corpus on compressed (or compressive) sensing (CS) in the signal processing literature stemming from the work of (Candes & Tao, 2006; Donoho, 2006), there is not a great deal of theory on applying CS results to learning. The prototype for the majority of the theoretical papers in the machine learning literature is the seminal work of Arriaga and Vempala on RP perceptron (Arriaga & Vempala, 2006) where high probability guarantees on uniform geometry preservation via the JLL are the main tool employed. Contemporaneous work

[1] Rather surprisingly they also found that, to keep the distortion in interpoint distances at a comparable level, when applying PCA to the same data required retaining the first 600 principal eigenvectors!

by (Indyk & Motwani, 1998) gave the first algorithm for approximate nearest neighbour classification which was polynomial in both its time and space requirements. Around the same time (Garg et al., 2002) considered the use of random projections as a tool to develop PAC guarantees for non-projected classifiers and, more recently, the effect of RP on regression was considered in (Maillard & Munos, 2009). Semi-supervised learning of large margin classifiers is considered in (Balcan et al., 2006) where random projection is treated as an alternative to the 'kernel trick'. There is also theory for the unsupervised learning of Gaussian mixtures, where the first algorithm for learning a separated Gaussian mixture with high probability guarantees (Dasgupta, 1999) used random projections, as does the more recent algorithm for learning a non-separated mixture proposed by (Kalai et al., 2012) (which is, at present, a theoretical construct). The unsupervised learning of low-dimensional manifolds was also analyzed in (Baraniuk & Wakin, 2009). Very recently work by (Shi et al., 2012) considered margin preservation under random projection, while unpublished manuscripts by (Paul et al., 2012) and (Zhang et al., 2012) consider the question of how close the optimal linear classifier in the randomly-projected space is to the optimum in the dataspace and speeding up SVM learning using random projection respectively. For sparse data, where the guarantees derive from CS-type theory, exemplars are learning an SVM from RP sparse data (Calderbank et al., 2009) and learning a privacy-preserving regressor from RP sparse data (Zhou et al., 2009). A more extensive list of examples of empirical and theoretical papers on random projection and its application to learning (and related problems) can be found in the references to the slides of my ECML-PKDD 2012 tutorial 'Random Projections for Machine Learning and Data Mining: Theory and Applications'(Durrant & Kabán, 2012a). These are available on the internet at `https://sites.google.com/site/rpforml/`.

Finally we note, with an eye on chapter 8 of this thesis, that kernel classifiers have a similar form to the randomly-projected classifier we consider here; the key difference being that in kernel classifiers the role of the random projection is filled instead by orthogonal projection on to the span of the (feature mapped) training examples. We do not survey the very substantial literature on kernel learning here, rather we note that in the case of Kernel Fisher Discriminant (KFLD) - the setting we will later analyze - there are to the best of our knowledge only two existing results considering the generalization error of this classifier. The earliest attempts are in (Mika, 2002) which approaches the analysis of KFLD from the starting point of the KFLD objective function and its algorithmic solution as an eigenproblem, and try to quantify the quality of the eigenvector estimates. Unfortunately these still leave unanswered the question of the generalization performance of KFLD. A more informative bound appears in (Diethe et al., 2009) but for a sparse variant of KFLD. Moreover they require that the induced distribution of the feature-mapped data has bounded support and their bounds are somewhat loose.

3.2 Inadequacies in previous work and our Contribution

There are several deficiencies in the existing literature on learning from randomly-projected data. In particular, there is a significant disconnection between the theoretical and empirical work and, furthermore, the random projections are considered in isolation from the classification setting in which the randomly-projected data will be employed.

The empirical work is essentially ad hoc: Practitioners show that approach A works on dataset B, and conclude that therefore approach A is a good one. In reality though, the results are limited to the data and evaluation criteria used, and these are all different. There is little or no attempt made in the experimental papers to understand the results in the context of existing (or new) theoretical knowledge, which could illuminate the findings, and so these papers do not reveal why performance is good or when we can expect it to remain good. We need theory to understand what is going on, and to come up with improved algorithms that do better. On the other hand, the theory developed to date does not answer all of the questions that one might like to see answered. There is already considerable theory giving guarantees for the case of learning from finite training data outside of randomly-projected domains, and this has enhanced our understanding of statistical learning in a variety of settings, but there is little theoretical work to date incorporating the effect of dimensionality in these bounds. Given the increasing prevalence of high-dimensional data and the widespread application of dimensionality reduction techniques, this appears to be a subject that could be of particular interest.

We can broadly summarise previous theoretical work on learning in randomly-projected domains as belonging to one of two camps: In the first camp are guarantees based on geometry preservation via the JLL. These proceed by showing that with high probability all norms and dot products that appear as quantities in the learning algorithm in the randomly-projected space take values close to the corresponding quantities in the data space. The whole of the effect of the random projection is disposed of in this way, and then one simply works with these distorted quantities in the data space to derive guarantees. The major drawback to this uniform geometry preservation approach is that one then obtains a bound that, contrary to experience and expectation, grows looser as the number of training observations increases.

In the second camp are guarantees for sparse data based on the RIP of CS. These are similar in flavour to the JLL-based approaches but, by assuming sparse data and working with the RIP property of random projection matrices, one removes the unwanted logarithmic dependence on the number of observations and therefore the unnatural behaviour of bounds employing this approach, but at the cost of a linear dependence on the number of non-zero entries of the sparse data instead. However, this is the same dependence as is required for perfect reconstruction of randomly projected data and classification should be a far simpler task than that. For classification it is reasonable to believe that either stronger results should be achievable for sparse data, or that the sparsity condition on the data can be relaxed without substantially weakening the guarantee, or perhaps both.

This thesis attempts to advance theoretical understanding of issues related to learning, and specifically we aim to answer some questions that may be relevant to practical applications.

4
Fisher's Linear Discriminant

Summary In this chapter, we consider Fisher's Linear Discriminant (FLD) - a popular choice of linear classifier.

We first provide some preliminary background, in which we introduce the classification problem and describe the decision rule implemented by FLD. We show that the FLD decision rule for multiclass settings can be decomposed as a sequence of two-class classification problems, and therefore we focus mainly on two-class classification problems in the remainder of this thesis. We develop upper bounds on the classification performance of FLD working in the data space of two types: Firstly, we give an upper bound on the FLD generalization error when the class-conditional distributions are subgaussian; Secondly, we provide elementary proofs of the results of Bickel & Levina (2004); Pattison & Gossink (1999) giving the exact generalization error of FLD conditional on a fixed training set when the query point class-conditionals are Gaussian with identical covariance.

These two data space bounds (and variants of similar form) will be used frequently in future chapters when we analyze randomly-projected FLD (RP-FLD), ensembles of RP-FLD classifiers, and kernel FLD.

Our main aim in deriving these results is to eventually quantify the effect of random projection on the performance of FLD classification, therefore here and in the subsequent chapter 5 we take the training set to be fixed. In the final two chapters 7 and 8 we consider the effect of random training sets of fixed size.

4.1 Preliminaries

4.1.1 The classification problem

In an m-class classification problem we observe N examples of labelled training data $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$ where $(x_i, y_i) \overset{i.i.d}{\sim} \mathcal{D}_{x,y}$ a (usually unknown) distribution. For a given class of functions $\mathcal{H}$, our goal is to learn from $\mathcal{T}_N$ the function $\hat{h} \in \mathcal{H}$ with the lowest possible *generalization error* in terms of some loss function $\mathcal{L}$. That is, find $\hat{h}$ such that $\mathcal{L}(\hat{h}(x_q), y_q) = \min_{h \in \mathcal{H}} \mathrm{E}_{x_q,y_q}[\mathcal{L}(h(x_q), y_q)]$, where $(x_q, y_q) \sim \mathcal{D}_{x,y}$ is a query point with unknown label y_q.
In the case we consider here, the class of functions $\mathcal{H}$ consists of instantiations of FLD learned from the training data, and we use the $(0,1)$-loss $\mathcal{L}_{(0,1)} : \{0,1\} \times \{0,1\} \to \{0,1\}$ as a measure of performance defined by:

$$\mathcal{L}_{(0,1)}(h(x_q), y_q) = \begin{cases} 0 & \text{if } h(x_q) = y_q \\ 1 & \text{otherwise.} \end{cases}$$

We seek to quantify the generalization error or expected $(0,1)$-loss of FLD which, for the $(0,1)$-loss, is equivalent to the probability that an arbitrarily drawn query point $(x_q, y_q) \sim \mathcal{D}_{x,y}$ is misclassified by the learned classifier. Our approach is therefore to upper bound:

$$\mathrm{E}_{x_q,y_q}[\mathcal{L}_{(0,1)}(\hat{h}(x_q), y_q)] = \mathrm{Pr}_{x_q,y_q}[\hat{h}(x_q) \neq y_q : (x_q, y_q) \sim \mathcal{D}_{x,y}]$$

Where $\hat{h}$ is the classifier decision rule learned from the training data. For concreteness and tractability, we do this for Fisher's Linear Discriminant (FLD) classifier, which we briefly review in the next Section 4.1.2. With our eye on classification in the randomly-projected domain, which is our main interest and the subject of later chapters, our motivation for choosing RP-FLD as our object of study was threefold: In the first place, FLD and closely-related variants are a popular and successful family of classification methods with applications in diverse areas including Medicine, Economics, Spectroscopy, and Face Recognition (Dudoit et al., 2002; Guo et al., 2007; Kim & Kittler, 2005; Lu et al., 2005; McLachlan, 2004; Pang et al., 2005; Wu et al., 1996), therefore results concerning RP-FLD are likely to be of interest. Secondly, discriminative linear classifiers working on randomly projected data have been studied before in (Arriaga & Vempala, 2006; Calderbank et al., 2009) (where the authors considered Perceptron and SVM respectively) and it is therefore interesting to consider a randomly-projected linear generative classifier – RP-FLD is such a classifier. Thirdly, for discriminative classifiers we found it hard to see how to avoid using the geometry-preserving properties of RP as the basis for any quantitative analysis, while on the other hand it seemed possible that a generative classifier might present opportunities for different proof techniques and therefore results with a quite different flavour from the existing ones. We shall see in the subsequent chapters 5, 7 and 8 that this is indeed the case.

4.1.2 Fisher's Linear Discriminant

FLD is a generative classifier that seeks to model, given training data $\mathcal{T}_N$, the optimal decision boundary between classes. Let Σ_0 and Σ_1 be the class-conditional covariance matrices and μ_0 and μ_1 be the class-conditional means of the data distribution $\mathcal{D}$. If $\Sigma = \Sigma_0 = \Sigma_1$ and μ_0 and μ_1 are known, then the optimal classifier is given by Bayes' rule (Bickel & Levina, 2004):

$$h(x_q) = \mathbf{1}\left\{\log \tfrac{f_1(x_q)}{f_0(x_q)} > 0\right\} \tag{4.1.1}$$

$$= \mathbf{1}\left\{(\mu_1 - \mu_0)^T \Sigma^{-1}\left(x_q - \tfrac{(\mu_0+\mu_1)}{2}\right) > 0\right\} \tag{4.1.2}$$

where $\mathbf{1}(P)$ is the indicator function that returns one if P is true and zero otherwise, and f_y is the Gaussian density $\mathcal{N}(\mu_y, \Sigma)$ with mean μ_y and covariance Σ, namely:

$$\left((2\pi)^{d/2} \det{(\Sigma)}^{1/2}\right)^{-1} \exp\left(-\frac{1}{2}(x-\mu_y)^T \Sigma^{-1}(x-\mu_y)\right)$$

To construct the FLD classifier from training data, one simply replaces the true parameters μ_y and Σ in equation (4.1.2) with estimates made from the data to obtain:

$$\hat{h}(x_q) = \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\left(x_q - \frac{(\hat{\mu}_0 + \hat{\mu}_1)}{2}\right) > 0\right\} \tag{4.1.3}$$

where $\hat{\Sigma}$ is a positive definite covariance matrix estimated from data or assigned a specific form due to prior knowledge or problem characteristics. An asymptotically unbiased choice for $\hat{\Sigma}$ is the maximum likelihood (ML) covariance estimate: $\hat{\Sigma} = \frac{1}{N}\sum_{y=1}^{0}\sum_{i=1}^{N_y}(x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T$ where the second summation is over the N_y training examples from class y. If the ML estimate is singular then either one can use a regularization scheme (for example, ridge regularization (Hastie et al., 2001) or shrinkage regularization (Friedman, 1989; Ledoit & Wolf, 2004)) or one can pseudo-invert to obtain a working classifier - both approaches are used in practice (Raudys & Duin, 1998). In the case of very high-dimensional data a simplified covariance model may instead be used for computational tractability, with common choices being a diagonal covariance estimate consisting of the feature sample variances: $\hat{\Sigma} = \text{diag}(\text{Var}(x_i))_{i=1}^{d}$ or the identity matrix. Other approaches which essentially interpolate between these diagonal schemes and the full ML covariance estimate have also been proposed (Ma et al., 2010).

In the following we shall generally assume equal class covariance matrices in the true data distribution, both for simplicity and to reduce notational overload. Although the FLD model constrains the covariance estimate to be shared across the classes, i.e. $\hat{\Sigma}_y = \hat{\Sigma}, \forall y$, different *true* class covariance matrices Σ_y can be introduced into the bounds we derive in a straightforward manner; we show in section 4.2, corollary 1, how this can be done. On the other hand allowing the *model* covariances to be different gives rise to a non-linear (quadratic) decision boundary – this class of classifiers is called the Quadratic Normal-based Discriminant Rule in McLachlan (2004) and Quadratic

Discriminant Analysis (QDA) in Friedman (1989); Lachenbruch & Goldstein (1979). In principle one would expect QDA to generally outperform FLD but it is known that, in practice, QDA rarely outperforms FLD unless the number of training observations for each class is large compared to the dimensionality and the class-conditional covariance matrices are very different (Friedman, 1989; Wu et al., 1996). Moreover empirical results in Friedman (1989); Wu et al. (1996) show that regularized variants of FLD can still outperform QDA even in those cases where QDA improves on the standard 'vanilla' FLD, while Lachenbruch & Goldstein (1979) shows that FLD is more robust to both systematic and random labelling errors in training data than QDA. In this thesis we therefore restrict our attention to FLD and we do not analyze the QDA model.
Finally we note that FLD can be extended from the two-class setting to the multi-class setting; if there are m class labels then the decision rule is given by:

$$\hat{h}(x_q) = j \iff j = \arg\max_i \{\Pr_y(y = i | x_q)\} \quad y, j, i \in \{0, 1, \ldots, m-1\}$$

We shall see in section 4.2.3 that, for multi-class FLD with this decision rule, upper bounds on the generalization error in the multi-class setting can be given by a sum of two-class generalization errors and so multi-class generalization error bounds can easily be derived from our two-class bounds.

4.2 Dataspace analysis of FLD

We use two approaches to quantify the generalization error in the dataspace:
In our first approach in Section 4.2.1 we derive an upper bound on the dataspace generalization error that holds for Gaussian and subgaussian classes. In our second approach in Section 4.2.2 we show that if the modelling assumptions of FLD are satisfied, that is we have a two-class classification problem where the classes are multivariate Gaussian with different means but equal covariance matrices, then we can give the exact generalization error of FLD. We note that since FLD is the Bayes' optimal classifier when its modelling assumptions hold, this exact error is therefore also a lower bound on the generalization error of FLD when the class-conditional distributions are not Gaussian and also for *any* two-class classifier when the assumption of Gaussianity holds.
We shall see that our upper bound obtained in Section 4.2.1 for subgaussians has the same form as the error obtained in Section 4.2.2 when the classes are Gaussian, and the cost of allowing for subgaussianity is a small multiplicative absolute constant. Specifically, we shall see that the bound for subgaussian classes we derive here is exactly twice a corresponding upper bound for Gaussian classes.

4.2.1 Upper bound on the generalization error of two-class FLD with subgaussian class conditionals

In this section we derive an upper bound on the generalization error of two-class FLD in the data space assuming subgaussian classes (in fact, we derive these bounds also under the assumption of Gaussian classes, but we highlight in the proofs how the relaxation to subgaussianity becomes possible). For this upper bound we require the condition $\alpha_y = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} - \hat{\mu}_y) > 0$ to hold; this condition arises from the optimization step in the proof of the bound, and amounts only to requiring that, $\forall y \in \{0,1\}$, μ_y and $\hat{\mu}_y$ lie on the same side of the classification hyperplane estimated by FLD in the dataspace.

Theorem 4 (Upper bound on generalization error of two-class FLD) *Let $x_q|y_q = y \sim \mathcal{N}(\mu_y, \Sigma)$ or a multivariate subgaussian with mean μ_y and covariance Σ. Let $\pi_y = Pr\{y_q = y\}$. Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Assume that we have sufficient data so that $\hat{\Sigma}$ is full-rank and $\alpha_y = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T\hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} - \hat{\mu}_y) > 0$ (i.e. μ_y and $\hat{\mu}_y$ lie on the same side of the decision hyperplane) $\forall y, \neg y \in \{0,1\}, y \neq \neg y$. Then the probability that x_q is misclassified is bounded above by:*

$$Pr_{x_q,y_q}[\hat{h}(x_q) \neq y_q] \leqslant \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T\hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_y)\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right) \tag{4.2.1}$$

with μ_y the mean of the class from which x_q was drawn, estimated class means $\hat{\mu}_0$ and $\hat{\mu}_1$, model covariance $\hat{\Sigma}$, and true class priors π_y.

Proof 1 (of Theorem 4)
We prove one term of the bound using standard techniques, the other term being proved similarly.
Without loss of generality let x_q have label $y_q = 0$. Then the probability that x_q is misclassified is given by $Pr_{x_q|y_q=0}[\hat{h}(x_q) \neq y_q|y_q = 0]$:

$$\begin{aligned} &= Pr_{x_q|y_q=0}\left[(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}\left(x_q - \tfrac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) > 0\right] \\ &= Pr_{x_q|y_q=0}\left[(\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}\left(x_q - \tfrac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) > 0\right] \end{aligned}$$

for all $\alpha_0 > 0$. Exponentiating both sides gives:

$$\begin{aligned} &= Pr_{x_q|y_q=0}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}\left(x_q - \tfrac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)\right) > 1\right] \\ &\leqslant E_{x_q|y_q=0}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}\left(x_q - \tfrac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)\right)\right] \end{aligned}$$

by Markov's inequality. Then, isolating terms in x_q we have:

$$\begin{aligned} &Pr_{x_q|y_q=0}[\hat{h}(x_q) \neq y_q|y_q = 0] \\ \leqslant\ & E_{x_q|y_q=0}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}x_q - \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1)\right)\right] \\ =\ & \exp\left(-\frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1)\right) E_{x_q|y_q=0}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}x_q\right)\right] \end{aligned}$$

This expectation is of the form of the moment generating function of a multivariate

Gaussian and so:

$$E_{x_q|y_q=0}\left[\exp\left((\hat{\mu}_1-\hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}x_q\right)\right] = \\ \exp\left(\frac{1}{2}(\hat{\mu}_1-\hat{\mu}_0)^T\alpha_0^2\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)+\mu_0^T\alpha_0\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\right) \quad (4.2.2)$$

where μ_0 is the true mean, and Σ is the true covariance matrix, of $\mathcal{D}_{x_q|y_q=0}$. Thus, we have the probability of misclassification is bounded above by the following:

$$\exp\left(-\frac{1}{2}(\hat{\mu}_1-\hat{\mu}_0)^T\alpha_0\hat{\Sigma}^{-1}(\hat{\mu}_0+\hat{\mu}_1)+\mu_0^T\alpha_0\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)+\frac{1}{2}(\hat{\mu}_1-\hat{\mu}_0)^T\alpha_0^2\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\right)$$

Now, since this holds for every $\alpha_0>0$ we may optimise the bound by choosing the best one. Since exponentiation is a monotonic increasing function, in order to minimize the bound it is sufficient to minimize its argument. Differentiating the argument w.r.t α_0 and equating to zero then yields:

$$\alpha_0=\frac{(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}{2(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)} \quad (4.2.3)$$

This is strictly positive as required, since the denominator is always positive (Σ is positive definite, then so is $\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}$), and the numerator is assumed to be positive as a precondition in the theorem. Substituting α_0 back into the bound then yields, after some algebra, the following:

$$Pr_{x_q|y_q=0}[\hat{h}(x_q)\neq 0|y_q=0]\leqslant\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)\right]^2}{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}\right)$$

The second term, for when $x_q\sim\mathcal{D}_{x_q|y_q=1}$, can be derived similarly and gives:

$$Pr_{x_q|y_q=1}[\hat{h}(x_q)\neq y_q|y_q=1]\leqslant\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_0-\hat{\mu}_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_1)\right]^2}{(\hat{\mu}_0-\hat{\mu}_1)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_0-\hat{\mu}_1)}\right)$$

Finally, putting these two terms together and applying the law of total probability we arrive at Theorem 4, i.e. that:

$$Pr_{x_q,y_q}[\hat{h}(x_q)\neq y_q]\leqslant\pi_0\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)\right]^2}{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}\right)+ \\ (1-\pi_0)\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_0-\hat{\mu}_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_1)\right]^2}{(\hat{\mu}_0-\hat{\mu}_1)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_0-\hat{\mu}_1)}\right)$$

Comment 1

We see from the proof above, that it is straightforward to introduce different true class-conditional covariance matrices into the bound in Theorem 4. In particular, when $x_q|y \sim \mathcal{N}(\mu_y, \Sigma_y)$ we see that Σ_y will appear in the bound via the step employing the class-conditional m.g.f, equation (4.2.2), and for the y-th class we may therefore replace Σ everywhere with Σ_y. In this case α_y then becomes:

$$\alpha_y := \frac{(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} - \hat{\mu}_y)}{2(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\Sigma_y \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}$$

which does not affect the constraint $\alpha_y > 0, \forall y$ at all since the sign of α_y is determined completely by its numerator. Therefore, under the same conditions as before, we have the following upper bound on the generalization error when $\Sigma_0 \neq \Sigma_1$:

Corollary 1 (Different Class-conditional Covariances)
Under the same conditions as theorem 4, except relaxing the requirement that $\Sigma_0 = \Sigma_1 = \Sigma$ (i.e. the class-conditional covariances need no longer be the same), the probability that x_q is misclassified is bounded above by:

$$Pr_{x_q,y_q}[\hat{h}(x_q) \neq y_q] \leqslant \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_y)\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\Sigma_y \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right) \tag{4.2.4}$$

Comment 2

We should confirm, of course, that the requirement that $\alpha_y > 0$ is a reasonable one. Because the denominator in (4.2.3) is always positive the condition $\alpha_y > 0$ holds when:

$$(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y) > 0$$

It can be seen that $\alpha_y > 0$ holds provided that for each class the true and estimated means are both on the same side of the decision hyperplane.

Comment 3

We note that, in equation (4.2.2) it is in fact sufficient to have inequality. Therefore our bound also holds when the true distributions of the data classes are such that they have a moment generating function no greater than that of the Gaussian. By Remark 5.6.2 of Vershynin (2012) any such distribution has the super-exponential tail decay characteristic of the Gaussian distribution (in fact, has all of the equivalent properties (1)-(3) of Lemma 5.5 in (Vershynin, 2012)) and such distributions are called subgaussian distributions.

4.2.2 Exact generalization error for two-class FLD with Gaussian class-conditionals

In this section we derive the exact generalization error of FLD when the classes have Gaussian distribution with identical covariance matrices.

Theorem 5 (Exact generalization error with Gaussian classes) *Let $\Sigma \in \mathcal{M}_{d\times d}$ be a full rank covariance matrix and let $x_q|y_q = y \sim \mathcal{N}(\mu_y, \Sigma)$. Let $\hat{\Sigma} \in \mathcal{M}_{d\times d}$ be a p.s.d covariance estimate and let $\hat{\Sigma}^{-1}$ be its inverse or pseudo-inverse. Then the exact generalization error of the FLD classifier (4.1.2) is given by:*

$$Pr_{x_q,y_q}[\hat{h}(x_q) \neq y_q] = \sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T\hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}}\right)$$

where Φ is the c.d.f of the standard Gaussian.

Proof 2 (of Theorem 5)
The proof of this theorem is similar in spirit to the ones given in Bickel & Levina (2004); Pattison & Gossink (1999). Without loss of generality let x_q have label 0. By assumption the classes have Gaussian distribution $\mathcal{N}(\mu_y, \Sigma)$ so then applying (4.1.2) the probability that x_q is misclassified by FLD is given by:

$$Pr_{x_q|y_q=0}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\} \tag{4.2.5}$$

Define $a^T := (\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}$ and observe that if $x_q \sim \mathcal{N}(\mu_0, \Sigma)$ then:

$$\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) \sim \mathcal{N}\left(\left(\mu_0 - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right), \Sigma\right)$$

and so:

$$a^T\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) \sim \mathcal{N}\left(a^T\left(\mu_0 - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right), a^T\Sigma a\right)$$

which is a univariate Gaussian. Therefore:

$$\frac{a^T\left(x_q - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) - a^T\left(\mu_0 - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)}{\sqrt{a^T\Sigma a}} \sim \mathcal{N}(0, 1)$$

Hence, for the query point x_q we have the probability (4.2.5) is given by:

$$\Phi\left(\frac{a^T\left(\mu_0 - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)}{\sqrt{a^T\Sigma a}}\right)$$
$$= \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_0)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}}\right)$$

where Φ is the c.d.f of the standard Gaussian.
A similar argument deals with the case when x_q belongs to class 1, and shows that if x_q belongs to class 1, then the probability of misclassification is given by:

$$\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_0 - \hat{\mu}_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_1)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}}\right)$$

Finally, applying the law of total probability: $Pr_{x_q,y_q}[\hat{h}(x_q) \neq y_q] = \sum_{y=0}^{1} Pr[x_q \sim \mathcal{N}(\mu_y, \Sigma)] \cdot Pr[\hat{h}(x_q) \neq y | x_q \sim \mathcal{N}(\mu_y, \Sigma)]$*, completes the proof.* □

Comment 4

We note that starting from Theorem 5 which holds for Gaussian classes, we can see how much the relaxation to subgaussian classes costs us (in terms of the tightness of our bound) by directly bounding the exact error of FLD derived there in a similar way. The exact error derived in Theorem 5 is:

$$\sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma_y \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}}\right)$$

where $\Phi(\cdot)$ is the Gaussian CDF. Then using equation (13.48) of Johnson et al. (1994) which bounds this quantity we get:

$$\Phi(-x) = 1 - \Phi(x) \leqslant 1 - \frac{1}{2}\left[1 + \sqrt{1 - e^{-x^2/2}}\right] \leqslant \frac{1}{2}\exp(-x^2/2) \tag{4.2.6}$$

The upper bound on the RHS follows from observing that $\sqrt{1 - e^{-x^2/2}} \geqslant 1 - e^{-x^2/2}$, and so we obtain a bound exactly half of that in Theorem 4.

4.2.3 Multi-class FLD

The multi-class version of FLD may be analyzed in extension to the two-class analyses above as follows:

Lemma 16

Let $\mathcal{C} = \{0, 1, \ldots, m\}$ be a collection of $m+1$ classes partitioning the data. Let $x_q|y \sim \mathcal{N}(\mu_y, \Sigma)$ (or $x_q|y$ drawn from a subgaussian distribution with mean μ_y and covariance matrix Σ).

Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Then, the probability that an unseen query point x_q is misclassified by FLD is upper bounded by:

$$Pr_{x_q,y_q}[\hat{h}(x_q) \neq y_q] \leqslant \sum_{y=0}^{m} \pi_y \sum_{i \neq y}^{m} \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_y - \hat{\mu}_i)^T \hat{\Sigma}^{-1}(\hat{\mu}_y - \hat{\mu}_i - 2\mu_y)\right]^2}{(\hat{\mu}_y - \hat{\mu}_i)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_y - \hat{\mu}_i)}\right) \tag{4.2.7}$$

Proof 3

The decision rule for FLD in the multi-class case is given by:

$$\hat{h}(x_q) = j \iff j = \underset{i}{arg\ max}\{Pr_{y_q}(y_q = i|x_q)\} \quad j, i \in \mathcal{C}$$

Without loss of generality, we again take the correct label of x_q to be 0. Then:

$$\hat{h}(x_q) = 0 \iff \bigwedge\nolimits_{i \neq 0}\{Pr_{y_q}(y_q = 0|x_q) \geqslant Pr_{y_q}(y_q = i|x_q)\} \tag{4.2.8}$$

$$\iff \bigwedge\nolimits_{i \neq 0}\left\{\tfrac{Pr_{y_q}(y_q=0|x_q)}{Pr_{y_q}(y_q=i|x_q)} \geqslant 1\right\} \tag{4.2.9}$$

and so misclassification occurs when:

$$\hat{h}(x_q) \neq 0 \iff \bigvee_{i \neq 0} \left\{ \frac{Pr_{y_q}(y_q = i|x_q)}{Pr_{y_q}(y_q = 0|x_q)} > 1 \right\}$$

Then since if $A \iff B$ then $Pr(A) = Pr(B)$, we have:

$$\begin{aligned} Pr_{x_q|y_q=0}[\hat{h}(x_q) \neq 0] &= Pr_{x_q|y_q=0} \left[\bigvee_{i \neq 0} \left\{ \frac{Pr_{y_q}(y_q = i|x_q)}{Pr_{y_q}(y_q = 0|x_q)} > 1 \right\} \right] \\ &\leqslant \sum_{i=1}^{m} Pr_{x_q|y_q=0} \left\{ \frac{Pr_{y_q}(y_q = i|x_q)}{Pr_{y_q}(y_q = 0|x_q)} > 1 \right\} \qquad (4.2.10) \\ &= \sum_{i=1}^{m} Pr_{x_q|y_q=0} \left\{ \log \frac{Pr_{y_q}(y_q = i|x_q)}{Pr_{y_q}(y_q = 0|x_q)} > 0 \right\} \qquad (4.2.11) \end{aligned}$$

where (4.2.10) *follows by the union bound. Writing out (4.2.11) via Bayes' rule, we find a sum of two-class error probabilities of the form that we have dealt with earlier, so (4.2.11) equals:*

$$\sum_{i=1}^{m} Pr_{x_q|y_q=0} \left\{ (\hat{\mu}_i - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_i}{2} \right) > 0 \right\} \qquad (4.2.12)$$

The result for the other possible values of $y_q \neq 0$ now follows by applying the bounding technique used for the two-class case m times to each of the m possible incorrect classes. The line of thought is then the same for $y = 1, \ldots, y = m$ in turn.

Comment 5

Owing to the straightforward way in which the multiclass error can be split into sums of two-class errors, as shown in lemma 16 above, it is therefore sufficient for the remainder of our analysis to be performed for the two-class case, and for $m + 1$ classes the error will always be upper bounded by m times the greatest of the two-class errors. We will use this fact later in Section 5.2.3.

4.3 Summary and Discussion

We prepared the way for the later chapters in this thesis by deriving upper bounds on the generalization error of Fisher's Linear Discriminant (FLD) for the large family of subgaussian distributions, under the mild constraint that we have enough observations to estimate the class means well enough to ensure that the sample mean in each class is closer to its true value than it is to the means of the other classes. In the two-class case this constraint on our bounds amounts to requiring only that, for both classes, the training examples for each class are not mostly outliers located in the same half space as the mean of the other class. We also derived the exact error for FLD in the case of Gaussian classes with the same covariance matrix, the situation in which FLD is Bayes optimal, without the constraint on the sample size. Finally we showed that we can bound the error of multi-class FLD by decomposition into a collection of two-class problems.

We note that for FLD we have left open the problems of data distributions where the classes are not Gaussian or subgaussian. Although we do not consider settings other than the Gaussian or subgaussian ones in this thesis we expect that for many unimodal class-conditional distributions, the error of FLD will exhibit qualitatively similar behaviour to that we find here. In particular, one could truncate such a distribution to obtain a two term bound containing a subgaussian part plus a constant term, the size of which will depend on how fat the tails of the distribution are. If we exclude distributions with particularly fat tails (e.g. classes that are t-distributed with very few degrees of freedom) then the sum of the two obtained terms will behave similarly to the bounds we derived here. We therefore see that our error bounds for subgaussian distributions should provide a reasonable characterization of the generalization error of FLD for many unimodal distributions.

5

Randomly-projected Fisher Linear Discriminant

Summary In this chapter, we consider random projections in conjunction with classification, specifically the analysis of Fisher's Linear Discriminant (FLD) classifier in randomly projected data spaces.

We focus on two-class classification problems since, as we saw from lemma 16 in the previous chapter, the FLD decision rule in multiclass settings can be decomposed as a sequence of two-class classification problems.

We are interested in quantifying the generalization error of randomly-projected FLD (RP-FLD) with respect to the $\{0, 1\}$-loss and discovering which are the key properties of a classification problem that make learning easy or hard in this setting.

We focus in this chapter on the setting where data are plentiful, which is a common situation in many problem domains including web-mining applications (e.g. search, data from social sites), science (e.g. Sloan Digital Sky Survey, Large Hadron Collider), and commerce (customer profiling, fraud detection)(Cukier, 2010). We defer dealing with the case when observations are few compared to the data dimensionality, another common situation, to our treatment of RP-FLD ensembles in chapter 7. In this chapter, and those that follow, we do not assume that the data have any special structural properties (such as sparsity). We derive average-case guarantees (w.r.t to random projection matrices) for the performance of RP-FLD and, in view of our findings that random projection reduces the condition number of the projected covariance matrix -vs- the data space covariance matrix, we also make an in-depth study of RP-FLD when the model covariance matrix is taken to be spherical.

5.1 Preliminaries

5.1.1 The randomly-projected classification problem

We recall the general problem setting of chapter 4 where, in a classification problem we observe N examples of labelled training data and we want to learn from this data a classifier which gives the lowest possible generalization error in terms of the $(0,1)$-loss $\mathcal{L}_{(0,1)}$. The case we consider here consists of instantiations of FLD learned on randomly-projected training data, $\mathcal{T}_N^R = \{(R(x_i), y_i)\}_{i=1}^N$ and we seek to bound the probability that an arbitrarily drawn query point $x_q \sim \mathcal{D}_{x|y}$ is misclassified by the learned classifier. In particular, we want to link the error of RP-FLD to the error of FLD learned in the dataspace, thereby quantifying the cost of working with the randomly projected data. Our approach is to upper bound the dataspace generalization error:

$$\mathrm{Pr}_{x_q,y_q}[\hat{h}(x_q) \neq y_q : (x_q, y_q) \sim \mathcal{D}_{x_q,y_q}] = \mathrm{E}_{x_q,y_q}[\mathcal{L}_{(0,1)}(\hat{h}(x_q), y_q)]$$

Where $\hat{h}$ is the classifier decision rule learned from the training data using the results derived in chapter 4, and then to bound the corresponding probability in a random projection of the data:

$$\mathrm{Pr}_{R,x_q,y_q}[\hat{h}^R(R(x_q)) \neq y_q : (x_q, y_q) \sim \mathcal{D}] = \mathrm{E}_{R,x_q,y_q}[\mathcal{L}_{(0,1)}(\hat{h}^R(Rx_q), y_q)]$$

Where $\hat{h}^R$ is the decision rule learned from the randomly-projected data. In the next section 5.1.2 we derive the decision rule for RP-FLD as the first step in this approach.

5.1.2 Decision rule for RP-FLD

Recall from chapter 4 that the FLD classifier learned from training data applies the decision rule:

$$\hat{h}(x_q) = \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \left(x_q - \frac{(\hat{\mu}_0 + \hat{\mu}_1)}{2}\right) > 0\right\} \tag{5.1.1}$$

When the training data are of the form $\mathcal{T}_N^R = \{(R(x_i), y_i)\}_{i=1}^N$ then, by linearity of the projection matrix R and of the expectation operator $\mathrm{E}[\cdot]$ we see that the decision rule for RP-FLD is given by:

$$\hat{h}^R(x_q) = \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left(R\hat{\Sigma}R^T\right)^{-1} R \left(x_q - \frac{(\hat{\mu}_0 + \hat{\mu}_1)}{2}\right) > 0\right\} \tag{5.1.2}$$

The derivation of the RP-FLD parameters is straightforward – details are in the appendix section A.2.

5.2 Analysis of RP-FLD

We now present two similar, but different, upper bounds on the generalization error of RP-FLD using theorem 4 and lemma 17 as our starting point. In this section we focus on the setting in which both the dimensionality, d, and the sample size N are large – these are typical properties of 'big data' problems such as data-mining of high throughput retail data. We consider the large d, small N case later in Chapter 7. For the bounds we derive here we consider finite random projection matrices with entries drawn i.i.d from the Gaussian $\mathcal{N}(0, \sigma^2)$, though we shall see that our final bounds also hold *mutatis mutandis* for random projection matrices with entries from zero-mean subgaussian distributions also. We use our first bound (Theorem 6), a version of which was originally published in Durrant & Kabán (2010b), to analyze the effect of covariance misspecification on RP-FLD and to quantify the projection dimension required to control the generalization error of RP-FLD. We find, in common with Dasgupta (1999), that random projection improves the conditioning of Σ_y and $\hat{\Sigma}$ and that therefore covariance misspecification, for example by approximating $\hat{\Sigma}$ with the identity I in the projected space, has a relatively benign effect on generalization error compared to a similar misspecification in the data space. Using the same bound we also quantify the projection dimension required for performance guarantees and in doing so we improve considerably on previous theoretical work which used the Johnson-Lindenstrauss lemma (JLL) or compressed sensing (CS) to derive generalization guarantees; we show that, unlike those settings where one takes the projection dimension logarithmic in the number of observations (JLL) or linear in the number of non-zero components (CS), in order to control the generalization error of RP-FLD it is enough to take the projection dimension, k, logarithmic in the number of *classes* which is of course typically a much smaller quantity. Finally we show that as the projection dimension $k \nearrow d$ the generalization error of RP-FLD decreases nearly exponentially; this is similar to the behaviour observed by Davenport et al. (2010); Haupt et al. (2006) where the authors analyze m-ary hypothesis testing of signals, but here we explicitly consider learning a classifier from data whereas in those works the authors assume the set of possible signals and all parameters to be estimated are perfectly known.

Our second bound, which was published in Durrant & Kabán (2010a; 2011) is simpler but is also tighter than our first bound when we take $\hat{\Sigma}$ to be spherical, and we therefore use this bound to analyze more carefully the case when $\hat{\Sigma} = \sigma^2 I$ (where $\sigma^2 > 0$ is a real-valued scalar) in the projected space.

We begin this section by decomposing the FLD bound of theorem 4 into two terms, one of which will go to zero as the number of training examples increases. This gives us the opportunity to assess the contribution of these two sources of error separately.

5.2.1 Decomposition of data space FLD bound as sum of 'estimated error' and 'estimation error'

Here, and from now on, we will write $\mathcal{D}_{x,y}$ for the joint distribution of data points and labels and $\mathcal{D}_{x|y}$ for the class-conditional distribution of x. We will assume that either x_q has a multivariate Gaussian class-conditional distribution so then $x_q|y_q = y \sim \mathcal{N}(\mu_y, \Sigma_y)$ with $\Pr\{y_q = y\} = \pi_y$, or else x_q is class-conditionally multivariate subgaussian with mean μ_y and covariance matrix Σ_y.

Lemma 17

Let $(x_q, y_q) \sim \mathcal{D}_{x,y}$ and let $\mathcal{H}$ be the class of FLD functions and $\hat{h}$ be the instance *learned from the training data* $\mathcal{T}_N$*. Write for the* estimated error*:*

$$\hat{B} := \sum_{y=0}^{1} \pi_y \hat{B}_y = \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma_y \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)} \right) \tag{5.2.1}$$

Similarly write $B := \sum_{y=0}^{1} \pi_y B_y$ *for the right hand side of the corollary to theorem 4, namely* (4.2.4)*. Then:*

$$Pr_{x_q, y_q}[\hat{h}(x_q) \neq y_q] \leqslant \hat{B} + C \cdot \sum_{y,i} |\hat{\mu}_{yi} - \mu_{yi}| \tag{5.2.2}$$

with $C := \max_{y,i} \sup \left\{ \left| \frac{\partial B_y}{\partial \mu_{yi}} \right| \right\}$ *a constant,* μ_y *the mean of the class from which* x_q *was drawn, estimated class means* $\hat{\mu}_y$ *with* $\hat{\mu}_{yi}$ *the* i*-th component, and model covariance* $\hat{\hat{\Sigma}}$*.*

Proof 4
We will use the mean value theorem[1]*, so we start by differentiating* $B = \sum_{y=0}^{1} \pi_y B_y$ *with respect to* μ_y*. Recalling that* $\hat{B}_0$ *and* $\hat{B}_1$ *are the two exp terms in (5.2.1), we have for the case* $y_q = 0$*:*

$$\nabla_{\mu_0} B = \pi_0 \hat{B}_0 \times \frac{1}{2} \alpha_0 \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \tag{5.2.3}$$

Now, provided that $\|\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0\| < +\infty$ *and* $0 < \|\hat{\mu}_1 - \hat{\mu}_0\| < +\infty$ *then* $\pi_0 \hat{B}_0$ *is bounded between zero and one and the supremum of the* i*-th component of this gradient exists. Therefore we have that:*

$$B \leqslant \pi_0 \hat{B}_0 + \max_i \sup \left\{ \left| \frac{\partial B_y}{\partial \mu_{0i}} \right| \right\} \sum_i |\hat{\mu}_{0i} - \mu_{0i}| \ldots$$
$$\ldots + (1 - \pi_0) Pr_{x_q | y_q = 1}[\hat{h}(x_q) \neq 1] \tag{5.2.4}$$

Then by applying the mean value theorem again w.r.t. μ_1 *with* $\|\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_1\| < +\infty$ *we can approach the* $y_q = 1$ *case similarly, and taking the maximum over both classes then yields lemma 17.*

We call the two terms obtained in (5.2.2) the 'estimated error' and 'estimation error' respectively. The estimation error can be bounded using Chernoff bounding techniques, and converges to zero as the number of training examples, N, increases. In particular, for the Gaussian and subgaussian distributions we consider here this convergence to zero is exponentially fast as a function of increasing N since, by lemma 5.9 of Vershynin (2012), the components $\hat{\mu}_{yi} - \mu_{yi}$ are zero-mean subgaussian random variables with the tails of their distribution decaying like those of the Gaussian $\mathcal{N}(0, \sigma_{yi}/N)$.

[1] Mean value theorem in several variables: Let f be differentiable on S, an open subset of $\mathbb{R}^d$, let x and y be points in S such that the line between x and y also lies in S. Then $\exists t \in (0, 1)$, such that $f(y) - f(x) = (\nabla f((1 - t)x + ty))^T (y - x)$

We now have the groundwork necessary to prove the main results in this section, namely bounds on this estimated misclassification probability if we choose to work with a k-dimensional random projection of the original data. From the results of lemma 17 and lemma 16, in order to study the behaviour of our bounds, we may restrict our attention to the two-class case and we focus on bounding the estimated error term which, provided sufficient training data, is the main source of error. From hereon we will use $\hat{\Pr} := \pi_0 \hat{B}_0 + (1-\pi_0)\hat{B}_1$ to denote the estimated error.

5.2.2 Main results: Generalization error bounds on RP-FLD

Our first bound on the (estimated) generalization error of RP-FLD is the following Theorem 6:

Theorem 6 *Let $(x_q, y_q) \sim \mathcal{D}_{x,y}$. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with entries drawn i.i.d from the univariate Gaussian $\mathcal{N}(0,\sigma^2)$. Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}^R$ be the instance learned from the randomly-projected training data $\mathcal{T}_N^R$.*
Assume that $\alpha_y^R := (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_{\neg y} - \hat{\mu}_y) > 0$, $\forall y$.
Then the estimated misclassification error $\hat{Pr}_{R,x_q,y_q}[\hat{h}^R(Rx_q) \neq y_q]$ is bounded above by:

$$\sum_{y=0}^{1} \pi_y \left(1 + \frac{1}{4} g(\hat{\Sigma}^{-1}\Sigma_y) \cdot \frac{1}{d}\frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{\lambda_{\max}(\Sigma_y)}\right)^{-k/2} \tag{5.2.5}$$

with μ_y the mean of class y, π_y the prior probability that x_q is drawn from class y, estimated class means $\hat{\mu}_0$ and $\hat{\mu}_1$, shared model covariance estimate $\hat{\Sigma}$, and $g(Q) = 4 \cdot \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \cdot \left(1 + \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}\right)^{-2}$.

Proof 5 (of Theorem 6)
We will start our proof in the dataspace, highlighting the contribution of covariance misspecification in the estimated error, and then make a move to the projected space with the use of a result (lemma 18) that shows that this component is always non-increasing under the random projection.
Without loss of generality we take $x_q \sim \mathcal{N}(\mu_0, \Sigma_0)$ then, by theorem 4 and lemma 17, the estimated misclassification error in this case is upper bounded by:

$$\exp\left(-\frac{1}{8} \cdot \frac{\left[(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\Sigma_0\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right) \tag{5.2.6}$$

Now, in the Kantorovich inequality (lemma 8) we can take:

$$v = \hat{\Sigma}^{-1/2}(\hat{\mu}_1 - \hat{\mu}_0)$$

where we use the fact from lemma 2 that since $\hat{\Sigma}^{-1}$ is symmetric positive definite it has *a unique symmetric positive definite square root:*

$$\hat{\Sigma}^{-1/2} = \left(\hat{\Sigma}^{-1}\right)^{1/2} = \left(\hat{\Sigma}^{1/2}\right)^{-1} = \left(\hat{\Sigma}^{-1/2}\right)^{T}$$

and, again by the properties given in lemma 2, we can take our positive definite Q to be $Q = \hat{\Sigma}^{-1/2}\Sigma_0\hat{\Sigma}^{-1/2}$. Then, by lemma 8 we have the expression (5.2.6) is less than or equal to:

$$\exp\left(-\frac{1}{8}\cdot(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1/2}\left[\hat{\Sigma}^{-1/2}\Sigma_0\hat{\Sigma}^{-1/2}\right]^{-1}\hat{\Sigma}^{-1/2}(\hat{\mu}_1-\hat{\mu}_0)\ldots\right.$$
$$\left.\ldots\times 4\cdot\frac{\lambda_{\max}\left(\hat{\Sigma}^{-1}\Sigma_0\right)}{\lambda_{\min}\left(\hat{\Sigma}^{-1}\Sigma_0\right)}\cdot\left(1+\frac{\lambda_{\max}\left(\hat{\Sigma}^{-1}\Sigma_0\right)}{\lambda_{\min}\left(\hat{\Sigma}^{-1}\Sigma_0\right)}\right)^{-2}\right) \tag{5.2.7}$$

where the change in argument for the eigenvalues comes from the use of the identity given in lemma 3 eigenvalues(AB) = eigenvalues(BA). After simplification we can write this as:

$$\exp\left(-\frac{1}{8}\cdot(\hat{\mu}_1-\hat{\mu}_0)^T\Sigma_0^{-1}(\hat{\mu}_1-\hat{\mu}_0)\cdot g(\hat{\Sigma}^{-1}\Sigma_0)\right) \tag{5.2.8}$$

The term $g(\hat{\Sigma}^{-1}\Sigma_0)$ is a function of the model covariance misspecification, e.g. due to the imposition of diagonal or spherical constraints on $\hat{\Sigma}$. The following lemma shows that this term of the error can only decrease or stay the same after random projection.

Lemma 18 (Non-increase of covariance misspecification error in projected space) *Let Q be a positive definite matrix. Let $\kappa(Q) = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \in [1,\infty)$ be the condition number of Q. Let $g(Q)$ be as given in the theorem 6. Then if we take $Q := (R\hat{\Sigma}R^T)^{-1/2}R\Sigma_yR^T(R\hat{\Sigma}R^T)^{-1/2}$ we have, for any fixed $k\times d$ matrix R with full row rank:*

$$g((R\hat{\Sigma}R^T)^{-1}R\Sigma_yR^T) \geqslant g(\hat{\Sigma}^{-1}\Sigma_y) \tag{5.2.9}$$

Proof: We will show that $g(\cdot)$ is monotonic decreasing with κ on $[1,\infty)$, then show that $\kappa((R\hat{\Sigma}R^T)^{-1}R\Sigma_yR^T) \leqslant \kappa(\hat{\Sigma}^{-1}\Sigma_y)$, and hence $g((R\hat{\Sigma}R^T)^{-1}R\Sigma_yR^T) \geqslant g(\hat{\Sigma}^{-1}\Sigma_y)$.

Step 1 *We show that g is monotonic decreasing:*
First note that for positive definite matrices $0 < \lambda_{\min} \leqslant \lambda_{\max}$, and so κ is indeed in $[1,\infty)$. Differentiating $g(\cdot)$ with respect to κ we get:

$$\frac{dg}{d\kappa} = \frac{4(1+\kappa)-8\kappa}{(1+\kappa)^3} = \frac{4(1-\kappa)}{(1+\kappa)^3}$$

Here the denominator is always positive on the range of κ while the numerator is always non-positive with maximum 0 at $\kappa = 1$. Hence $g(\cdot)$ is monotonic decreasing on $[1,\infty)$.

Step 2 *We show that $\kappa((R\hat{\Sigma}R^T)^{-1}R\Sigma_yR^T) \leqslant \kappa(\hat{\Sigma}^{-1}\Sigma_y)$:*
We will show that if $\hat{\Sigma}$ and Σ_y are symmetric positive definite and R is a matrix with full row rank then:

$$\lambda_{\min}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma_yR^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{5.2.10}$$

$$\geqslant \lambda_{\min}(\hat{\Sigma}^{-1}\Sigma_y) = \lambda_{\min}(\hat{\Sigma}^{-1/2}\Sigma_y\hat{\Sigma}^{-1/2}) \tag{5.2.11}$$

and

$$\lambda_{\max}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma_yR^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{5.2.12}$$

$$\leqslant \lambda_{\max}(\hat{\Sigma}^{-1}\Sigma_y) = \lambda_{\max}(\hat{\Sigma}^{-1/2}\Sigma_y\hat{\Sigma}^{-1/2}) \tag{5.2.13}$$

Combining these inequalities then gives:

$$\kappa((R\hat{\Sigma}R^T)^{-1}R\Sigma_yR^T) \leqslant \kappa(\hat{\Sigma}^{-1}\Sigma_y)$$

We give a proof of the first inequality, the second being proved similarly.

First, by lemma 5:

$$\lambda_{\min}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma_yR^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{5.2.14}$$

$$= \min_{u\in\mathbb{R}^k}\left\{\frac{u^T[R\hat{\Sigma}R^T]^{-1/2}R\Sigma_yR^T[R\hat{\Sigma}R^T]^{-1/2}u}{u^Tu}\right\} \tag{5.2.15}$$

Writing $v = [R\hat{\Sigma}R^T]^{-1/2}u$ so that $u = [R\hat{\Sigma}R^T]^{1/2}v$ then we may rewrite the expression (5.2.15), as the following:

$$= \min_{v\in\mathbb{R}^k}\left\{\frac{v^TR\Sigma_yR^Tv}{v^TR\hat{\Sigma}R^Tv}\right\} \tag{5.2.16}$$

Writing $w = R^Tv$, and noting that the span of all possible vectors w is a k-dimensional subspace of $\mathbb{R}^d$, we can bound the expression 5.2.16 by allowing the minimal vector $w \in \mathbb{R}^d$ not to lie in this subspace:

$$\geqslant \min_{w\in\mathbb{R}^d}\left\{\frac{w^T\Sigma_yw}{w^T\hat{\Sigma}w}\right\} \tag{5.2.17}$$

Now put $y = \hat{\Sigma}^{1/2}w$, with $y \in \mathbb{R}^d$. This y exists uniquely since $\hat{\Sigma}^{1/2}$ is invertible, and we may rewrite (5.2.17) as the following:

$$= \min_{y\in\mathbb{R}^d}\left\{\frac{y^T\hat{\Sigma}^{-1/2}\Sigma_y\hat{\Sigma}^{-1/2}y}{y^Ty}\right\} \tag{5.2.18}$$

$$= \lambda_{\min}(\hat{\Sigma}^{-1/2}\Sigma_y\hat{\Sigma}^{-1/2}) = \lambda_{\min}(\hat{\Sigma}^{-1}\Sigma_y) \tag{5.2.19}$$

This completes the proof of the first inequality, and a similar approach proves the second. Taken together the two inequalities imply $\kappa(\hat{\Sigma}^{-1}\Sigma_y) \geqslant \kappa([R\hat{\Sigma}R^T]^{-1}R\Sigma_yR^T)$ as required. Finally putting the results of steps 1 and 2 together gives the lemma 18. □

Back to the proof of theorem 6, we now move into the low dimensional space defined by any fixed instantiation of the random projection matrix R (i.e. with entries drawn from $\mathcal{N}(0, \sigma^2)$). By lemma 18, we can upper bound the projected space counterpart of (5.2.8) by the following:

$$\exp\left(-\frac{1}{8} \cdot (\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left[R\Sigma_0 R^T\right]^{-1} R(\hat{\mu}_1 - \hat{\mu}_0) \cdot g(\hat{\Sigma}^{-1}\Sigma_0)\right) \tag{5.2.20}$$

This holds for ***any*** *fixed matrix R with full row rank, so it also holds with probability 1 for any instantiation of a Gaussian random projection matrix R. Note, in the dataspace we bounded $Pr_{x_q,y_q}[\hat{h}(x_q) \neq y]$ but in the projected space we want to bound:*

$$\begin{aligned} Pr_{R,x_q,y_q}[\hat{h}^R(Rx_q) \neq y_q] &= E_{R,x_q,y_q}[\mathcal{L}_{(0,1)}(\hat{h}^R(Rx_q), y_q)] && (5.2.21)\\ &= E_R[E_{x_q,y_q}[\mathcal{L}_{(0,1)}(\hat{h}^R(Rx_q), y_q)]|R] && (5.2.22)\end{aligned}$$

This is the expectation of (5.2.20) w.r.t. the random choices of R. Restricting to the estimated error, and recalling that here $y_q = 0$ by assumption, we now have:

$$\begin{aligned}
&\hat{Pr}_{R,x_q,y_q}[\hat{h}^R(Rx_q) \neq 0|y_q = 0]\\
&\leqslant E_R\left[\exp\left(-\frac{1}{8} \cdot (\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left[R\Sigma_0 R^T\right]^{-1} R(\hat{\mu}_1 - \hat{\mu}_0) g(\hat{\Sigma}^{-1}\Sigma_0)\right)\right]\\
&= E_R\left[\exp\left(-\frac{1}{8} \cdot (\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left[RR^T\right]^{-1/2} \left[RR^T\right]^{1/2} \left[R\Sigma_0 R^T\right]^{-1} \left[RR^T\right]^{1/2} \ldots\right.\right.\\
&\qquad\qquad \left.\left.\ldots \left[RR^T\right]^{-1/2} R(\hat{\mu}_1 - \hat{\mu}_0) g(\hat{\Sigma}^{-1}\Sigma_0)\right)\right]\\
&= E_R\left[\exp\left(-\frac{1}{8} \cdot (\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left[RR^T\right]^{-1/2} \left[\left[RR^T\right]^{-1/2} R\Sigma_0 R^T \left[RR^T\right]^{-1/2}\right]^{-1} \ldots\right.\right.\\
&\qquad\qquad \left.\left.\ldots \left[RR^T\right]^{-1/2} R(\hat{\mu}_1 - \hat{\mu}_0) g(\hat{\Sigma}^{-1}\Sigma_0)\right)\right]\\
&\leqslant E_R\left[\exp\left(-\frac{1}{8} \cdot \frac{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left[RR^T\right]^{-1/2} \left[RR^T\right]^{-1/2} R(\hat{\mu}_1 - \hat{\mu}_0)}{\lambda_{\max}(\Sigma_0)} \cdot g(\hat{\Sigma}^{-1}\Sigma_0)\right)\right]
\end{aligned} \tag{5.2.23}$$

where the last step substituting $1/\lambda_{\max}(\Sigma_0)$ for $\left[\left[RR^T\right]^{-1/2} R\Sigma_0 R^T \left[RR^T\right]^{-1/2}\right]^{-1}$ is justified by lemma 7, since $\left[RR^T\right]^{-1/2} R$ has orthonormal rows. Now, by corollary 5 given in the Appendix, the moment generating function (5.2.23) is bounded above by the similar m.g.f:

$$E_R\left[\exp\left(-\frac{1}{8d} \cdot \frac{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T R(\hat{\mu}_1 - \hat{\mu}_0)}{\lambda_{\max}(\Sigma_0)\sigma^2} \cdot g(\hat{\Sigma}^{-1}\Sigma_0)\right)\right] \tag{5.2.24}$$

where the entries of R are drawn i.i.d from $\mathcal{N}(0, \sigma^2)$.

Then the term $(\hat{\mu}_1 - \hat{\mu}_0)^T R^T R(\hat{\mu}_1 - \hat{\mu}_0)/\sigma^2 = \|R(\hat{\mu}_1 - \hat{\mu}_0)\|^2/\sigma^2$ is χ^2_k distributed and (5.2.24) is therefore the moment generating function of a χ^2_k distribution (Weisstein). Hence we can upper bound (5.2.23) by the moment generating function of a χ^2 to *obtain:*

$$E_R\left[\exp\left(-\tfrac{1}{8}\cdot\tfrac{(\hat{\mu}_1-\hat{\mu}_0)^T R^T R(\hat{\mu}_1-\hat{\mu}_0)}{\lambda_{\max}(\Sigma_0)}\cdot g(\hat{\Sigma}^{-1}\Sigma_0)\right)\right] \tag{5.2.25}$$

$$\leqslant \left[1+\tfrac{1}{4}\cdot g(\hat{\Sigma}^{-1}\Sigma_0)\cdot\tfrac{\|\hat{\mu}_1-\hat{\mu}_0\|^2}{d\cdot\lambda_{\max}(\Sigma_0)}\right]^{-k/2} \tag{5.2.26}$$

A similar sequence of steps proves the other side, when $x_q \sim \mathcal{N}(\mu_1, \Sigma_1)$, and gives an expression of the same form except with Σ_0 replaced by Σ_1. Then putting the two terms together, applying the law of total probability with $\sum_y \pi_y = 1$ finally gives theorem 6.

A High Probability Guarantee

Following similar steps to those in the proof above it is possible to take a different route to obtain high probability guarantees with respect to the random projection matrix R on the estimated generalization error of RP-FLD. Starting from the expression (5.2.24) instead of taking expectation we could instead use Rayleigh quotient to obtain:

$$\exp\left(-\frac{1}{8d}\cdot\frac{(\hat{\mu}_1-\hat{\mu}_0)^T R^T R(\hat{\mu}_1-\hat{\mu}_0)}{\sigma^2\lambda_{\max}(\Sigma_y)}\cdot g(\hat{\Sigma}^{-1}\Sigma_y)\right) \tag{5.2.27}$$

$$= \exp\left(-\frac{1}{8d}\cdot\frac{\|R(\hat{\mu}_1-\hat{\mu}_0)\|^2}{\sigma^2\lambda_{\max}(\Sigma_y)}\cdot g(\hat{\Sigma}^{-1}\Sigma_y)\right) \tag{5.2.28}$$

Now using the Johnson-Lindenstrauss lemma 3, which controls the magnitude of the distortion under random projection of $\|\hat{\mu}_0 - \hat{\mu}_1\|^2$, we obtain the following high probability bound on the estimated generalization error of RP-FLD:

$$\Pr_R\left\{\hat{\Pr}_{x_q,y_q}[\hat{h}^R(Rx_q)\neq y_q]\leqslant\exp\left(-\frac{1}{8d}\cdot\frac{\sigma^2 k(1-\epsilon)\|\hat{\mu}_1-\hat{\mu}_0\|^2}{\sigma^2\lambda_{\max}(\Sigma_y)}\cdot g(\hat{\Sigma}^{-1}\Sigma_y)\right)\right\}\geqslant 1-\exp(-k\epsilon^2/8) \tag{5.2.29}$$

Comment: Other projections R

Although we have taken the entries of R to be drawn from $\mathcal{N}(0, \sigma^2)$ this was used only in the final step, in the form of the moment generating function of the χ^2 distribution. In consequence, other distributions that produce inequality in the step from equation (5.2.23) to equation (5.2.26) suffice. Such distributions include subgaussians; this is quite a rich class which includes, for example, Gaussian distributions and any distribution with bounded support (Vershynin, 2012) and some examples of suitable distributions may be found in (Achlioptas, 2003). Whether any deterministic projection R can be found that is both non-adaptive (i.e. makes no use of the training labels) and still yields a non-trivial guarantee for FLD in terms of only the data statistics seems a difficult open problem.

5.2.3 Bounds on the projected dimensionality k and discussion

For both practical and theoretical reasons, we would like to know to which dimensionality k we can project our original high dimensional data and still expect to recover good classification performance from RP-FLD. This may be thought of as a measure of the difficulty of the classification task.

By setting our bound in Theorem 6 on the average estimated generalization error to be no more than $\delta \in (0,1)$ and solving for k we can obtain such a bound on k for RP-FLD that guarantees that the expected misclassification probability (w.r.t. R) in the projected space remains below δ.[2]

Corollary 2 (to Theorem 6)
Let k, d, $g(\cdot)$, $\hat{\mu}_y$, Σ_y, $\hat{\Sigma}$ be as given in theorem 6. Let $\mathcal{C} = \{0, 1, \ldots, m-1\}$ be a set indexing the different classes. Then, in order that the probability of misclassification in the projected space remains below δ it is sufficient to take:

$$k \geqslant 8 \cdot \frac{1}{\min\limits_{i,j \in \mathcal{C}, i \neq j} \|\hat{\mu}_i - \hat{\mu}_j\|^2} \cdot \max_{y \in \mathcal{C}} \left\{ \frac{d\lambda_{\max}(\Sigma_y)}{g(\hat{\Sigma}^{-1}\Sigma_y)} \right\} \cdot \log(m/\delta) \tag{5.2.30}$$

Proof 6 (of corollary 2)
In the 2-class case we have:

$$\delta \geqslant \left[1 + \tfrac{1}{4} \cdot \min_{y \in \{0,1\}} \left\{ g(\hat{\Sigma}^{-1}\Sigma_y) \cdot \tfrac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{d\lambda_{\max}(\Sigma_y)} \right\} \right]^{-k/2} \iff \tag{5.2.31}$$

$$\log(1/\delta) \leqslant \tfrac{k}{2} \log \left[1 + \tfrac{1}{4} \cdot g(\hat{\Sigma}^{-1}\Sigma_y) \cdot \tfrac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{d\lambda_{\max}(\Sigma_y)} \right] \tag{5.2.32}$$

then using the inequality $(1 + x) \leqslant e^x$, $\forall\, x \in \mathbb{R}$ we obtain:

$$k \quad \geqslant 8 \cdot \tfrac{d\lambda_{\max}(\Sigma_y)}{\|\hat{\mu}_1 - \hat{\mu}_0\|^2} \cdot \tfrac{1}{g(\hat{\Sigma}^{-1}\Sigma_y)} \cdot \log(1/\delta) \tag{5.2.33}$$

Using (5.2.33) and lemma 16, it is then easy to see that to expect no more than δ error from FLD in an $m+1$-class problem, the required dimension of the projected space need only be:

$$k \geqslant 8 \cdot \frac{d\lambda_{\max}(\Sigma_y)}{\min_{i,j \in \mathcal{C}, i \neq j} \|\hat{\mu}_i - \hat{\mu}_j\|^2} \cdot \frac{1}{g(\hat{\Sigma}^{-1}\Sigma_y)} \cdot \log(m/\delta) \tag{5.2.34}$$

as required.

We find it interesting to compare our k bound with that given in the seminal paper of Arriaga and Vempala (Arriaga & Vempala, 1999). The analysis in (Watanabe et al., 2005) shows that the bound in (Arriaga & Vempala, 1999) for randomly projected 2-class perceptron classifiers means requiring that the projected dimensionality

$$k = \mathcal{O}\left(72 \cdot \frac{L}{l^2} \cdot \log(6N/\delta)\right) \tag{5.2.35}$$

where δ is the user-specified tolerance of misclassification probability, N is the number of training examples, and L/l^2 is the diameter of the data ($L = \max_{n=1,\ldots,N} \|x_n\|^2$)

[2] Although we do not do so here, we note that a variant of corollary 2, that holds with high probability, can be derived from the inequality (5.2.29) and lemma 16 using very similar steps.

divided by the margin (or 'robustness', as they term it). In our bound, $g(\cdot)$ is a function that encodes the quality of the model covariance specification, δ and k are the same as in (Arriaga & Vempala, 1999) and the factor $d\lambda_{\max}(\Sigma_y) \cdot \|\hat{\mu}_1 - \hat{\mu}_0\|^{-2}$ – which, should be noted, is exactly the reciprocal of the squared class separation as defined by Dasgupta in (Dasgupta, 1999) – may be thought of as the 'generative' analogue of the data diameter divided by the margin in (5.2.35).

Observe, however, that (5.2.35) grows with the log of the training set size, whereas ours (5.2.33) grows with the log of the number of classes. This is not to say, by any means, that FLD is superior to perceptrons in the projected space. Instead, the root and significance of this difference lies in the assumptions (and hence the methodology) used in obtaining the bounds. The result in (5.2.35) was derived from the precondition that all pairwise distances between the training points must be approximately preserved *uniformly* cf. the Johnson-Lindenstrauss lemma (Dasgupta & Gupta, 2002). It is well understood (Alon, 2003) that examples of data sets exist for which the $k = \mathcal{O}(\log N)$ dimensions are indeed required for this. However, we conjecture that, for learning, this starting point is too strong a requirement. Learning should not become harder with more training points, assuming of course that the additional examples add 'information' to the training set.

Our derivation is so far specific to FLD, but it is able to take advantage of the class structure inherent in the classification setting in that the misclassification error probability is down to very few key distances only – the ones between the class centres.

Despite this difference from (Arriaga & Vempala, 1999) and approaches based on uniform distance preservation, in fact our conclusions should not be too surprising. Earlier work in theoretical computer science in (Dasgupta, 1999) proves performance guarantees with high probability (over the choice of R) for the *unsupervised* learning of a mixture of Gaussians which also requires k to grow logarithmically with the number of centres only. Moreover, our finding that the error from covariance misspecification is always non-increasing in the projection space is also somewhat expected, in the light of the finding in (Dasgupta, 1999) that projected covariances tend to become more spherical.

It is also worth noting that the extensive empirical results in e.g. (Dasgupta, 2000a) and (Fradkin & Madigan, 2003b) also suggest that classification (including non-sparse data) requires a much lower projection dimension than that which is needed for global preservation of all pairwise distances via the JLL. We therefore conjecture that, all other things being equal, the difficulty of a classification task should be a function only of selected distances, and preserving those may be easier that preserving every pairwise distance uniformly. Investigating this more generally remains for further research.

5.2.4 Numerical validation

We present three numerical tests that illustrate and confirm our main results.

Lemma 18 showed that the error contribution of a covariance misspecification is always no worse in the low dimensional space than in the high dimensional space. Figure 5.1 shows the quality of fit between a full covariance Σ and its diagonal approximation $\hat{\Sigma}$ when projected from a $d = 100$ dimensional data space into successively lower dimensions k. We see the fit is poor in the high dimensional space, and it keeps improving as k

gets smaller. The error bars span the minimum and maximum of $g([R\hat{\Sigma}R^T]^{-1}R\Sigma_y R^T)$ observed over 40 repeated trials for each k. The second set of experiments demonstrates Corollary 2 of our Theorem 6, namely that for good generalization of FLD in the projected space, the required projection dimension k is logarithmic in the number of classes.

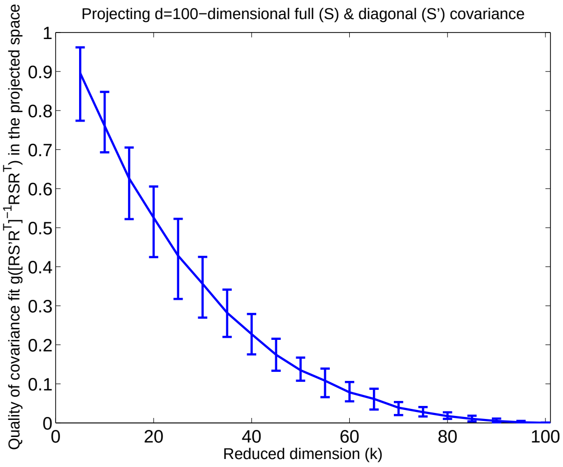

Figure 5.1: Experiment confirming Lemma 18; the error contribution of a covariance misspecification is always no worse in the projected space than in the data space. The best possible value on the vertical axis is 1, the worst is 0. We see the quality of fit is poor in high dimensions and improves dramatically in the projected space, approaching the best value as k decreases.

We randomly projected m equally distanced spherical unit variance 7-separated Gaussian classes ($\|\mu_i - \mu_j\| = 7,\ \forall i \neq j \in \{0, 1, \ldots, m-1\},\ \Sigma_i = I,\ \forall i$)from $d = 100$ dimensions and chose the target dimension of the projected space as $k = 12\log(m)$. The boxplots in figure 5.2 show, for each m tested, the distribution of the empirical error rates over 100 random realisations of R, where for each R the empirical error was estimated from 500 independent query points. Other parameters being unchanged, we see the classification performance is indeed maintained with this choice of k.

The third experiment shows the effect of reducing k for a 10-class problem in the same setting as experiment two. As expected, the classification error in figure 5.3 decreases nearly exponentially as the projected dimensionality k tends to the data dimensionality d. We note also, from these empirical results, that the variability in the classification performance also decreases with increasing k. Finally, we observe that the worst performance in the worst case is still a weak learner that performs better than chance.

We now come to our second bound on RP-FLD, which is derived similarly to Theorem 6 but using a slightly more straightforward approach. This bound is generally quantitatively sharper and especially so when $\hat{\Sigma}$ (or Σ) is spherical, and we therefore use this bound to study in depth the case of spherical model covariance. We shall see that in this case we can also quantify exactly the probability that, if the condition

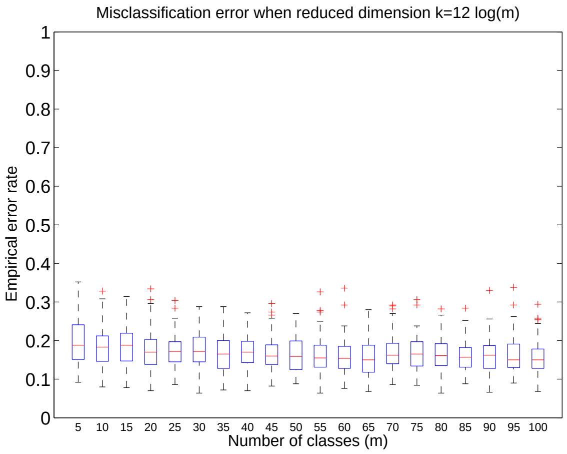

Figure 5.2: Experiment illustrating Theorem 6 & its Corollary 2. With the choice $k = 12\log(m)$ and $\|\mu_i - \mu_j\| = 7$, $\forall i \neq j$, the classification performance is kept at similar rates while the number of classes m varies.

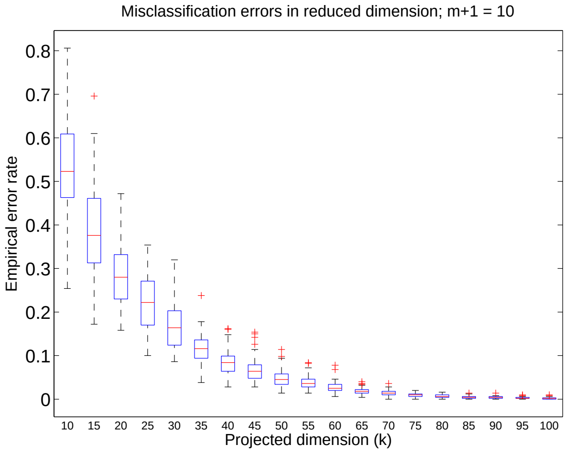

Figure 5.3: Experiment illustrating Theorem 6. We fix the number of classes, $m+1 = 10$, and the data dimensionality, $d = 100$, and vary the projection dimensionality k. The classification error decreases nearly exponentially as $k \to d$.

$\alpha_y > 0$, $\forall y$ holds in the data space (i.e. the estimated and true means for each class are both on the same side of the decision boundary as one another in the data space), then the corresponding condition in the projected space $\alpha_y^R > 0$, $\forall y$ fails with a corresponding reduction in generalization performance. That is, we give the exact formula for $\Pr\{\exists y : \alpha_y^R \leqslant 0\}$.

Theorem 7 *Under the same conditions as theorem 6 the estimated misclassification* *error $\hat{Pr}_{R,x_q,y_q}[\hat{h}^R(Rx_q) \neq y_q]$ is bounded above by:*

$$\begin{aligned}&\hat{Pr}_{R,x_q,y_q}[\hat{h}^R(Rx_q) \neq y_q] \\ \leqslant\ & \sum_{y=0}^{1} \pi_y \exp\left(-\frac{k}{2}\log\left(1+\frac{1}{4d}\cdot\|\hat{\mu}_y-\hat{\mu}_{\neg y}\|^2\cdot\frac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right)\right)\end{aligned} \tag{5.2.36}$$

Working as before, we bound the estimated error and the proof of this bound is very similar to that employed for Theorem 6: We first obtain a bound that holds for any fixed instantiation of the random projection matrix R, and finally on average over all R.

Proof 7 (of Theorem 7)
By theorem 4 and lemma 17, the estimated error in the projected space defined by any given instance of R is upper bounded by:

$$\begin{aligned}&\hat{Pr}_{x_q,y_q}\left[\hat{h}^R(Rx_q) \neq y_q\right] \\ \leqslant & \sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\frac{\left[(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T R^T (R\hat{\Sigma}R^T)^{-1}R(\hat{\mu}_{\neg y}-\hat{\mu}_y)\right]^2}{(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T R^T (R\hat{\Sigma}R^T)^{-1}R\Sigma_y R^T(R\hat{\Sigma}R^T)^{-1}R(\hat{\mu}_{\neg y}-\hat{\mu}_y)}\right)\end{aligned} \tag{5.2.37}$$

We would like to analyse the expectation of (5.2.37) w.r.t the random choices of R in terms of the quantities of the original space. To this end, we first proceed by rewriting and further bounding (5.2.37) using majorization of the numerator by the Rayleigh quotient (lemma 5), where we take $v = \left(R\hat{\Sigma}R^T\right)^{-1/2} R(\hat{\mu}_{\neg y}-\hat{\mu}_y)$ and take, for the y-th class, our positive definite Q_y to be $Q_y = \left(R\hat{\Sigma}R^T\right)^{-1/2} R\Sigma_y R^T \left(R\hat{\Sigma}R^T\right)^{-1/2}$ where we use lemma 2 and take $\left(R\hat{\Sigma}R^T\right)^{-1/2}$ to be the unique symmetric positive definite square root of $\left(R\hat{\Sigma}R^T\right)^{-1}$. Then, we have (5.2.37) is less than or equal to:

$$\sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\frac{\left[(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T R^T \left(R\hat{\Sigma}R^T\right)^{-1} R(\hat{\mu}_{\neg y}-\hat{\mu}_y)\right]^2}{\lambda_{\max}(Q_y)(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T R^T \left(R\hat{\Sigma}R^T\right)^{-1} R(\hat{\mu}_{\neg y}-\hat{\mu}_y)}\right) \tag{5.2.38}$$

Simplifying and using the fact that, whenever both multiplications are defined, the non-zero eigenvalues of the matrix AB are the same as the non-zero eigenvalues of the matrix BA, for each term in the summation we may write $\lambda_{\max}(Q_y) = \lambda_{\max}\left((R\hat{\Sigma}R^T)^{-1}R\Sigma_y R^T\right)$

and we may now further bound the expression (5.2.37) from above with:

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8}\cdot\frac{(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T R^T\left(R\hat{\Sigma}R^T\right)^{-1}R(\hat{\mu}_{\neg y}-\hat{\mu}_y)}{\lambda_{\max}\left((R\hat{\Sigma}R^T)^{-1}R\Sigma_y R^T\right)}\right) \tag{5.2.39}$$

$$\leqslant \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8}\cdot\frac{\|R(\hat{\mu}_{\neg y}-\hat{\mu}_y)\|^2}{\lambda_{\max}(\hat{\Sigma})}\frac{1}{\lambda_{\max}(\left(R\hat{\Sigma}R^T\right)^{-1}R\Sigma_y R^T)}\right) \tag{5.2.40}$$

where in the last line we used minorization by Rayleigh quotient of the numerator and inserted $(RR^T)^{-1/2}(RR^T)^{1/2}$, as we did before in equation (5.2.23), in order to apply Poincaré separation theorem 6 to $\left(R\hat{\Sigma}R^T\right)^{-1}$.
Continuing, we upper bound equation (5.2.40) with:

$$\sum_{y=0}^{1} \pi_y \exp\left(-\tfrac{1}{8}\cdot\tfrac{\|R(\hat{\mu}_{\neg y}-\hat{\mu}_y)\|^2}{\lambda_{\max}(\hat{\Sigma})}\tfrac{1}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right) \tag{5.2.41}$$

$$= \sum_{y=0}^{1} \pi_y \exp\left(-\tfrac{1}{8}\cdot\|R(\hat{\mu}_{\neg y}-\hat{\mu}_y)\|^2\cdot\tfrac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right) \tag{5.2.42}$$

where the change of term in the denominator uses the fact that $\lambda_{\max}(R\Sigma_y R^T\hat{\Sigma}_R^{-1}) \leqslant \lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})$, which we proved earlier in lemma 18.

The bound in (5.2.42) holds deterministically, for any fixed projection matrix R. We can also see from (5.2.42) that, by the Johnson-Lindenstrauss lemma, with high probability (over the choice of R) the misclassification error will also be exponentially decaying, except with $\frac{k}{d}(1-\epsilon)\|(\hat{\mu}_1-\hat{\mu}_0)\|^2$ in place of $\|R(\hat{\mu}_1-\hat{\mu}_0)\|^2$. However, we are more interested in the misclassification probability on average over all random choices of R. To complete the proof we again upper bound this expression with the m.g.f of a χ_k^2 distribution using the corollary 5 given in the Appendix to obtain:

$$\begin{aligned}
& \sum_{y=0}^{1} \pi_y E_R\left[\exp\left(-\tfrac{1}{8}\cdot\|R(\hat{\mu}_{\neg y}-\hat{\mu}_y)\|^2\cdot\tfrac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right)\right] \\
= \quad & \sum_{y=0}^{1} \pi_y \left(1+\left(\tfrac{1}{4d}\cdot\|(\hat{\mu}_{\neg y}-\hat{\mu}_y)\|^2\cdot\tfrac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right)\right)^{-k/2} \\
= \quad & \sum_{y=0}^{1} \pi_y \exp\left(-\tfrac{k}{2}\log\left(1+\tfrac{1}{4d}\cdot\|(\hat{\mu}_{\neg y}-\hat{\mu}_y)\|^2\cdot\tfrac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right)\right)
\end{aligned}$$

which, noting that under the $(0,1)$-loss the probability of an error coincides with the expected error, finally yields theorem 7.

5.2.5 Relation to Theorem 6

It is interesting to briefly compare the upper bound we just derived on the average estimated error of randomly projected FLD with the alternative bound we gave in theorem 6. Both bounds have the same preconditions, but theorem 6 has (after a little algebra) the following different form:

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{k}{2}\log\left(1+\frac{1}{4d}\cdot\|\mu_{\neg y}-\mu_y\|^2\cdot\frac{g(\Sigma_y\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y)}\right)\right) \tag{5.2.43}$$

where $g(Q) = 4\cdot\frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}\cdot\left(1+\frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}\right)^{-2}$.

We see by comparing the final forms of theorems 6 and 7 that the two bounds differ in that theorem 6 has the function $f_1(\hat{\Sigma},\Sigma) := g(\Sigma\hat{\Sigma}^{-1})/\lambda_{\max}(\Sigma)$ in the bound whereas in theorem 7 we have $f_2(\hat{\Sigma},\Sigma) := \lambda_{\min}(\hat{\Sigma}^{-1})/\lambda_{\max}(\Sigma\hat{\Sigma}^{-1})$ in its place. Note that therefore both bounds are invariant to scalings of $\hat{\Sigma}$, but monotonic in the eigenvalues of Σ. This is a desirable property in this setting, because it mirrors the behaviour of the FLD classifier. Denote by f_1^*, f_2^* the maximum values taken by these functions (that is, when the bounds are tightest). Then both f_1 and f_2 take their maximum value when $\hat{\Sigma} = \Sigma$ and we then have:

$$f_1^* = f_1(\hat{\Sigma}=\Sigma,\Sigma) = \frac{1}{\lambda_{\max}(\Sigma)} = f_2(\hat{\Sigma}=\Sigma,\Sigma) = f_2^* \tag{5.2.44}$$

so both bounds coincide when $\hat{\Sigma} = \Sigma$. For $\hat{\Sigma} \neq \Sigma$ in turn f_1 becomes smaller (the bound becomes larger) and this property makes it useful for studying the effects of covariance misspecification in the projected space, as we saw earlier in this chapter.

On the other hand, the bound of theorem 7 is quantitatively sharper in particular covariance settings, most notably it also takes its best value when $\hat{\Sigma}$ is estimated to be spherical (i.e. a scalar multiple of the identity matrix). Indeed in this case the $\lambda_{\max}$ term in the denominator of theorem 7 factorizes and we have

$$f_2(\hat{\Sigma}=I,\Sigma) = \frac{1}{\lambda_{\max}(\Sigma)} = f_1^* \tag{5.2.45}$$

since the $\lambda_{\min}(\hat{\Sigma}^{-1})$ in the numerator cancels with the $\lambda_{\max}(\hat{\Sigma}^{-1})$ in the denominator. Hence, it is natural to use the bound of theorem 7 to study the spherical model covariance setting in more detail.

Furthermore, this setting is one of particular interest since our earlier analysis above showed that the error arising from covariance misspecification in the projected space is never greater than the corresponding error in the data space, and therefore a simplified covariance model in the projected space has a relatively benign effect on classification performance compared to a similar covariance misspecification in the data space.

For these two reasons the remainder of this chapter will consider randomly projected FLD in the spherical model setting in more depth, and we will see that in the spherical covariance setting we can bound the average estimated error tightly even if we relax the condition required on theorem 7. Figure 5.2.5 gives a comparison of the bounds of theorem 6 and theorem 7 against empirical estimates of the misclassification error of randomly projected FLD. The misclassification error is estimated from 2000 random query points drawn from one of two Gaussian classes with identical covariance matrices and averaged over 2500 random projections. The data dimensionality is $d = 100$ in each case and the projected dimensionality is $k \in \{1, 10, 20, \ldots, 100\}$. The constant $c := \frac{\|\mu_0 - \mu_1\|}{\sqrt{d \cdot \lambda_{\max}(\Sigma)}}$ is the class separation metric used by Dasgupta in (Dasgupta, 1999; 2000a).

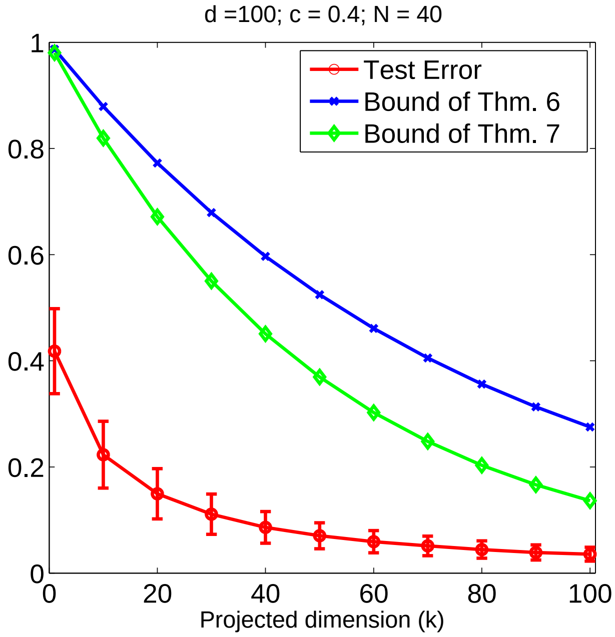

Figure 5.4: Comparison of our bounds from theorem 6 and 7 against empirical estimates of misclassification error for two identical $c = 0.4$-separated Gaussian classes with $\lambda_{\max}(\Sigma) \simeq 4$. We ran 2500 trials, fixing the data dimensionality at $d = 100$ while k varies. Error bars mark one standard deviation. In each trial the empirical error is estimated from 2000 randomly drawn query points.

5.2.6 The mean flipping problem

In theorem 7 we give the probability of misclassification error in the projected space, conditional on $\alpha_y^R > 0$. We mentioned that this was equivalent to requiring that none of the class means were ‘flipped’ by random projection, which requires some explanation. Recall that in our data space bound we make the (reasonable) assumption that if we have sufficient data then in the pairs of true and estimated means for each class, both means lie on the same side of the decision boundary. However, in the projected space it is not at all obvious that this remains a reasonable assumption; in fact it seems quite possible that the true mean vector could be ‘flipped’ across the decision boundary by random projection. It is interesting to consider if this is in fact the case and, if so, can we quantify the likelihood of this event? We are in fact able to find the exact probability of this event in the case that we replace $\hat{\Sigma}$ with a spherical model covariance (i.e. a scalar multiple of the identity) as is sometimes done in practice for computational reasons. However from simulations (see figure 6.3) it appears that for

non-spherical $\hat{\Sigma}$ the flipping probability is typically greater than in the spherical case and also far less well-behaved.
We therefore once again restrict our attention in the following discussion to the case of spherical $\hat{\Sigma}$ where we can show that for any fixed pair of vectors $n = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)$ and $m = (\hat{\mu}_{\neg y} - \hat{\mu}_y) \in \mathbb{R}^d$ with angular separation $\theta \in [0, \pi/2]$ in the data space, the probability of flipping: (i) reduces exponentially with increasing k and is typically very small even when k is very small (for example, when $k = 5$ the two vectors must be separated by about 30° for the flip probability to be much above machine precision), and (ii) is independent of the original data dimensionality d.
We will recall these properties shortly, when we combine our estimated error bound with the flip probability in section 5.2.7. For now, we state the theorem:

Theorem 8 (Flip Probability) *Let $n,\ m \in \mathbb{R}^d$ with angular separation $\theta \in [0, \pi/2]$. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with entries $r_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and let $Rn,\ Rm \in \mathbb{R}^k$ be the projections of $n,\ m$ into $\mathbb{R}^k$ with angular separation θ_R.
Then the 'flip probability' $Pr_R[\theta_R > \pi/2|\theta] = Pr_R[(Rn)^T Rm < 0|n^T m \geqslant 0] = Pr_R[\alpha_y^R < 0|\alpha_y \geqslant 0]$ is given by:*

$$Pr_R[\theta_R > \pi/2|\theta] = \frac{\int_0^\theta \sin^{k-1}(\phi)\ d\phi}{\int_0^\pi \sin^{k-1}(\phi)\ d\phi} \tag{5.2.46}$$

The proof of theorem 8 is given in the next chapter 6. Note particularly the surprising fact that the flip probability in theorem 8 depends only on the angular separation of the true and sample means in a particular class and on the projection dimensionality k. In fact equation (5.2.46) decays exponentially with increasing k. To see this, we note that this probability can be interpreted geometrically as the proportion of the surface of the k-dimensional unit sphere covered by a spherical cap subtending an angle of 2θ (see section 6.6 in the next chapter), and this quantity is bounded above by $\exp\left(-\frac{1}{2}k\cos^2(\theta)\right)$ ((Ball, 1997), Lemma 2.2, Pg 11).

5.2.7 Corollary to theorems 7 and 8

Taking into account the flip probability, we may now give the following bound on the estimated error when $\hat{\Sigma}$ is replaced by a multiple of the identity matrix and assuming only that $\alpha_y > 0$ holds in the data space (i.e. with no similar condition in the projected space):

Corollary 3
Let $(x_q, y_q) \sim \mathcal{D}_{x,y}$. Assume there is sufficient training data so that in the data space

$\alpha_y > 0$. Let $\hat{\Sigma}$ be spherical. Then with the notation of theorems 7 and 8 we have:

$$\begin{aligned} &Pr_{x_q,y_q,R}[\hat{h}^R(Rx_q) \neq y_q] \leqslant \\ &\sum_{y=0}^{1} \pi_y Pr_R[\alpha_y^R > 0] \cdot \exp\left(-\frac{k}{2}\log\left(1+\frac{1}{4d}\cdot\|\hat{\mu}_y-\hat{\mu}_{\neg y}\|^2\cdot\frac{1}{\lambda_{\max}(\Sigma_y)}\right)\right) \\ &+\sum_{y=0}^{1}\pi_y(1-Pr_R[\alpha_y^R>0]) \end{aligned} \tag{5.2.47}$$

where $\pi_y = Pr[y = y_q]$

Proof 8
Consider x_q drawn from class $y_q \in \{0, 1\}$. We have, by the law of total probability:

$$\begin{aligned} Pr_{x_q,y_q,R}[\hat{h}^R(Rx_q) \neq y_q] = \ & \textstyle\sum_{y=0}^{1} \pi_y \Big(Pr_R[\alpha_y^R > 0] \cdot Pr_{x_q,R}[\hat{h}^R(Rx_q) \neq y_q | y_q = y, \alpha_y^R > 0] \\ & +(1 - Pr_R[\alpha_y^R > 0]) \cdot Pr_{x_q,R}[\hat{h}^R(Rx_q) \neq y_q | y_q = y, \alpha_y^R \leqslant 0]\Big) \end{aligned} \tag{5.2.48}$$

Then expanding the bracket and taking the worst case when flipping occurs, we get the stated bound.

Note that the first sum is always no greater than the bound given in Theorem 7 since $\Pr_R[\alpha_y^R > 0]$ is always smaller than 1. Furthermore, the second sum $\sum_{y=0}^{1} \pi_y(1 - \Pr_R[\alpha_y^R > 0])$ is a convex combination of flip probabilities, and this term is typically small because it is independent of d and decays exponentially with increasing k.
We conclude that, provided we have a sufficient number of observations to ensure that $\alpha_y > 0$ in the data space, the problem of flipping typically makes a very small contribution to the error (on average, over the random picks of R) of the projected FLD classifier unless k is chosen to be extremely small (for example, $k = 1$).

5.3 Summary and Discussion

This chapter presented our findings concerning the effect of dimensionality reduction using random projection on the performance of FLD. In this chapter our conceptual view of random projection was as a form of lossy data compression, and we aimed to give a characterization of how lossy it is in the context of classification.
We restricted our attention to the setting of high-dimensional data and sufficient training examples, in which setting our bound on the 'estimated error' captures faithfully the generalization performance of FLD and RP-FLD. In particular, in the randomly projected domain we have fewer parameters to estimate than in the data space, but the same amount of data with which to estimate them, and so we can expect the 'estimation error' in the projected domain to be generally small – the estimated error of course increases.
Our main aim in this chapter was to quantify the performance cost of working with

randomly-projected data, at least for FLD, and this was achieved by assuming a fixed training set and evaluating the effect of random projection on the estimated error term. Our central motivation for taking this approach was the observation that, in a classification setting, often some distances are more important than others and so it should be possible to preserve classification performance provided one could preserve only those important distances. In particular uniform approximate preservation of data geometry through the JLL, and the consequent unnatural behaviour w.r.t the number of observations in bounds utilising it, appeared to us to be too strong a requirement in order to give guarantees on classification performance.

We conjectured that one should therefore be able to give guarantees on classifier performance in the randomly projected domain where, all other things being equal, the performance guarantee depends only in some simple way on the projection dimensionality and the number of important distances. In the case of RP-FLD this is indeed the case, and the number of important distances is the same as the number of classes because of the particularly simple structure of this classifier. Most other classification regimes have a significantly more complex structure than FLD but, since other generative classifiers still use the notion of the distance between a query point and some modelled distribution in order to assign a label to the query point, we believe that it should be possible to extend this approach to (at least some of) these more complex scenarios. We recognize however that such extensions are unlikely to be straightforward, even if one were to restrict the possible data distributions as we have here.

Finally, we see that the generalization error of RP-FLD depends on the data distribution and the parameters estimated from the data. Therefore a further alternative approach to bounding the generalization error would be to work with the data space bound (Thm. 4) or the exact error (Thm. 5) from Chapter 4 and then to quantify how good the parameter estimates can be expected to be with high probability over training sets of size N. From there we could derive high probability guarantees on the worst-case generalization error for subgaussian classes in terms of the sample size, data and projected dimensions, and true parameters. In fact this approach did not occur to us until we had already completed the work covered in this chapter, although with hindsight it looks very natural. We follow this approach in Chapter 7 where we consider an RP-FLD ensemble to address the issue of a small training sample, and in Chapter 8 when we study the generalization error of Kernel FLD. The techniques applied in those two chapters can be applied in an obvious way to extend the results in this chapter to give high-probability guarantees, but to avoid unnecessary repetition we have not done so here.

6

Flip Probabilities for Random Projections of θ-separated Vectors

Summary In theorems 6 and 7 we gave average case bounds (w.r.t Gaussian random projection matrices R) on the estimated error of RP-FLD for a fixed training set, provided that the condition $\alpha_y^R > 0$ held in the projected space. Although we see that this condition $\alpha_y > 0$ is not a particularly restrictive one in the data space, given our assumption in chapter 5 of sufficient data, in order to understand better how restrictive our condition is in the projected space we would like to better quantify the probability that $\alpha_y^R \leqslant 0$ in the projected space when $\alpha_y > 0$ in the data space.

In this chapter we therefore derive the exact probability that α_y and α_y^R have different signs in the restricted case when $\hat{\Sigma} = \sigma^2 I$ is a scalar multiple of the identity matrix (e.g. taken to be a scalar matrix for computational reasons).

6.1 Preliminaries

We consider the case of two vectors $n, m \in \mathbb{R}^d$ with the angle between them $\theta \in [0, \pi/2]$ which we randomly project by premultiplying them with a random matrix R with entries drawn i.i.d from the Gaussian $\mathcal{N}(0, \sigma^2)$. As a consequence of the Johnson-Lindenstrauss lemma, the angle between the projected vectors Rn, Rm is approximately θ with high probability (see e.g. (Arriaga & Vempala, 1999)), so the images of the vectors n, m under the same random projection are *not* independent. We want to find the probability that following random projection the angle between these vectors $\theta_R > \pi/2$, i.e. switches from being acute to being obtuse, which we call the 'flip probability'. In the context of RP-FLD and our condition $\alpha_y > 0$, if we take $\hat{\Sigma} = \sigma^2 I$ then we have $n = \hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y$ and $m = \sigma^2(\hat{\mu}_{\neg y} - \hat{\mu}_y)$, and without loss of generality we can assume the positive scaling constant $\sigma^2 = 1$ since it does not affect the sign of the dot product $n^T m$.

The proof proceeds by carrying out a whitening transform on each coordinate of the pair of projected vectors, and will make use of techniques inspired from the study of random triangles in (Eisenberg & Sullivan, 1996) to derive the probability. We obtain the exact expression for the flip probability in the form of an integral that has no general analytic closed form, although for any particular choice of k and θ this integral can be evaluated (in principle) using integration by parts. However this integral does turn out to have a natural geometrical interpretation as the quotient of the surface area of a (hyper-)spherical cap by the surface area of the corresponding (hyper-)sphere.

Before commencing the proof proper we make some preliminary observations. First recall, from the definition of the dot product, $n^T m = \|n\|\|m\| \cos\theta$, where $\theta \in [-\pi/2, \pi/2]$ is the principal angle between n and m, we have $n^T m < 0 \Leftrightarrow \cos\theta < 0$ and so the dot product is positive if and only if the principal angle between the vectors n and m is $\theta \in [0, \pi/2]$.

Hence, for $\theta \in [0, \pi/2]$ in the original d-dimensional space and θ_R in the k-dimensional randomly projected space we have $\Pr_R[\theta_R > \pi/2] = \Pr_R[(Rn)^T Rm < 0]$, and this is the probability of our interest[1]. Regarding random Gaussian matrices we note that, for any non-zero vector $x \in \mathbb{R}^d$, the event: $Rx = 0$ has probability zero with respect to the random choices of R. This is because the null space of R, $\ker(R) = R(\mathbb{R}^d)^\perp$, is a linear subspace of $\mathbb{R}^d$ with dimension $d - k < d$, and therefore $\ker(R)$ has zero Gaussian measure in $\mathbb{R}^d$. Hence $\Pr_R\{x \in \ker(R)\} = \Pr_R\{Rx = 0\} = 0$.

In a similar way, R almost surely has rank k. Denote the i-th row of R by $(r_{i1}, \ldots, r_{id})$, then the event: $\text{span}\{(r_{i1}, \ldots, r_{id})\} = \text{span}\{(r_{i'1}, \ldots, r_{i'd})\}$, $i \neq i'$ has probability zero since $\text{span}\{(r_{i1}, \ldots, r_{id})\}$ is a 1-dimensional linear subspace of $\mathbb{R}^d$ with measure zero. By induction, for finite $k < d$, the probability that the j-th row is in the span of the first $j - 1$ rows is likewise zero. In this setting we may therefore safely assume that $n, m \notin \ker(R)$ and that R has rank k.

Finally, since we are concerned only with the angles between n, m and Rn, Rm which are unaffected by their norms, we may assume without loss of generality that $\|n\| =$

[1] In fact the arguments for the proof of our first two parts of our theorem will not rely on the condition $\theta \in [0, \pi/2]$, which is only needed for the third part of the theorem.

$\|m\| = 1$.
With these preliminaries out of the way, we begin our proof.

6.2 Statement of Theorem and Proofs

Theorem 9 (Flip Probability) *Let $n,\ m \in \mathbb{R}^d$ and let the angle between them be $\theta \in [0, \pi/2]$. Without loss of generality take $\|n\| = \|m\| = 1$.*
Let $R \in \mathcal{M}_{k\times d}$, $k < d$, be a random projection matrix with entries $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ and let $Rn,\ Rm \in \mathbb{R}^k$ be the images of $n,\ m$ with angular separation θ_R.

1. *Denote by $f_k(\theta)$ the 'flip probability' $f_k(\theta) := Pr[\theta_R > \pi/2] = Pr[(Rn)^T Rm < 0]$. Then:*
$$f_k(\theta) = \frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\psi} \frac{z^{(k-2)/2}}{(1+z)^k} dz \tag{6.2.1}$$
where $\psi = (1 - \cos(\theta))/(1 + \cos(\theta))$.

2. *The expression above can be shown to be of the form of the quotient of the surface area of a hyperspherical cap subtending an angle of 2θ by the surface area of the corresponding hypersphere:*
$$f_k(\theta) = \frac{\int_0^{\theta} \sin^{k-1}(\phi)\ d\phi}{\int_0^{\pi} \sin^{k-1}(\phi)\ d\phi} \tag{6.2.2}$$
This form recovers Lemma 3.2 of (Goemans & Williamson, 1995) where the flip probability θ/π for $k = 1$ was given, and extends it for $k \geqslant 1$ showing that the flip probability is polynomial of order k in θ.

3. *The flip probability is monotonic decreasing as a function of k: Fix $\theta \in [0, \pi/2]$, then $f_k(\theta) \geqslant f_{k+1}(\theta)$.*

6.3 Proof of Theorem 9

Proof of part 1.

First we expand out the terms of $(Rn)^T Rm$:

$$\mathrm{Pr}_R[(Rn)^T Rm < 0] = \mathrm{Pr}_R\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} r_{ij}m_j\right)\left(\sum_{j=1}^{d} r_{ij}n_j\right) < 0\right] \tag{6.3.1}$$

Recall that the entries of R are independent and identically distributed with $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ and make the change of variables $u_i = \sum_{j=1}^{d} r_{ij}m_j$ and $v_i = \sum_{j=1}^{d} r_{ij}n_j$. A linear combination of Gaussian variables is again Gaussian, however u_i and v_i are now no longer independent since they both depend on the same row of R. On the other hand, for $i \neq j$ the vectors (u_i, v_i) and (u_j, v_j) are independent of each other since the

i-th row of R is independent of its j-th row. Moreover $(u_i, v_i) \sim (u_j, v_j)$, $\forall i, j$ so it is enough to consider a single term of the outer sum in (6.3.1). We have:

$$\left(\begin{array}{c} u_i \\ v_i \end{array} \right) \sim \mathcal{N} \left(\mathrm{E}_R \left[\left(\begin{array}{c} u_i \\ v_i \end{array} \right) \right], \mathrm{Cov}_R \left[\left(\begin{array}{c} u_i \\ v_i \end{array} \right) \right] \right)$$

Since u_i and v_i are zero mean, the expectation of this distribution is just $(0,0)^T$, and its covariance is:

$$\Sigma_{u,v} = \left[\begin{array}{cc} \mathrm{Var}(u_i) & \mathrm{Cov}(u_i, v_i) \\ \mathrm{Cov}(u_i, v_i) & \mathrm{Var}(v_i) \end{array} \right] \tag{6.3.2}$$

Then:

$$\begin{aligned} \mathrm{Var}(u_i) &= \mathrm{E}[(u_i - \mathrm{E}(u_i))^2] \\ &= \mathrm{E}[(u_i)^2] \text{ since } \mathrm{E}(u_i) = 0 \\ &= \mathrm{E}\left[\left(\sum_{j=1}^{d} r_{ij} n_j \right)^2 \right] \\ &= \mathrm{E}\left[\sum_{\substack{j=1 \\ j'=1}}^{d} r_{ij} r_{ij'} n_j n_{j'} \right] \\ &= \sum_{\substack{j=1 \\ j'=1}}^{d} n_j n_{j'} \mathrm{E}\left[r_{ij} r_{ij'} \right] \end{aligned}$$

Now, when $j \neq j'$, r_{ij} and $r_{ij'}$ are independent, and so $\mathrm{E}[r_{ij} r_{ij'}] = \mathrm{E}[r_{ij}]\mathrm{E}[r_{ij'}] = 0$. On the other hand, when $j = j'$ we have $\mathrm{E}[r_{ij} r_{ij'}] = \mathrm{E}[r_{ij}^2] = \mathrm{Var}(r_{ij}) = \sigma^2$, since $r_{ij} \sim \mathcal{N}(0, \sigma^2)$. Hence:

$$\mathrm{Var}(u_i) = \sum_{j=1}^{d} \sigma^2 n_j^2 = \sigma^2 \|n\|^2 = \sigma^2 \tag{6.3.3}$$

since $\|n\| = 1$. A similar argument gives $\mathrm{Var}(v_i) = \sigma^2$.
Now we want to find the covariance $\mathrm{Cov}(u_i, v_i)$:

$$\begin{aligned} \mathrm{Cov}(u_i, v_i) &= \mathrm{E}\left[(u_i - \mathrm{E}[u_i]) (v_i - \mathrm{E}[v_i]) \right] = \mathrm{E}[u_i v_i] \\ &= \mathrm{E}\left[\left(\sum_{j=1}^{d} r_{ij} n_j \right) \left(\sum_{j=1}^{d} r_{ij} m_j \right) \right] \\ &= \sum_{\substack{j=1 \\ j'=1}}^{d} n_j m_{j'} \mathrm{E}[r_{ij} r_{ij'}] \end{aligned} \tag{6.3.4}$$

Now, when $j \neq j'$ the expectation is zero, as before, and similarly when $j = j'$ we have for (6.3.4):

$$= \sum_{j=1}^{d} n_j m_j \mathrm{E}[(r_{ij})^2] = \sum_{j=1}^{d} n_j m_j \mathrm{Var}(r_{ij}) = \sigma^2 n^T m \tag{6.3.5}$$

Hence for each $i \in \{1, \ldots, k\}$ the covariance matrix is:

$$\Sigma_{u,v} = \sigma^2 \begin{bmatrix} 1 & n^T m \\ n^T m & 1 \end{bmatrix}$$

and so, $(u_i, v_i)^T \overset{\text{i.i.d}}{\sim} (0, \Sigma_{u,v})$. Now we can rewrite the probability in (6.3.1) as:

$$\Pr\left\{ \sum_{i=1}^{k} u_i v_i < 0 \right\} \tag{6.3.6}$$

Next, it will be useful to use the identity $u_i v_i = (u_i, v_i)\left(\begin{smallmatrix} 0 & 1/2 \\ 1/2 & 0 \end{smallmatrix}\right)(u_i, v_i)^T$ and rewrite the probability (6.3.6) as:

$$\Pr\left\{ \sum_{i=1}^{k} (u_i, v_i) \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \begin{pmatrix} u_i \\ v_i \end{pmatrix} < 0 \right\} \tag{6.3.7}$$

where the probability is now over the distribution of $(u_i, v_i)^T$. We will make a further change of variables and write:

$$(x_i, y_i)^T = \Sigma_{u,v}^{-1/2} (u_i, v_i)^T \tag{6.3.8}$$

so that the new variables x_i, y_i are independent unit variance spherical Gaussian variables, $(x_i, y_i)^T \overset{iid}{\sim} \mathcal{N}(0, I)$. Substituting back into (6.3.7) the probability we want to find is then:

$$\Pr\left\{ \frac{1}{2} \sum_{i=1}^{k} (x_i, y_i)\, \Sigma_{u,v}^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \Sigma_{u,v}^{1/2} \begin{pmatrix} x_i \\ y_i \end{pmatrix} < 0 \right\} \tag{6.3.9}$$

where the probability now is w.r.t the standard Gaussian distribution. Now diagonalizing the positive definite matrix $\Sigma_{u,v}^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \Sigma_{u,v}^{1/2}$ as $U \Lambda U^T$ with $UU^T = U^T U = I$ and Λ a diagonal matrix of its (necessarily positive) eigenvalues we rewrite (6.3.9) as:

$$\Pr\left\{ \frac{1}{2} \sum_{i=1}^{k} (x_i, y_i) U \Lambda U^T \begin{pmatrix} x_i \\ y_i \end{pmatrix} < 0 \right\} \tag{6.3.10}$$

and note that, as the standard Gaussian distribution is invariant under orthogonal transformations, the form of U does not affect this probability. Therefore without loss of generality we can take it to be the identity matrix to rewrite (6.3.10) as:

$$\Pr\left\{\frac{1}{2}\sum_{i=1}^{k}(x_i,y_i)\Lambda\left(\begin{array}{c}x_i\\ y_i\end{array}\right)<0\right\}$$

Now we need the entries of Λ. These are the eigenvalues of:

$$\Sigma_{u,v}^{1/2}\left[\begin{array}{cc}0 & 1\\ 1 & 0\end{array}\right]\Sigma_{u,v}^{1/2}$$

and using the fact that the eigenvalues of AB are the same as the eigenvalues of BA (Horn & Johnson, 1985) these are the eigenvalues of

$$\sigma^2\left[\begin{array}{cc}1 & n^Tm\\ n^Tm & 1\end{array}\right]\left[\begin{array}{cc}0 & 1\\ 1 & 0\end{array}\right]=\sigma^2\left[\begin{array}{cc}n^Tm & 1\\ 1 & n^Tm\end{array}\right]=\sigma^2\left[\begin{array}{cc}\cos(\theta) & 1\\ 1 & \cos(\theta)\end{array}\right]$$

which are $\lambda=\sigma^2(\cos(\theta)\pm 1)$.

Substituting this back into the inequality (6.3.10) we may drop the positive scaling constant $\frac{1}{2}\sigma^2$ since it does not affect the sign of the left hand side, and so the probability to find can now be written as:

$$\begin{aligned}
&\Pr\left\{\sum_{i=1}^{k}(x_i,y_i)^T\left[\begin{array}{cc}\cos(\theta)+1 & 0\\ 0 & \cos(\theta)-1\end{array}\right]\left(\begin{array}{c}x_i\\ y_i\end{array}\right)<0\right\}\\
&=\Pr\left\{\sum_{i=1}^{k}((\cos(\theta)+1)x_i^2+(\cos(\theta)-1)y_i^2)<0\right\}\\
&=\Pr\left\{(\cos(\theta)+1)\sum_{i=1}^{k}x_i^2+(\cos(\theta)-1)\sum_{i=1}^{k}y_i^2<0\right\}\\
&=\Pr\left\{\frac{(\cos(\theta)+1)\sum_{i=1}^{k}x_i^2}{(\cos(\theta)-1)\sum_{i=1}^{k}y_i^2}+1<0\right\}\\
&=\Pr\left\{\frac{\sum_{i=1}^{k}x_i^2}{\sum_{i=1}^{k}y_i^2}<\frac{1-\cos(\theta)}{1+\cos(\theta)}\right\}
\end{aligned}\tag{6.3.11}$$

Now, x_i and y_i are standard univariate Gaussian variables, hence $x_i^2, y_i^2 \overset{iid}{\sim} \chi^2$, and so the left hand side of (6.3.11) is F-distributed ((Mardia et al., 1979), Appendix B.4, pg 487). Therefore:

$$\Pr_R[(Rn)^TRm<0]=\frac{\Gamma(k)}{(\Gamma(k/2))^2}\int_0^{\psi}\frac{z^{(k-2)/2}}{(1+z)^k}\mathrm{d}z$$

where $\psi=(1-\cos(\theta))/(1+\cos(\theta))$ and $\Gamma(\cdot)$ is the Gamma function. This proves the first part of Theorem 9. □

6.4 Proof of part 2.

Note that $\psi = \tan^2(\theta/2)$ and make the substitution $z = \tan^2(\theta/2)$. Then, via the trigonometric identity $\sin(\theta) = 2\tan(\theta)/(1+\tan^2(\theta))$ and $\frac{\mathrm{d}z}{\mathrm{d}\theta} = \tan(\theta/2)(1+\tan^2(\theta/2))$, we obtain:

$$f_k(\theta) = \frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2} \int_0^\theta \sin^{k-1}(\phi)\mathrm{d}\phi \tag{6.4.1}$$

To put the expression (6.4.1) in the form of the second part of the theorem, we need to show that the gamma term outside the integral is the reciprocal of $\int_0^\pi \sin^{k-1}(\phi)\mathrm{d}\phi$. This can be shown in a straightforward way using the beta function.

Recall that the beta function is defined by (e.g. (Abramowitz & Stegun, 1972), 6.2.2, pg 258):

$$\mathrm{B}(w,z) = \frac{\Gamma(w)\Gamma(z)}{\Gamma(w+z)} = 2\int_0^{\pi/2} \sin^{2w-1}(\theta)\cos^{2z-1}(\theta)\mathrm{d}\theta, \quad \mathrm{Re}(w), \mathrm{Re}(z) > 0 \tag{6.4.2}$$

Then from (6.4.2) we have:

$$\frac{1}{2}\mathrm{B}\left(\frac{k}{2}, \frac{1}{2}\right) = \int_0^{\pi/2} \sin^{k-1}(\theta)\mathrm{d}\theta$$

and from the symmetry of the sine function about $\pi/2$, equation (6.4.2), and using $\Gamma(1/2) = \sqrt{\pi}$ we have:

$$\int_0^{\pi} \sin^{k-1}(\theta)\mathrm{d}\theta = 2\int_0^{\pi/2} \sin^{k-1}(\theta)\mathrm{d}\theta = \mathrm{B}\left(\frac{k}{2}, \frac{1}{2}\right) = \frac{\sqrt{\pi}\,\Gamma(k/2)}{\Gamma((k+1)/2)}$$

Now we just need to show that the left hand side of (6.4.1):

$$\frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2} = \frac{\Gamma((k+1)/2)}{\sqrt{\pi}\,\Gamma(k/2)} \tag{6.4.3}$$

To do this we use the duplication formula ((Abramowitz & Stegun, 1972), 6.1.18, pg 256):

$$\Gamma(2z) = (2\pi)^{-\frac{1}{2}} 2^{2z-\frac{1}{2}} \Gamma(z)\Gamma((2z+1)/2)$$

with $z = k/2$.
Then the left hand side of (6.4.3) is equal to:

$$\frac{2^{k-\frac{1}{2}}\Gamma(k/2)\Gamma((k+1)/2)}{\sqrt{2\pi}\;2^{k-1}(\Gamma(k/2))^2} = \frac{\Gamma(k/2)\Gamma((k+1)/2)}{\sqrt{\pi}\;(\Gamma(k/2))^2} = \frac{\Gamma((k+1)/2)}{\sqrt{\pi}\;\Gamma(k/2)}$$

as required. Putting everything together, we arrive at the alternative form for (6.4.1) given in equation (6.2.2), namely:

$$\Pr_R[(Rn)^T Rm < 0] = \frac{\int_0^\theta \sin^{k-1}(\phi)\ \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\ \mathrm{d}\phi} \tag{6.4.4}$$

This proves the second part of Theorem 9. □

We will give a geometric interpretation of this form of the flip probability in the next section.

Remark. It is easy to verify that (6.4.4) recovers the known result for $k = 1$, namely θ/π, as given in (Goemans & Williamson, 1995) (Lemma 3.2). Further, for general k, the expression (6.4.4) is polynomial of order k in θ. This can be seen by using the fact that $\sin(\phi) \leqslant \phi$, which gives us the upper bound $\Pr_R[(Rn)^T Rm < 0] \leqslant \frac{\theta^k}{k \cdot \int_0^\pi \sin^{k-1}(\phi)\, \mathrm{d}\phi}$.

6.5 Proof of part 3.

Finally, we prove that the flip probability is monotonic decreasing in the projection dimension k. Note that although the value of the expressions in (6.2.2) and (6.2.1) can be calculated exactly for any given k and θ (e.g. using integration by parts) there is no general closed form for either the integral or the gamma term and, as k grows, this becomes increasingly inconvenient. The final part of the theorem, bounding the flip probability in the $(k+1)$-dimensional case above by the flip probability in the k-dimensional case, is therefore useful in practice.

To prove the final part of the theorem we will show that for all $\theta \in [0, \pi/2]$, the ratio of successive flip probabilities:

$$\frac{f_{k+1}(\theta)}{f_k(\theta)} = \frac{\left(\frac{\int_0^\theta \sin^k(\phi)\, \mathrm{d}\phi}{\int_0^\pi \sin^k(\phi)\, \mathrm{d}\phi}\right)}{\left(\frac{\int_0^\theta \sin^{k-1}(\phi)\, \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\, \mathrm{d}\phi}\right)} \leqslant 1 \tag{6.5.1}$$

which is sufficient.

Let us rewrite the ratio (6.5.1) above as:

$$\frac{\left(\frac{\int_0^\theta \sin^k(\phi)\, \mathrm{d}\phi}{\int_0^\theta \sin^{k-1}(\phi)\, \mathrm{d}\phi}\right)}{\left(\frac{\int_0^\pi \sin^k(\phi)\, \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\, \mathrm{d}\phi}\right)} \tag{6.5.2}$$

Call the numerator of (6.5.2) $g_k(\theta)$, and notice that the denominator is nothing but $g_k(\pi)$. Now observe that the denominator, $g_k(\pi) = g_k(\pi/2)$. Since:

$$\frac{\int_0^\pi \sin^k(\phi)\, \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\, \mathrm{d}\phi} = \frac{2\int_0^{\pi/2} \sin^k(\phi)\, \mathrm{d}\phi}{2\int_0^{\pi/2} \sin^{k-1}(\phi)\, \mathrm{d}\phi} = \frac{\int_0^{\pi/2} \sin^k(\phi)\, \mathrm{d}\phi}{\int_0^{\pi/2} \sin^{k-1}(\phi)\, \mathrm{d}\phi}$$

where the first equality follows from the symmetry of the sine function about $\pi/2$. Thus we see that the whole expression (6.5.2) is equal to 1 when $\theta = \pi/2$. It remains now to show that the $g_k(\theta) \leqslant g_k(\pi/2), \forall \theta \in [0, \pi/2]$ and $k \in \mathcal{N}$. In fact more is true: We show that $\forall k, g_k(\theta)$ is monotonic increasing as a function of θ on $[0, \pi/2]$. From this the required inequality follows, and therefore the expression (6.5.2) has its maximum

value of 1 on this domain, and the result follows.
To show monotonicity, we differentiate the function $g_k(\theta)$ with respect to θ to obtain:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} g_k(\theta) = \frac{\sin^k(\theta)\int_0^\theta \sin^{k-1}(\phi)\mathrm{d}\phi - \sin^{k-1}(\theta)\int_0^\theta \sin^k(\phi)\mathrm{d}\phi}{\left(\int_0^\theta \sin^{k-1}(\phi)\mathrm{d}\phi\right)^2} \tag{6.5.3}$$

Then (6.5.3) is greater than zero when its numerator is, and:

$$\begin{aligned}
&\sin^k(\theta)\int_0^\theta \sin^{k-1}(\phi)\mathrm{d}\phi - \sin^{k-1}(\theta)\int_0^\theta \sin^k(\phi)\mathrm{d}\phi \\
&= \sin^{k-1}(\theta)\left[\sin(\theta)\int_0^\theta \sin^{k-1}(\phi)\mathrm{d}\phi - \int_0^\theta \sin^k(\phi)\mathrm{d}\phi\right] \\
&= \sin^{k-1}(\theta)\left[\int_0^\theta \sin(\theta)\sin^{k-1}(\phi)\mathrm{d}\phi - \int_0^\theta \sin(\phi)\sin^{k-1}(\phi)\mathrm{d}\phi\right] \geqslant 0
\end{aligned}$$

Where the last step follows from monotonicity of the sine function on $[0, \pi/2]$ and so $\sin(\theta) \geqslant \sin(\phi)$ for $\theta \geqslant \phi > 0$, $\theta \in [0, \pi/2]$. It follows now that the numerator of (6.5.2) is monotonic increasing with $\theta \in [0, \pi/2]$ and so the whole expression (6.5.1) takes its maximum value of 1 when $\theta = \pi/2$. This proves the final part of Theorem 9 and completes the proofs. ■

Remark. We note that for $\theta \in [\pi/2, \pi]$ it is easy to show, using the symmetry of sine about $\pi/2$, that the sense of the inequality in part 3 of the theorem is reversed. Then: $f_{k+1}(\theta) \geqslant f_k(\theta)$, $\theta \in [\pi/2, \pi]$.

6.6 Geometric Interpretation

In the case $k = 1$, the flip probability θ/π (given also in (Goemans & Williamson, 1995)) is the quotient of the length of the arc subtending an angle of 2θ by the circumference of the unit circle 2π. In the form of (6.4.4) our result generalizes this geometric interpretation in a natural way, as follows. Recall that the surface area of the unit hypersphere in $\mathbb{R}^{k+1}$ is given by (Kendall, 2004):

$$2\pi \cdot \prod_{i=1}^{k-1} \int_0^\pi \sin^i(\phi)\mathrm{d}\phi$$

(which is is also $(k+1)/r$ times the volume of the same hypersphere.) This expression is the extension to $\mathbb{R}^{k+1}$ of the standard 'integrating slabs' approach to finding the volume of the 3-dimensional sphere S^2, and so the surface area of the hyperspherical cap subtending angle 2θ is simply:

$$2\pi \cdot \prod_{i=1}^{k-2} \int_0^\pi \sin^i(\phi)\mathrm{d}\phi \cdot \int_0^\theta \sin^{k-1}(\phi)\mathrm{d}\phi$$

If we now take the quotient of these two areas all but the last terms cancel and we obtain:

$$\frac{2\pi \cdot \prod_{i=1}^{k-2} \int_0^\pi \sin^i(\phi) \mathrm{d}\phi \cdot \int_0^\theta \sin^{k-1}(\phi) \mathrm{d}\phi}{2\pi \cdot \prod_{i=1}^{k-1} \int_0^\pi \sin^i(\phi) \mathrm{d}\phi} = \frac{\int_0^\theta \sin^{k-1}(\phi)\ \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\ \mathrm{d}\phi}$$

which is exactly our flip probability as given in (6.4.4).

Hence, the probability that a dot product flips from being positive to being negative (equivalently the angle flips from acute to obtuse) after Gaussian random transformation is given by the ratio of the surface area in $\mathbb{R}^{k+1}$ of a hyperspherical cap subtending an angle of 2θ to the surface area of the unit hypersphere. Note particularly the surprising fact that therefore the flip probability depends only on the angular separation of the two vectors and on the projection dimensionality k: It is independent of the embedding dimensionality d. Moreover our geometric interpretation shows that equation (6.4.4) decays exponentially with increasing k as the proportion of the surface of the k-dimensional unit sphere covered by a spherical cap subtending an angle of 2θ is known to be bounded above by $\exp\left(-\frac{1}{2}k\cos^2(\theta)\right)$ ((Ball, 1997), Lemma 2.2, Pg 11).

6.7 Empirical Validation

Our results seem quite counterintuitive, especially the fact that the flip probability is independent of the embedding dimensionality d. To confirm our theoretical findings we ran Monte Carlo trials to estimate the flip probability as follows: We let $d \in \{50, 100, \ldots, 500\}$, $k \in \{1, 5, 10, 15, 20, 25\}$ and $\theta \in \{0, \pi/128, \ldots, t \cdot \pi/128, \ldots, \pi/2\}$. For each (d, θ) tuple we generated 2 randomly oriented d-dimensional θ-separated unit length vectors m, n. For each (k, d, θ) tuple, we generated 5000 $k \times d$ random projection matrices R with which we randomly projected m and n. Finally we counted the number of times, N, that the dot product $(R(m))^T R(n) < 0$ and estimated the flip probability by $N/5000$.

We give plots of the results: Figure 6.1 shows the close match between our theoretical values and empirical estimates of the flip probabilities, while figure 6.2 gives empirical validation of the fact that the flip probability is independent of d. We note that, for non-spherical $\hat{\Sigma}$ empirical trials show that the flip probability can be significantly higher than in the spherical case and also less well-behaved – see figure 6.3. The following upper bound (6.7.1) on the flip probability in the non-spherical case, which is tight since in the spherical case (i.e. when we estimate $R\hat{\Sigma}R^T = I$ in the projected space) it recovers our theorem 9, indicates why this is the case:

$$\begin{aligned}
&\Pr_R\left[\left((R\hat{\Sigma}R^T)^{-1/2}Rn\right)^T((R\hat{\Sigma}R^T)^{-1/2}Rm) < 0\right] \\
= \;&\Pr_R\left[\frac{1}{2}\left(\|(R\hat{\Sigma}R^T)^{-1/2}Rn\|^2 + \|(R\hat{\Sigma}R^T)^{-1/2}Rm\|^2 - \|(R\hat{\Sigma}R^T)^{-1/2}(Rn-Rm)\|^2\right) < 0\right] \\
\leqslant \;&\Pr_R\left[\frac{1}{2}\left(\lambda_{\min}((R\hat{\Sigma}R^T)^{-1})\left(\|Rn\|^2 + \|Rm\|^2\right) - \lambda_{\max}((R\hat{\Sigma}R^T)^{-1})\|Rn-Rm\|^2\right) < 0\right] \\
= \;&\Pr_R\left[\frac{1}{2}\left(\left(\|Rn\|^2 + \|Rm\|^2\right) - \kappa((R\hat{\Sigma}R^T)^{-1})\|Rn-Rm\|^2\right) < 0\right] \\
= \;&\Pr_R\left[\frac{1}{2}\left(\left(\|Rn\|^2 + \|Rm\|^2\right) - \kappa(R\hat{\Sigma}R^T)\|Rn-Rm\|^2\right) < 0\right] \\
\leqslant \;&\Pr_R\left[\frac{1}{2}\left(\left(\|Rn\|^2 + \|Rm\|^2\right) - \kappa(\hat{\Sigma})\|Rn-Rm\|^2\right) < 0\right] \qquad (6.7.1)
\end{aligned}$$

where the step from the first to the second line is the parallelogram law, the second last step uses the identity $\kappa(A) = \kappa(A^{-1})$ for invertible matrices A, and the last step follows from our findings in chapter 5 that the covariance estimated in the randomly-projected space is better conditioned than its data space counterpart. Note that we also recover our original theorem when $k = 1$, since then $\kappa(R\hat{\Sigma}R^T) = 1$.

6.8 Summary and Discussion

We derived the exact probability of 'label flipping' as a result of random projection, for the case where Σ is estimated by a spherical covariance matrix $\hat{\Sigma} = \alpha I$, and proved a simple, yet tight, upper bound on this probability for the general setting when $\hat{\Sigma}$ is allowed to be non-spherical. The inequality 6.7.1 agrees with theoretical and empirical findings of (Dasgupta, 1999; 2000a) in particular when d is large, even if $\hat{\Sigma}$ is very eccentric, it can be that $\kappa(R\hat{\Sigma}R^T) \ll \kappa(\hat{\Sigma})$ where the difference in the two condition numbers may be several orders of magnitude. In practice therefore it seems that for non-spherical $\hat{\Sigma}$ there is likely to be a trade off between reducing k which reduces $\kappa(R\hat{\Sigma}R^T)$ and increasing k which makes the flip probability smaller in the spherical case and, therefore, presumably also does the same in the non-spherical case. Testing this intuition by better quantifying the flip probability for non-spherical $\hat{\Sigma}$ remains for future research.

We also note that, via existing VC-type bounds, our flip probability implies an upper bound on the $(0, 1)$-generalization error of any linear classifier trained by empirical risk minimization (ERM) in a randomly projected space. More specifically, the empirical risk minimizer learned in the projected space has training error no greater than the projection of the ERM classifier learned in the data space, and the training error of the projection of the data space ERM classifier is simply the training error of the data space ERM classifier plus the flip probability. The same observation was made, for

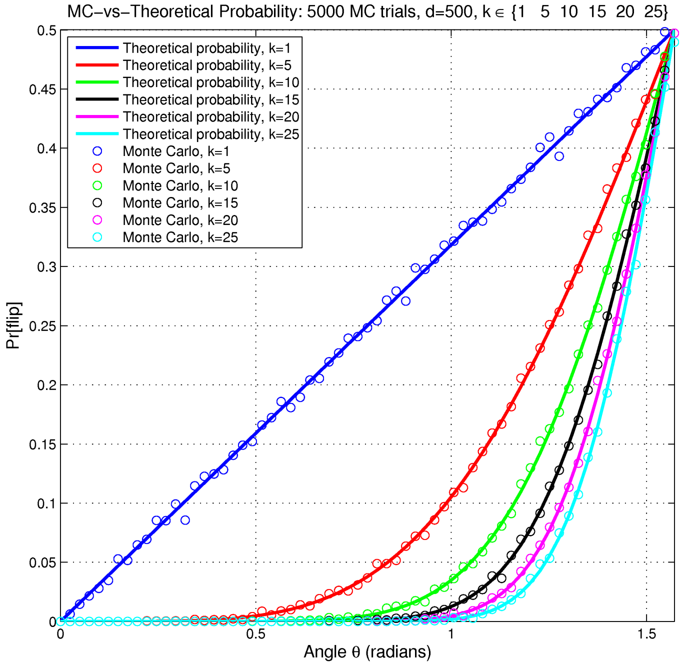

Figure 6.1: Experiment illustrating match between probability calculated in theorem 9 and empirical trials. We fixed $d = 500$ and allowed k to vary. For each trial we generated two random unit-length vectors in $\mathbb{R}^{500}$ with angular separation θ and for each (θ, k) pair we randomly projected them with 5000 different random projection matrices to estimate the empirical flip probability. Circles show the empirical flip probabilities, lines show the theoretical flip probability.

different reasons, by Garg et. al. in (Garg & Roth, 2003; Garg et al., 2002): In those papers however the authors' motivation was to provide bounds on the generalization error of the *data space* classifier, and they obtained only a loose upper bound on the flip probability.

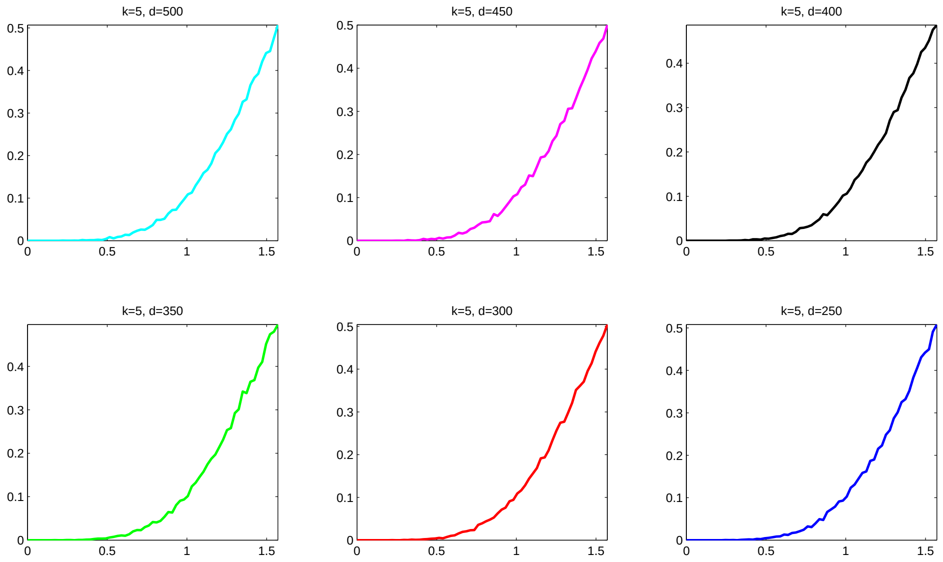

Figure 6.2: Experiment illustrating d-invariance of the flip probability of theorem 9. We fixed $k = 5$ and allowed d to vary, estimating the empirical flip probability with 5000 different random projections from $\mathbb{R}^d$ into $\mathbb{R}^5$ for each (θ, d) pair. The results for each of six choices of d are plotted on separate graphs, highlighting the similarity of the outcomes.

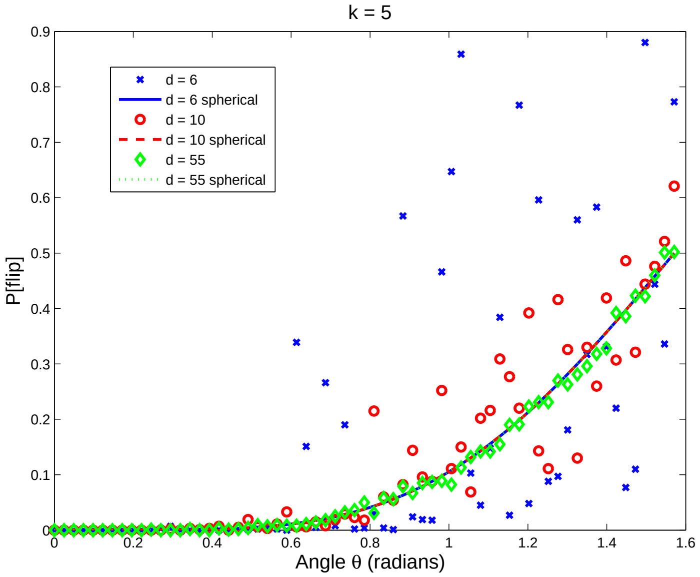

Figure 6.3: Experiment illustrating irregularity in flip probability when the two vectors are drawn from an eccentric Gaussian. We fixed $k = 5$ and allowed d to vary. For each set of trials we randomly generated an eccentric Gaussian and estimated the flip probability using Monte Carlo trials as before.

7
Voting Ensembles of RP-FLDs

Summary In this chapter we derive theory for, and empirically evaluate the performance of, an ensemble of randomly projected Fisher Linear Discriminant classifiers. We focus on the case when there are fewer training observations than data dimensions, which is a common situation in a range of problem domains such as radiology, biomedical imaging, and genomics (Candes & Tao, 2007). Our ensemble is learned from a sequence of randomly projected representations of the original high dimensional data and so, for this approach, data could be collected, stored and processed in such a compressed form.

The specific form and simplicity of the ensemble we consider permits a direct and much more detailed analysis than can be obtained using existing generic tools applied in previous works. In particular, we are able to derive the exact form of the generalization error of our ensemble, conditional on the training set, and based on this we give theoretical guarantees which directly link the performance of the ensemble to that of the corresponding FLD classifier learned in the full data space. Furthermore we show that our randomly projected ensemble implicitly implements a sophisticated regularization scheme to FLD learned in the original data space and this prevents overfitting in conditions of small sample size where pseudoinverted FLD learned in the data space is provably poor. To the best of our knowledge these are the first theoretical results to provide such an explicit link between any classifier and classifier ensemble pair.

We confirm the utility of our ensemble approach with a range of experiments on real datasets from the bioinformatics domain; for these data the number of observations in each dataset is orders of magnitude smaller than the data dimensionality, yet our ensemble still achieves performance comparable with the state-of-the-art. Moreover desirable properties of our ensemble approach include (i) fitting the regularization parameter is straightforward, and (ii) a well-performing ensemble can be constructed for generally lower computational cost than working in the full data space.

7.1 Preliminaries

As in the previous chapters of this thesis, we consider a binary classification problem in which we observe N i.i.d examples of labelled training data $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$ where the $x_i \in \mathbb{R}^d$ are d-dimensional real valued observations. Unlike the previous chapters of this thesis, here we assume that $N \ll d$, and we are interested in comparing the performance of an ensemble of RP-FLD classifiers working in the projected space $\mathbb{R}^k$, $k < d$, to the performance achievable by the corresponding FLD classifier working in the data space $\mathbb{R}^d$.
Recall that the decision rule for FLD learned from training data is given by:

$$\hat{h}(x_q) := \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\}$$

where $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\Sigma}$ are maximum likelihood (ML) estimates of the class-conditional means and (shared) covariance matrix respectively, and $\mathbf{1}(\cdot)$ is the indicator function which returns 1 if its argument is true and 0 otherwise.

In the setting considered here we assume that $N \ll d$ and therefore $\hat{\Sigma}$ will be singular. In order to obtain a working decision rule one can either pseudo-invert or regularize $\hat{\Sigma}$; both approaches are used in practice (Raudys & Duin, 1998).
To construct the randomly projected ensemble, we choose the number of ensemble members M and the projection dimensionality k, and generate M random matrices $R \in \mathcal{M}_{k\times d}$ with i.i.d entries $r_{ij} \sim \mathcal{N}(0, \sigma^2)$, and we take $\sigma^2 = 1$ without loss of generality.[1].
Pre-multiplying the data with one of the matrices R maps the training examples to a k-dimensional subspace of the data space $\mathbb{R}^d$ and, by linearity of expectation and of the projection operator, the decision rule for a single randomly projected classifier is then given by:

$$\hat{h}_R(x_q) := \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\}$$

For an ensemble, various different combination rules can be applied. The most common choices include majority voting (when there is an odd number of classifiers in the ensemble) and linear or convex combination (e.g. Brown, 2009). We may interpret the magnitude of the output of an individual RP-FLD classifier as a measure of the confidence in that classifier's decision, and we want to make use of these confidence estimates. For this reason we choose to employ the averaged linear decisions of M base learners – which gives the following ensemble decision function:

$$\hat{h}_{ens}(x_q) := \mathbf{1}\left\{\frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i\left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\}$$

[1] We find empirically that, as one would expect, other common choices of random projection matrix with zero-mean i.i.d subgaussian entries (e.g. Achlioptas, 2003) do not affect the ensemble performance.

This decision rule is called 'voting' in the ensemble literature[2]. Unlike majority voting, this choice of decision rule does not require the number of classifiers in the ensemble to be odd for good generalization and, as we shall see, it also has the advantage of analytical tractability.

7.2 Theory: Linking the ensemble error to the full data space error

We are interested in the generalization error of our voting ensemble in the setting where $\text{rank}(\Sigma) = d$, but $\text{rank}(\hat{\Sigma}) = \rho \ll d$ and, especially, in linking the performance of this ensemble to the corresponding data space FLD. We will start by examining the expected performance of the RP-FLD ensemble when the training set is fixed, which is central to linking the ensemble and data space classifiers, and then later in Theorem 10 we will consider random instantiations of the training set.

To begin, observe that by the law of large numbers the LHS of the argument of the decision rule of our ensemble converges to the following:

$$\begin{aligned} & \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} (\hat{\mu}_{\neg y} - \hat{\mu}_y)^T R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i (\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y) \\ = \quad & (\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \text{E}\left[R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\right] (\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y) \qquad (7.2.1) \end{aligned}$$

provided that this limit exists. It will turn out that for $R \in \mathcal{M}_{k\times d}$ having i.i.d zero-mean Gaussian entries $r_{ij} \sim \mathcal{N}(0,1)$, if $k \in \{1, ..., \rho-2\} \cup \{\rho+2, ..., d\}$, then this expectation is indeed defined for each entry. From equation (7.2.1) we see that, for a fixed training set, in order to quantify the error of the ensemble it is enough to consider the expectation (w.r.t random matrices R):

$$\text{E}\left[R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\right] \qquad (7.2.2)$$

To smooth our way, we will emulate Marzetta et al. (2011) and make use of the following two lemmas in the next subsections:

Lemma 19 (Orthogonal invariance)
Let $R \in \mathcal{M}_{k\times d}$ with $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$. Let $\hat{\Sigma}$ be any symmetric positive semi-definite matrix, and let $\hat{U}$ be an orthogonal matrix such that $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$, where $\hat{\Lambda}$ is a diagonal matrix with the eigenvalues of $\hat{\Sigma}$ in descending order along the diagonal. Then:

$$E\left[R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\right] = \hat{U} E\left[R^T \left(R\hat{\Lambda}R^T\right)^{-1} R\right] \hat{U}^T$$

[2] Voting is distinct from majority voting: In majority voting one considers only the sign, and not the magnitude, of the decisions of the individual classifiers comprising the ensemble. In voting both the sign and magnitude of the output from each individual ensemble member is taken into account.

Proof 9

(of Lemma 19) We use the facts that $\hat{U}\hat{U}^T = \hat{U}^T\hat{U} = I$ and that the rows of R are drawn from a spherical Gaussian, and therefore their distribution is invariant under orthogonal transformations. Hence:

$$\begin{aligned}
& E\left[R^T\left(R\hat{U}\hat{\Lambda}\hat{U}^T R^T\right)^{-1} R\right] \\
= \ & E\left[\hat{U}\hat{U}^T R^T\left(R\hat{U}\hat{\Lambda}\hat{U}^T R^T\right)^{-1} R\hat{U}\hat{U}^T\right] \\
= \ & \hat{U}E_{\tilde{R}}\left[\tilde{R}^T\left(\tilde{R}\hat{\Lambda}\tilde{R}^T\right)^{-1}\tilde{R}\right]\hat{U}^T \\
= \ & \hat{U}E\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1} R\right]\hat{U}^T
\end{aligned}$$

since $\tilde{R} = R\hat{U} \sim R$

Next the easily proved fact that if A is a square matrix then A is diagonal if and only if $VAV^T = A$ for all diagonal orthogonal matrices $V = \text{diag}(\pm 1)$ yields the following useful lemma:

Lemma 20 (Expected preservation of eigenvectors)
Let $\hat{\Lambda}$ be a diagonal matrix, then $E\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1} R\right]$ is a diagonal matrix.

Furthermore, if $\hat{U}$ diagonalizes $\hat{\Sigma}$ as $\hat{U}\hat{\Lambda}\hat{U}^T$, then $\hat{U}$ also diagonalizes $E\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1} R\right]$.

We omit the proof of Lemma 20, which is very similar to that for lemma 19 using the facts that V is orthogonal and that if $\hat{U}V$ diagonalizes $\hat{\Sigma}$ then so does $\hat{U}$. It follows from lemmas 19 and 20 that at convergence our ensemble preserves the eigenvectors of $\hat{\Sigma}$, and so we only need to consider the diagonal entries (i.e. the eigenvalues) of $\text{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1} R\right]$, which we do in the next subsection.

Before starting our analysis we should note that, for the cases $k \in \{1, ..., \rho - 2\}$ and taking R to be a random projection matrix with the rows orthonormalized, Marzetta et al. (2011) provide a (rather complicated) procedure to compute this expectation exactly. Instead, in our context, we are more interested in how this expectation relates to characteristics of $\hat{\Sigma}$. In particular, we are interested in how the ensemble reduces the ill-conditioning of this matrix since we shall see in Section 7.3 that this has a direct impact on the generalization error of the FLD classifier. We answer this question by bounding this expectation from both sides in the positive semi-definite ordering. Furthermore, Marzetta et al. (2011) did not consider the case $k > \rho + 1$, when the expectation can still be computed exactly and has a meaningful interpretation in our context. Finally, in the cases $k \in \{\rho - 1, \rho, \rho + 1\}$ we will see that although this expectation is no longer defined for all of the matrix entries, we can still interpret its limiting form in terms of a pseudoinverted data space covariance.

7.2.1 Analysis of $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$

There are three cases to consider:

Case $k < \rho - 1$

To fix ideas we will look first at the case $k = 1$, when we are projecting the high dimensional data on to a single line for each classifier in the ensemble. In this case the i-th diagonal element of $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$ is $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho}\lambda_j r_j^2}\right]$, where r_i is the i-th entry of the single row matrix R. This can be upper and lower bounded as:

$$\frac{1}{\lambda_{\max}}\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] \leqslant \mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho}\lambda_j r_j^2}\right] \leqslant \frac{1}{\lambda_{\min\neq 0}}\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right]$$

where $\lambda_{\min\neq 0}$ denotes the smallest non-zero eigenvalue of $\hat{\Lambda}$ (and of $\hat{\Sigma}$), and $\lambda_{\max}$ its largest eigenvalue.
Recall that as a result of lemmas 19 and 20 we only need consider the diagonal entries of this expectation as the off-diagonal terms are known to be zero.
Now, we evaluate the remaining expectation. There are two cases: If $i > \rho$ then r_i is independent from the denominator and we have $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] = \mathrm{E}\left[r_i^2\right]\mathrm{E}\left[1/\sum_{j=1}^{\rho} r_j^2\right] = \frac{1}{\rho-2}$, $\rho > 2$ where we used the expectation of the inverse-χ^2 with ρ degrees of freedom, and the fact that $\mathrm{E}\left[r_i^2\right] = 1$. When $i \leqslant \rho$, then in turn we have $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] = \mathrm{E}\left[\frac{r_i^2}{\|r\|^2}\right] = \frac{1}{\rho}$. That is,

$$\mathrm{diag}\left(\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right]\right) = \left[\begin{array}{c|c} \frac{1}{\rho}I_\rho & 0 \\ \hline 0 & \frac{1}{\rho-2}I_{d-\rho} \end{array}\right]$$

and so $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$ is full rank, hence invertible. Its inverse may be seen as a regularized covariance estimate in the data space, and its condition number, κ, is upper bounded by:

$$\kappa \leqslant \frac{\rho}{\rho-2}\cdot\frac{\lambda_{\max}}{\lambda_{\min\neq 0}} \tag{7.2.3}$$

whereas in the setting $N < d$ the ML covariance estimate has unbounded condition number.
For the general $k < \rho - 1$ case we write R as a concatenation of two matrices $R = [P, S]$ where P is $k \times \rho$ and S is $k \times (d - \rho)$, so that $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$ can be decomposed as two diagonal blocks:

$$\left[\begin{array}{c|c} \mathrm{E}[P^T\left(P\underline{\hat{\Lambda}}P^T\right)^{-1}P] & 0 \\ \hline 0 & \mathrm{E}[S^T\left(P\underline{\hat{\Lambda}}P^T\right)^{-1}S] \end{array}\right] \tag{7.2.4}$$

Where here in $P\hat{\underline{\Lambda}}P^T$ we use $\hat{\underline{\Lambda}}$ to denote the $\rho \times \rho$ positive definite upper block of the positive semi-definite matrix $\hat{\Lambda}$. Now, rewrite the upper block to orthonormalize P as the following: $\mathrm{E}[P^T \left(P\hat{\underline{\Lambda}}P^T\right)^{-1} P] =$

$$\mathrm{E}[P^T(PP^T)^{-\frac{1}{2}} \left((PP^T)^{-\frac{1}{2}}P\hat{\underline{\Lambda}}P^T(PP^T)^{-\frac{1}{2}}\right)^{-1} (PP^T)^{-\frac{1}{2}}P]$$

Denoting by P_i the i-th column of P, we can write and bound the i-th diagonal element as:

$$\mathrm{E}[P_i^T(PP^T)^{-\frac{1}{2}} \left((PP^T)^{-\frac{1}{2}}P\hat{\underline{\Lambda}}P^T(PP^T)^{-\frac{1}{2}}\right)^{-1} (PP^T)^{-\frac{1}{2}}P_i]$$

$$\begin{aligned} &\leqslant \mathrm{E}\left[\frac{P_i^T(PP^T)^{-1}P_i}{\lambda_{\min}((PP^T)^{-\frac{1}{2}}P\hat{\underline{\Lambda}}P^T(PP^T)^{-\frac{1}{2}})}\right] \\ &\leqslant \mathrm{E}\left[\frac{P_i^T(PP^T)^{-1}P_i}{\lambda_{\min\neq 0}}\right] \end{aligned}$$

with $\lambda_{\min\neq 0}$ the smallest non-zero eigenvalue of $\hat{\Lambda}$ as before, and where we used the Rayleigh quotient and the Poincaré separation theorem respectively (e.g. Horn & Johnson, 1985, Thm 4.2.2, Corr 4.3.16). This holds for all i, so then replacing we have:

$$\mathrm{E}[P^T(PP^T)^{-1}P]/\lambda_{\min\neq 0} \succcurlyeq \mathrm{E}\left[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P\right] \tag{7.2.5}$$

where $A \succcurlyeq B$ denotes $A - B$ is positive semi-definite.

Now the remaining expectation can be evaluated using the expectation of the ρ-dimensional Wishart matrix P^TP with k degrees of freedom:

$$\mathrm{E}[P^T(PP^T)^{-1}P] = \mathrm{E}[P^TP]/\rho = \frac{k}{\rho} \cdot I_\rho \tag{7.2.6}$$

Similarly to eq. (7.2.5) we can also show that:

$$\mathrm{E}\left[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P\right] \succcurlyeq \mathrm{E}[P^T \left(PP^T\right)^{-1} P]/\lambda_{\max} \tag{7.2.7}$$

in much the same way. Put together, the diagonal elements in the upper block are all in the interval:

$$\left[\frac{1}{\lambda_{\max}}\frac{k}{\rho}, \frac{1}{\lambda_{\min\neq 0}}\frac{k}{\rho}\right]$$

Hence, we see that in this upper block the condition number is reduced in comparison to that of $\hat{\Lambda}$ in its column space.

$$\frac{\lambda_{\max}(\mathrm{E}[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P])}{\lambda_{\min}(\mathrm{E}[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P])} \leqslant \frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min\neq 0}(\hat{\Lambda})}$$

That is, in the range of $\hat{\Sigma}$, the ensemble has the effect of a shrinkage regularizer (Friedman, 1989; Ledoit & Wolf, 2004). Next, we consider its effect in the null space of $\hat{\Sigma}$.

The lower block is $\mathrm{E}\left[S^T(P\underline{\hat{\Lambda}}P^T)^{-1}S\right] = \mathrm{Tr}\left(\mathrm{E}\left[(P\underline{\hat{\Lambda}}P^T)^{-1}\right]\right) \cdot I_{d-\rho}$ since S is independent of P. We again rewrite this to orthonormalize P. Going through similar steps, we obtain: $\mathrm{Tr}\left(\mathrm{E}\left[(P\underline{\hat{\Lambda}}P^T)^{-1}\right]\right) =$

$$\mathrm{Tr}\left(\mathrm{E}\left[\left(PP^T\right)^{-\frac{1}{2}}\left(\left(PP^T\right)^{-\frac{1}{2}}P\underline{\hat{\Lambda}}P^T\left(PP^T\right)^{-\frac{1}{2}}\right)^{-1}\left(PP^T\right)^{-\frac{1}{2}}\right]\right)$$

$$\leqslant \frac{\mathrm{Tr}\left(\mathrm{E}\left[\left(PP^T\right)^{-1}\right]\right)}{\lambda_{\min\neq 0}} = \frac{k}{\rho-k-1}\cdot\frac{1}{\lambda_{\min\neq 0}}$$

where we used the Poincaré inequality, lemma 6, followed by taking the expectation of the inverse Wishart. Likewise,

$$\mathrm{Tr}\left(\mathrm{E}\left[\left(P\underline{\hat{\Lambda}}P^T\right)^{-1}\right]\right) \geqslant \frac{k}{\rho-k-1}\cdot\frac{1}{\lambda_{\max}} \tag{7.2.8}$$

Hence, the lower block is a multiple of $I_{d-\rho}$ with the coefficient in the interval:

$$\left[\frac{k}{\rho-k-1}\frac{1}{\lambda_{\max}}, \frac{k}{\rho-k-1}\frac{1}{\lambda_{\min\neq 0}}\right]$$

That is, in the null space of $\hat{\Sigma}$ the ensemble acts as a ridge regularizer (Hastie et al., 2001), and the strength of the regularization depends on k and ρ, and the condition number of $\hat{\Sigma}$ restricted to its range. Specifically, $\frac{k}{\rho-k-1}$ increases monotonically with k (and decreases with ρ). Since we are talking about an inverse covariance estimate, this implies that the extent of regularization decreases with increasing k (and increases when ρ gets larger). Hence, k takes the role of the regularization parameter and the analysis in this and the following subsections provides us with insight for setting this parameter.

Putting everything together, the condition number of the covariance (or inverse covariance) estimate is upper bounded by:

$$\kappa \leqslant \frac{\rho}{\rho-k-1}\cdot\frac{\lambda_{\max}}{\lambda_{\min\neq 0}} \tag{7.2.9}$$

which we see reduces to eq.(7.2.3) when $k = 1$.

Case $k > \rho + 1$

In this case the single RP-FLD is known to have an error that increases at the rate ρ/k (Bickel & Levina, 2004).

We use the form in eq. (7.2.4) again, with P a $k \times \rho$ matrix and S a $k \times (d-\rho)$

matrix. Since here we have $k > \rho + 1$ we replace $\left(P\underline{\hat{\Lambda}}P^T\right)^{-1}$ by its pseudo-inverse. Then we can rewrite this as:

$$\begin{aligned}
\left(P\underline{\hat{\Lambda}}P^T\right)^{+} &= \left[\left(P\underline{\hat{\Lambda}}^{1/2}\right)\left(P\underline{\hat{\Lambda}}^{1/2}\right)^{T}\right]^{+} \\
&= \left[\left(P\underline{\hat{\Lambda}}^{1/2}\right)^{T}\right]^{+}\left[\left(P\underline{\hat{\Lambda}}^{1/2}\right)\right]^{+} && (7.2.10) \\
&= \left[\left(P\underline{\hat{\Lambda}}^{1/2}\right)^{+}\right]^{T}\left[\left(P\underline{\hat{\Lambda}}^{1/2}\right)\right]^{+} && (7.2.11) \\
&= P\underline{\hat{\Lambda}}^{1/2}\left(\underline{\hat{\Lambda}}^{1/2}P^TP\underline{\hat{\Lambda}}^{1/2}\right)^{-1}\left(\underline{\hat{\Lambda}}^{1/2}P^TP\underline{\hat{\Lambda}}^{1/2}\right)^{-1}\underline{\hat{\Lambda}}^{1/2}P^T && (7.2.12) \\
&= P(P^TP)^{-1}\underline{\hat{\Lambda}}^{-1}(P^TP)^{-1}P^T && (7.2.13)
\end{aligned}$$

Where (7.2.10) and (7.2.11) use lemmas 1.5 and 1.2 of Penrose (1955) respectively, and (7.2.12) follows since $\rho < k$ and so $\underline{\hat{\Lambda}}^{1/2}P^TP\underline{\hat{\Lambda}}^{1/2}$ is invertible w.p. 1. Then using (7.2.13), the first diagonal block becomes:

$$\begin{aligned}
\mathrm{E}\left[P^T\left(P\underline{\hat{\Lambda}}P^T\right)^{+}P\right] &= \mathrm{E}\left[P^TP(P^TP)^{-1}\underline{\hat{\Lambda}}^{-1}(P^TP)^{-1}P^TP\right] \\
&= \underline{\hat{\Lambda}}^{-1} && (7.2.14)
\end{aligned}$$

The second diagonal block evaluates as $\mathrm{E}\left[S^T(P\underline{\hat{\Lambda}}P^T)^{+}S\right]$:

$$\begin{aligned}
&= \mathrm{Tr}\left(\mathrm{E}\left[\left(P\underline{\hat{\Lambda}}P^T\right)^{+}\right]\right)\cdot I_{d-\rho} \\
&= \mathrm{E}\left[\mathrm{Tr}\left(\left(P\underline{\hat{\Lambda}}P^T\right)^{+}\right)\right]\cdot I_{d-\rho} \\
&= \mathrm{E}\left[\mathrm{Tr}\left(P^TP\left(P^TP\right)^{-1}\underline{\hat{\Lambda}}^{-1}\left(P^TP\right)^{-1}\right)\right]\cdot I_{d-\rho} \\
&= \frac{\mathrm{Tr}(\underline{\hat{\Lambda}}^{-1})}{k-\rho-1}\cdot I_{d-\rho} && (7.2.15)
\end{aligned}$$

where we used the expectation of the inverse Wishart matrix $(P^TP)^{-1}$ in the last step, and the property $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ in the previous step.

Hence, in this case we obtained the exact form:

$$\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{+}R\right] = \left[\begin{array}{c|c} \underline{\hat{\Lambda}}^{-1} & 0 \\ \hline 0 & \frac{\mathrm{Tr}(\underline{\hat{\Lambda}}^{-1})}{k-\rho-1}\cdot I_{d-\rho} \end{array}\right] \qquad (7.2.16)$$

It follows that implicitly when $k > \rho + 1$ the data space covariance estimate gets regularized only in its null space by the ensemble, with zero eigenvalues replaced by $\frac{k-\rho-1}{\mathrm{Tr}(\underline{\hat{\Lambda}}^{-1})}$.

We see that, unlike the previous case, here the amount of regularization increases as we *increase* k.

Case $k \in \{\rho - 1, \rho, \rho + 1\}$

For a single RP-FLD, the choice $k = \rho$ or $k = \rho \pm 1$ is bad because the ML covariance estimate in the projection space remains poorly conditioned. Detailed analysis in Hoyle (2011) has shown that having the number of points or the rank of covariance equal to the dimensionality (which is k for RP-FLD) performs even worse than pseudo-inverse FLD would when having fewer points than dimensions (which is also bad, cf. the analysis in Section 7.4 and in Bickel & Levina (2004)).

However we will show that the ensemble of RP-FLDs with the choice $k = \rho$ in fact implements an unregularized FLD *in the data space*.

To make this connection, we again use the block-diagonal form of equation (7.2.4). Now, because $k = \rho$, P is a square matrix with independent $\mathcal{N}(0, 1)$ entries and therefore it is invertible with probability 1. Hence, the upper block is

$$\mathrm{E}[P^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} P] = \mathrm{E}[P^T \left(P^T\right)^{-1} \underline{\hat{\Lambda}}^{-1} P^{-1} P] = \underline{\hat{\Lambda}}^{-1} \tag{7.2.17}$$

The lower block is $\mathrm{E}[S^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} S] = \mathrm{Tr}(\mathrm{E}[\left(P\underline{\hat{\Lambda}}P^T\right)^{-1}])$ and this expectation is infinity when $k = \rho$ (and also for $k = \rho \pm 1$) since the expectation of the inverse Wishart is undefined when its dimension is not strictly greater than its degrees of freedom. We obtain:

$$\mathrm{E}\left[R^T \left(R\hat{\Lambda}R^T\right)^{-1} R\right] = \left[\begin{array}{c|c} \underline{\hat{\Lambda}}^{-1} & 0 \\ \hline 0 & \mathrm{diag}(+\infty) \end{array}\right] \tag{7.2.18}$$

Of course, in practice, a finite average still produces a finite large number, however this has a negligible regularization effect on the covariance estimate, therefore we have essentially a pseudo-inverse like effect. It should be noted that the pseudo-inverse in the data space is not necessarily bad, but it is bad except when the original data dimension is far from being twice the number of points (Raudys & Duin, 1998). We will see in the experiments section that indeed the performance of the ensemble can be good with the settings $k > \rho$ tested – however, because the individual ensemble members are so poor it takes a much larger ensemble for the average decision to reach a reasonable performance.

To complete the cases $k = \rho \pm 1$, it is easy to see (from the previous cases) that for $k = \rho - 1$ the upper left-hand block is of the form of the corresponding block in the $k < \rho - 1$ case, while the lower right-hand block has unbounded diagonal entries where the expectation is undefined. Similarly when $k = \rho + 1$, the upper left-hand block is $\underline{\hat{\Lambda}}^{-1}$ and the lower right-hand block has unbounded diagonal entries.

Summary

Summing up, we now see how the regularization implemented by the ensemble acts to improve the conditioning of the covariance estimate in small sample conditions, through a combination of regularization schemes all parameterized by the integer parameter k:

When $k \leqslant \rho - 1$ the ML estimate $\hat{\Sigma}$ is replaced with a shrinkage regularized upper block and a spherical lower block, and as $k \to 1$ the new estimate becomes more and more spherical. Hence when k is too small the ensemble can underfit the data and perform poorly, while on the other hand a careful choice of k can do well in this range. For the data sets we used in our experiments $k \simeq \rho/2$ appears to be a reasonable rule of thumb choice.

When $k \nearrow \rho - 1$ or $k \searrow \rho + 1$ the values in the upper block approach the non-zero eigenvalues of $\hat{\Sigma}^{+}$ while in the lower block the diagonal entries become extremely large, and we recover the data space pseudo-inverse performance. Hence when $k \simeq \rho$ we overfit about as badly as pseudo-inverting in the data space.

Finally, when $k \in [\rho{+}2, d]$ we recover the original non-zero eigenvalues of $\hat{\Sigma}$ in the upper block, but regularize in its null space. In this case the individual ensemble members still have singular covariances though the expectation does not, however noting that $\frac{1}{\lceil d/k \rceil} \sum_{i=1}^{\lceil d/k \rceil} R_i (R_i \hat{\Lambda} R_i^T)^{-1} R_i$ is full rank with probability 1 implies that an ensemble size of $\lceil d/k \rceil$ already achieves a full rank covariance estimate. Alternatively one can simply pseudo-invert in this range, as we did for our experiments, or indeed use some additional regularization scheme.

Note that when we plug the expectation examined above into the classifier ensemble, this is equivalent to an ensemble with infinitely many members and therefore, for any choice of $k \notin [\rho - 1, \rho + 1]$, although we can underfit (with a poor choice of k) we cannot overfit any worse than the unregularized (pseudo-inverse) FLD data space classifier regardless of the ensemble size, since we do not learn any combination weights from the data. This intuition will be made more precise in Sections 7.3 and 7.4. This is quite unlike adaptive ensemble approaches such as AdaBoost, where it is well-known that increasing the ensemble size can indeed lead to overfitting. Furthermore, we shall see from the experiments in Section 7.6 that this guarantee vs. the performance of pseudo-inversion appears to be a conservative prediction of the performance achievable by the randomly-projected ensemble.

7.2.2 Generalization error of the ensemble for a fixed training set

Traditionally ensemble methods are regarded as 'meta-learning' approaches and although bounds exist (e.g. Koltchinskii & Panchenko, 2002) there are, to the best of our knowledge, no results giving the exact analytical form of the generalization error of any particular ensemble. Indeed, in general it is not analytically tractable to evaluate the generalization error exactly, so one can only derive bounds. Because we deal with an FLD ensemble we are able to derive the exact generalization error of the ensemble in the case of Gaussian classes with shared covariance Σ, the setting in which FLD is Bayes' optimal. This allows us to explicitly connect the performance of the ensemble to its data space analogue. We note that an upper bound on generalization error with similar behaviour can be derived for the much larger class of subgaussian distributions (see e.g. Durrant & Kabán, 2010b), therefore this Gaussianity assumption is not crucial.

Theorem 10 (Exact generalization error with Gaussian classes) *Let* $x_q | y_q \sim$

$\mathcal{N}(\mu_y, \Sigma)$, where $\Sigma \in \mathcal{M}_{d\times d}$ is a full rank covariance matrix. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with i.i.d. Gaussian entries and denote $\hat{S}^{-1} := E_R\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$. Then the exact generalization error of the converged randomly projected ensemble classifier (7.1) is given by:

$$\Pr_{x_q,y_q}\{\hat{h}_{ens}(x_q) \neq y_q\} = \sum_{y=0}^{1}\pi_y\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T\hat{S}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_y)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}}\right) \tag{7.2.19}$$

The proof of this theorem is similar in spirit to the one for a single FLD we gave in chapter 4. For completeness we give it below.

Proof of Theorem 10

Without loss of generality let x_q have label 0. By assumption the classes have Gaussian distribution $\mathcal{N}(\mu_y, \Sigma)$ so then the probability that x_q is misclassified by the converged ensemble is given by:

$$\Pr_{x_q|y_q=0}\left\{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)>0\right\} \tag{7.2.20}$$

Define $a^T := (\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}$ and observe that if $x_q \sim \mathcal{N}(\mu_0, \Sigma)$ then:

$$\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) \sim \mathcal{N}\left(\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right), \Sigma\right)$$

and so:

$$a^T\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) \sim \mathcal{N}\left(a^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right), a^T\Sigma a\right)$$

which is a univariate Gaussian. Therefore:

$$\frac{a^T\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)-a^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)}{\sqrt{a^T\Sigma a}} \sim \mathcal{N}(0,1)$$

Hence, for the query point x_q we have the probability (7.2.20) is given by:

$$\Phi\left(\frac{a^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)}{\sqrt{a^T\Sigma a}}\right)$$
$$= \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_0)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}}\right)$$

where Φ is the c.d.f of the standard Gaussian.
A similar argument deals with the case when x_q belongs to class 1, and applying the law of total probability then completes the proof.
Indeed equation (7.2.19) has the same form as the error of the data space FLD, as we saw in chapter 4, and the converged ensemble, inspected in the original data space, produces exactly the same mean estimates and covariance matrix eigenvector estimates

as FLD working on the original data set. However it has different eigenvalue estimates that result from the sophisticated regularization scheme that we analyzed in section 7.2.1.

Comment: On the effect of eigenvector estimates

We have seen that the quality of the eigenvector estimates are not affected (in expectation) by this regularization approach. These depend on the eigengaps (differences between ordered eigenvalues) of $\Sigma = \sum_{y=0}^{1} \pi_y \Sigma_y$ as well as on the data dimension and number of training examples. Although this fact is well-known from perturbation theory, the following simple but powerful example from Horn & Johnson (1985) shows clearly both the problem and the importance of eigenvalue separation. Let:

$$\Sigma = \begin{bmatrix} 1-\epsilon & 0 \\ 0 & 1+\epsilon \end{bmatrix}$$

so that Σ has eigenvalues $1 \pm \epsilon$ and eigenvectors $(1,0)^T$, $(0,1)^T$. On the other hand consider the following perturbed matrix (where the perturbation could arise from, say, estimation error or noise):

$$\Sigma + E = \begin{bmatrix} 1-\epsilon & 0 \\ 0 & 1+\epsilon \end{bmatrix} + \begin{bmatrix} \epsilon & \epsilon \\ \epsilon & -\epsilon \end{bmatrix} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}$$

This matrix also has eigenvalues $1 \pm \epsilon$, but has eigenvectors $\frac{1}{\sqrt{2}}(1,1)^T, \frac{1}{\sqrt{2}}(1,-1)^T$ regardless of how small ϵ is.

Data-dependent guarantees for the quality of eigenvector estimates from finite samples are an active area of current research (for example, see Johnstone & Lu, 2009; Paul & Johnstone, 2012; Vu, 2011; Vu & Lei, 2012), and we do not review this theory here. However, in order to give a flavour of how bad the eigenvector estimates could be in this setting and how they are affected by the eigengaps of Σ we note that in the sparse setting considered in Johnstone & Lu (2009) for high probability guarantees on the quality of the first j eigenvector estimates one requires $N \in \mathcal{O}\left(\frac{1}{(\lambda_i - \lambda_{i+1})^2} \log d\right)$, $\forall i \leqslant j$ as well as that the remaining $d-j$ eigenvalues are not far from zero. We therefore see that in the small sample setting we consider here if the eigengaps of Σ are too small we can expect bad estimates of eigenvectors and therefore, following the argument made in Ledoit & Wolf (2004), a small value of k is likely to be a good choice since the more spherical regularized covariance will tend to reduce the effect of the poor eigenvector estimates. Conversely, following Johnstone & Lu (2009) if the eigengaps are large then better eigenvector estimates are likely from the same sample size and a larger k should then work better. We will make this intuition more concrete later, and show how it affects the generalization error of the ensemble, in Section 7.3 (and particularly in Section 7.3) where we will show that the generalization error of the ensemble can be bounded above by an expression that depends on covariance misestimation only through the condition number $\kappa(\Sigma(\Sigma+E)^{-1})$. For the above toy example, the eigenvalues of $\Sigma(\Sigma+E)^{-1}$ are $\frac{1 \pm \epsilon\sqrt{2-\epsilon^2}}{1-\epsilon^2}$, so its condition number is $\frac{1+\epsilon\sqrt{2-\epsilon^2}}{1-\epsilon\sqrt{2-\epsilon^2}}$. For small ϵ this remains fairly close to one – meaning that eigenvector misestimation has a negligible effect on classification performance when the eigengap of this toy example of Σ is small.

7.3 Tail bound on the generalization error of ensemble when $k < \rho - 1$

The previous section gave the exact generalization error of our ensemble conditional on a given training set. In this section our goal is to derive an upper bound with high probability on the ensemble generalization error w.r.t. random draws of the training set. We restrict ourselves to the choice $k < \rho - 1$ (in Section 7.3), which is arguably the most interesting one in practice; and the range where we empirically observe the best classification performance for the smallest computational cost.

We will use the following concentration lemma:

Lemma 21 (Concentration inequalities for exponential random variables) *Let $X = (X_1, X_2, X_3, \ldots, X_d)$ be a sequence of Gaussian random variables in $\mathbb{R}^d$ with mean vector $E[X] = \mu$ and covariance matrix Σ. Let $\epsilon > 0$. Then:*

$$Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(Tr(\Sigma) + \|\mu\|^2\right)\right\} \leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon} - 1\right)^2\right) \quad (7.3.1)$$

Furthermore, if $\epsilon \in (0,1)$:

$$Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(Tr(\Sigma) + \|\mu\|^2\right)\right\} \leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon} - 1\right)^2\right) \quad (7.3.2)$$

This result follows immediately as a special case of the more general result, lemma 23, used in the next chapter 8. We give the proof in the appendix. Now we can bound the generalization error of the RP-FLD ensemble. We begin by decomposing the numerator of the generalization error term (for a single class) obtained in Theorem 10 as follows:

$$\begin{aligned} & (\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \\ = \; & (\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) + 2(\hat{\mu}_0 - \mu_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \end{aligned} \quad (7.3.3)$$

Using this decomposition we can rewrite the argument of the first term in Theorem 10 in the following form:

$$\Phi\left(-\frac{1}{2}[A - B]\right)$$

Where:

$$A = \frac{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} \Sigma \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}} \quad (7.3.4)$$

and:

$$B = \frac{2(\mu_0 - \hat{\mu}_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} \Sigma \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}} \quad (7.3.5)$$

We will lower bound A and upper bound B to bound the whole term from above and, since Φ is monotonic increasing in its argument, this will give the upper bound on generalization error.

Lower-bounding the term A

Applying the Kantorovich inequality, lemma 8, A is lower bounded by:

$$\|\Sigma^{-\frac{1}{2}}\left(\hat{\mu}_1-\hat{\mu}_0\right)\| \cdot \frac{2\sqrt{\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})}}{1+\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})} \tag{7.3.6}$$

where $\kappa(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ denotes the condition number of the matrix A.

Next, since $\Sigma^{-\frac{1}{2}}\hat{\mu}_1$ and $\Sigma^{-\frac{1}{2}}\hat{\mu}_0$ are independent with $\Sigma^{-\frac{1}{2}}\hat{\mu}_y \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\mu_y, I_d/N_y)$, we have $\Sigma^{-\frac{1}{2}}(\hat{\mu}_1-\hat{\mu}_0) \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0), N/(N_0N_1)\cdot I_d)$.

Applying the concentration bound Lemma 21, (7.3.2), we have:

$$\|\Sigma^{-\frac{1}{2}}(\hat{\mu}_1-\hat{\mu}_0)\| \geqslant \sqrt{(1-\epsilon)\left(\frac{d\cdot N}{N_0N_1}+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2\right)} \tag{7.3.7}$$

with probability at least:

$$1-\exp\left(-\frac{d+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2 N_0N_1/N}{2}\left(\sqrt{1-\epsilon}-1\right)^2\right) \tag{7.3.8}$$

To complete the bounding of the term A, we denote $g(a) := \frac{\sqrt{a}}{1+a}$, and observe that this is a monotonic decreasing function on $[1,\infty)$. So, replacing a with the condition number $\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}}) \in [1,\infty)$ we need to upper bound this condition number in order to lower bound g. Denoting this upper bound by $\bar{\kappa}$, which will be quantified in Section 7.3, then the term A is lower bounded with high probability by:

$$A \geqslant 2g(\bar{\kappa})\sqrt{(1-\epsilon)\left(\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2+\frac{d\cdot N}{N_0N_1}\right)} \tag{7.3.9}$$

Upper-bounding the term B

We can rewrite B by inserting $\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}$, and using Cauchy-Schwarz in the numerator to give:

$$B \leqslant \frac{2\|\Sigma^{-\frac{1}{2}}(\mu_0-\hat{\mu}_0)\|\cdot\|\Sigma^{\frac{1}{2}}\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\|}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}} \tag{7.3.10}$$

After cancellation, this simplifies to:

$$= 2\|\Sigma^{-\frac{1}{2}}(\mu_0-\hat{\mu}_0)\| \tag{7.3.11}$$

and so by Lemma 21, 7.3.1, we have:

$$B \leqslant 2\sqrt{(1+\epsilon)d/N_0} \tag{7.3.12}$$

with probability at least $1-\exp(-\frac{d}{2}(\sqrt{1+\epsilon}-1)^2)$.

Upper-bounding $\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})$ for $k < \rho - 1$

Upper-bound on largest eigenvalue

By Jensen's inequality, and noting that $\lambda_{\max}(\cdot)$ is a convex function, we have:

$$\begin{aligned}
& \lambda_{\max}(\Sigma^{\frac{1}{2}}\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R]\Sigma^{\frac{1}{2}}) \\
\leqslant\ & \mathrm{E}_R[\lambda_{\max}(\Sigma^{\frac{1}{2}}R^T(R\hat{\Sigma}R^T)^{-1}R\Sigma^{\frac{1}{2}})] \\
=\ & \mathrm{E}_R[\lambda_{\max}((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T] \\
=\ & \mathrm{E}_R[\lambda_{\max}((R\Sigma R^T)^{\frac{1}{2}}(R\hat{\Sigma}R^T)^{-1}(R\Sigma R^T)^{\frac{1}{2}})] \\
=\ & \mathrm{E}_R\left[\frac{1}{\lambda_{\min}((R\Sigma R^T)^{-\frac{1}{2}}R\hat{\Sigma}R^T(R\Sigma R^T)^{-\frac{1}{2}})}\right] \\
\leqslant\ & \frac{N}{(\sqrt{N-2}-\sqrt{k}-\epsilon)^2}
\end{aligned}$$

with probability at least $1-\exp(-\epsilon^2/2), \forall\epsilon > 0$, where throughout we use the fact that the non-zero eigenvalues of AB are the same as non-zero eigenvalues of BA, in the second to last step we used the fact that for invertible matrices A we have $\lambda_{\max}(A) = 1/\lambda_{\min}(A^{-1})$, and in the last step we used that for any full row-rank matrix R, $(R\Sigma R^T)^{-\frac{1}{2}}R\hat{\Sigma}R^T(R\Sigma R^T)^{-\frac{1}{2}}$ is distributed as a k-dimensional Wishart with $N-2$ degrees of freedom and scale matrix I_k (e.g. Mardia et al., 1979, Corr. 3.4.1.2), and used the high probability lower-bound for the smallest eigenvalue of such a matrix, Eq. (2.3) in Vershynin (2012).

Lower-bound on smallest eigenvalue

Dealing with the smallest eigenvalue is less straightforward. Although $\lambda_{\min}(\cdot)$ is a concave function, Jensen's inequality does not help with lower bounding the smallest eigenvalue of the expectation since the matrix $\hat{\Sigma}$ in the argument of this expectation is singular. We therefore take a different route and start by rewriting as follows:

$$\begin{aligned}
& \lambda_{\min}(\Sigma^{\frac{1}{2}}\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)]\Sigma^{\frac{1}{2}}) \\
=\ & \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}(\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)])^{-1}\Sigma^{-\frac{1}{2}})} \\
=\ & \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}\{\hat{\Sigma}+(\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)])^{-1}-\hat{\Sigma}\}\Sigma^{-\frac{1}{2}})} \qquad (7.3.13)
\end{aligned}$$

Now, using Weyl's inequality, and the SVD decomposition $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$ combined with Lemma 19, the denominator in (7.3.13) is upper-bounded by:

$$\begin{aligned} & \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{U}\{(\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1} - \hat{\Lambda}\}\hat{U}^T\Sigma^{-\frac{1}{2}}) \\ \leqslant\ & \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \lambda_{\max}((\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1} - \hat{\Lambda})/\lambda_{\min}(\Sigma) \end{aligned} \tag{7.3.14}$$

Now observe from Lemma 20 that the matrix $\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1} - \hat{\Lambda}$ is diagonal and, from our analysis in Section 7.2.1, it has the upper ρ diagonal entries in the interval:

$$\left[(\frac{\rho}{k} - 1)\lambda_{\min\neq 0}(\hat{\Lambda}), (\frac{\rho}{k} - 1)\lambda_{\max}(\hat{\Lambda})\right]$$

and the lower $d - \rho$ diagonal entries in the interval:

$$\left[\frac{\rho - k - 1}{k}\lambda_{\min\neq 0}(\hat{\Lambda}), \frac{\rho - k - 1}{k}\lambda_{\max}(\hat{\Lambda})\right]$$

. Hence, $\lambda_{\max}((\mathrm{E}_R[R^T(R\hat{\Lambda}R)^{-1}R])^{-1} - \hat{\Lambda}) \leqslant \frac{\rho}{k}\lambda_{\max}(\hat{\Lambda})$ and so the lower-bounding of (7.3.14) continues as:

$$\geqslant \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \frac{\rho}{k}\frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\Sigma)}} \tag{7.3.15}$$

Now observe that $\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}$ is a d-dimensional standard Wishart with $N - 2$ degrees of freedom and scale matrix I_d (e.g. Mardia et al., 1979, Corr. 3.4.1.2), and using the bound in Vershynin (2012) for largest eigenvalues of standard Wishart matrices we get (7.3.15) lower-bounded as

$$\geqslant \frac{1}{(\sqrt{N-2} + \sqrt{d} + \epsilon)^2/N + \frac{\rho}{k}\frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\Sigma)}} \tag{7.3.16}$$

w.p. at least $1 - \exp(-\epsilon^2/2)$.

Finally, we bound $\lambda_{\max}(\hat{\Lambda})$ as:

$$\begin{aligned} \lambda_{\max}(\hat{\Lambda}) &= \lambda_{\max}(\hat{\Sigma}) = \lambda_{\max}(\Sigma\Sigma^{-1}\hat{\Sigma}) \\ &\leqslant \lambda_{\max}(\Sigma)\lambda_{\max}(\Sigma^{-1}\hat{\Sigma}) = \lambda_{\max}(\Sigma)\lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) \\ &\leqslant \lambda_{\max}(\Sigma)(\sqrt{N-2} + \sqrt{d} + \epsilon)^2/N \end{aligned}$$

To complete the bound on the condition number we combine the eigenvalue estimates to get, after simple algebra:

$$\kappa = \frac{\lambda_{\max}(\Sigma^{\frac{1}{2}} \cdot \mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R] \cdot \Sigma^{\frac{1}{2}})}{\lambda_{\min}(\Sigma^{\frac{1}{2}} \cdot \mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R] \cdot \Sigma^{\frac{1}{2}})} \tag{7.3.17}$$

$$\leqslant \frac{(\sqrt{N-2} + \sqrt{d} + \epsilon)^2(1 + \rho/k \cdot \kappa(\Sigma))}{(\sqrt{N-2} - \sqrt{k} - \epsilon)^2} =: \bar{\kappa}(\epsilon) \tag{7.3.18}$$

w.p. at least $1 - 2\exp(-\epsilon^2/2)$.

Remarks about the effect of k

Before proceeding to assemble our results to give the promised tail bound on the generalization error, a comment about the obtained condition number bound in (7.3.18) is now in order. It is interesting to notice the trade-off for the projection dimension k, which describes very well its role of regularization parameter in the context of our RP-FLD ensemble and places our discussion in Section 7.2.2 on firm foundations: To make the numerator smaller k needs to be large while to make the denominator larger it needs to be small. We also see natural behaviour with N, d and the conditioning of the true covariance.

From equations (7.3.9) and (7.3.12) we see that the condition number bounded by equation (7.3.18) is the only term in the generalization error bound affected by the choice of k, so we can also partly answer the question left open in Marzetta et al. (2011) about how the optimal k depends on the problem characteristics, from the perspective of classification performance, by reading off the most influential dependencies that the problem characteristics have on the optimal k. The first term in the numerator of (7.3.18) contains d but does not contain k while the remaining terms contain k but do not contain d, so we infer that in the setting of $k < \rho - 1 < d$ the optimal choice of k is not affected by the dimensionality d. Noting that, for $N < d$ we have $\rho = N - 2$ with probability 1 we see that for small N or ρ the minimizer of this condition number is achieved by a smaller k (meaning a stronger regulariser), as well as for a small $\kappa(\Sigma)$. Conversely, when N, ρ, or $\kappa(\Sigma)$ is large then k should also be large to minimize the bound.

Putting everything together

Collating the results derived so far, and re-arranging, we can state the following non-asymptotic error bound.

Theorem 11 *Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$ be a set of training data of size $N = N_0 + N_1$, subject to $N < d$ and $N_y > 1$ $\forall y$ with Gaussian class-conditionals $x|y \sim \mathcal{N}(\mu_y, \Sigma)$. Let ρ be the rank of the maximum likelihood estimate of the covariance matrix and let $0 < k < \rho - 1$ be an integer. Then for any $\delta \in (0,1)$ and any training set of size N, the generalization error of the converged ensemble of randomly projected FLD classifiers is upper-bounded w.p. $1 - \delta$ over the random draws of training set of size $N = N_0 + N_1$ by the following:*

$$\Pr_{x_q}(\hat{h}_{ens}(x_q) \neq y_q) \leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right)\left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 + \frac{dN}{N_0 N_1}} \cdots\right.\right.\right.$$

$$\left.\left.\left.\cdots - \sqrt{\frac{2N}{N_0 N_1}\log\frac{5}{\delta}}\right]_+ - \sqrt{\frac{d}{N_y}}\left(1 + \sqrt{\frac{2}{d}\log\frac{5}{\delta}}\right)\right]\right)$$

where $\bar{\kappa}$ is given by eq. (7.3.18) and $g(\cdot)$ is the function $g(a) := \frac{\sqrt{a}}{1+a}$.

Having completed the groundwork, the proof is now simple algebraic manipulation – we give the details in the appendix. This bound motivates our focus on RP-FLD only in the case where $N \ll d$. Specifically, it is easy to see (by plugging in the true parameters μ_y and Σ in the exact error (7.2.19)) that the Bayes' risk for FLD in the data space is $\sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|\right)$ but the expression in Theorem 11 converges to:

$$\sum_{y=0}^{1} \pi_y \Phi\left(-g\left(1 + \frac{d}{k}\kappa(\Sigma)\right)\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|\right)$$

where we recall that $g(1) = \frac{1}{2}$. In particular, we see from equation (7.3.18) that letting $N \to \infty$ (and so $\rho \to d$) while enforcing $k < d = \rho$ that our ensemble implements a biased estimator. When $N < d$ however, we see that the generalization error of our RP-FLD ensemble is upper bounded for any training sample containing at least one point for each class whereas, in the next Section 7.4, we show that this is not the case in the dataspace setting if we regularize by pseudo-inverting.

7.4 Lower bound on the error of pseudo-inverted FLD in the data space

In this section we give a lower bound on the error of pseudo-inverted FLD in data space by deriving a non-asymptotic extension to the negative result in Bickel & Levina (2004).

We start from the exact form of the error of FLD in the data space with a fixed training set. Using a similar approach to that employed in proving Theorem 10, this can easily be shown to be:

$$\Pr(\hat{h}_+(x_q) \neq y_q) = \sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^+ (\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^+ \Sigma \hat{\Sigma}^+ (\hat{\mu}_1 - \hat{\mu}_0)}}\right) \tag{7.4.1}$$

where $\hat{\Sigma}^+$ is the pseudo-inverse of the maximum likelihood covariance estimate.

Make the rank ρ SVD decomposition $\hat{\Sigma} = \underline{\hat{U}}\underline{\hat{\Lambda}}\underline{\hat{U}}^T$, where $\underline{\hat{U}}$ is the $d \times \rho$ matrix of eigenvectors associated with the non-zero eigenvalues, $\underline{\hat{U}}^T\underline{\hat{U}} = I_\rho$, and as before $\underline{\hat{\Lambda}}$ is the diagonal $\rho \times \rho$ matrix of non-zero eigenvalues. Then we have:

$$\begin{aligned}
&\frac{(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)^T\underline{\hat{U}}\underline{\hat{\Lambda}}^{-1}\underline{\hat{U}}^T(\hat{\mu}_1-\hat{\mu}_0)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\underline{\hat{U}}\underline{\hat{\Lambda}}^{-1}\underline{\hat{U}}^T\Sigma\underline{\hat{U}}\underline{\hat{\Lambda}}^{-1}\underline{\hat{U}}^T(\hat{\mu}_1-\hat{\mu}_0)}} \\
&\leqslant \frac{(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)^T\underline{\hat{U}}\underline{\hat{\Lambda}}^{-1}\underline{\hat{U}}^T(\hat{\mu}_1-\hat{\mu}_0)}{\sqrt{\lambda_{\min}(\Sigma)}\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\underline{\hat{U}}\underline{\hat{\Lambda}}^{-2}\underline{\hat{U}}^T(\hat{\mu}_1-\hat{\mu}_0)}} \\
&\leqslant \frac{\|\underline{\hat{U}}^T(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)\|\cdot\|\underline{\hat{\Lambda}}^{-1}\underline{\hat{U}}^T(\hat{\mu}_1-\hat{\mu}_0)\|}{\sqrt{\lambda_{\min}(\Sigma)}\|\underline{\hat{\Lambda}}^{-1}\underline{\hat{U}}^T(\hat{\mu}_1-\hat{\mu}_0)\|} \\
&= \frac{\|\underline{\hat{U}}^T(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)\|}{\sqrt{\lambda_{\min}(\Sigma)}}
\end{aligned}$$

where we used minorization by Rayleigh quotient and the fact that Σ was taken to be full rank, and the Cauchy-Schwartz inequality. We will use the well-known fact that $\hat{\Sigma}$ and $\hat{\mu}_1+\hat{\mu}_0$ are independent (Mardia et al., 1979). Observe that $\underline{\hat{U}}^T$ is a $\rho \times d$ random matrix with orthonormal rows representing the eigenvectors of the sample covariance of the canonical d-variate Gaussian distribution. Using the rotational invariance of this distribution, these eigenvectors are uniformly distributed over the d-dimensional hypersphere S^{d-1} and with high probability the action of this matrix on any d-dimensional vector, x, is therefore a Johnson-Lindenstrauss embedding of x into a random subspace of dimension ρ which approximately preserves its norm (Dasgupta & Gupta, 2002). Conditioning on $\hat{\mu}_1+\hat{\mu}_0$ to hold this quantity fixed, and using independence of $\underline{\hat{U}}$ and $\hat{\mu}_1+\hat{\mu}_0$, we have with probability at least $1-\exp(-N\epsilon^2/8)$ that:

$$\frac{\|\underline{\hat{U}}^T(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)\|}{\sqrt{\lambda_{\min}(\Sigma)}} \leqslant \sqrt{1+\epsilon}\sqrt{\frac{\rho}{d}}\frac{\|\hat{\mu}_1+\hat{\mu}_0-2\mu_0\|}{\sqrt{\lambda_{\min}(\Sigma)}}$$

Further, applying Lemma 21 (7.3.1) to the norm on the r.h.s and replacing in the generalization error expression, we have the following lower bound:

$$\Phi\left(-\frac{1}{2}\sqrt{(1+\epsilon_1)(1+\epsilon_2)}\sqrt{\frac{\rho}{d}\frac{\|\mu_1-\mu_0\|^2+\text{Tr}(\Sigma)\frac{N}{N_0N_1}}{\lambda_{\min}(\Sigma)}}\right)$$

with probability at least $1-[\exp(-N\epsilon_1^2/8)+\exp(-\frac{\text{Tr}(\Sigma)+\|\mu_1-\mu_0\|^2\frac{N_0N_1}{N}}{2\lambda_{\max}(\Sigma)}(\sqrt{1+\epsilon_2}-1)^2)]$.

Setting both of these exponential risk probabilities to $\delta/2$ and solving for ϵ_1 and ϵ_2, we have the following lower bound on the generalization error of pseudo-inverted FLD:

Theorem 12 (Lower bound on generalization error of pseudo-inverted FLD) *For any $\delta \in (0,1)$ and $\kappa(\Sigma) < \infty$, under the same assumptions as theorem 11, the generalization error of pseudo-inverted FLD is lower-bounded w.p. at least $1-\delta$ over the* *random draws of training set by:*

$$\Pr(\hat{h}_+(x_q) \neq y_q) \geqslant \Phi\left(-\frac{1}{2}\sqrt{1+\sqrt{\frac{8}{N}\log\frac{2}{\delta}}}\left(1+\sqrt{\frac{2\lambda_{\max}(\Sigma)\log(2/\delta)}{Tr(\Sigma)+\|\mu_1-\mu_0\|^2\frac{N_0N_1}{N}}}\right)\sqrt{\frac{\rho}{d}\frac{\|\mu_1-\mu_0\|^2+Tr(\Sigma)\frac{N}{N_0N_1}}{\lambda_{\min}(\Sigma)}}\right)$$

It is interesting to notice that this lower bound depends on the rank of the covariance estimate, not on its form or on its fit to the true covariance Σ. We see when $\rho < d$ the bound proves the bad performance of pseudo-inverted FLD since, as $\hat{\Sigma}$ is the ML estimate of Σ, the rank of $\hat{\Sigma}$ is at most $N-2$ and the lower bound (7.4.2) becomes tighter as ρ/d decreases. Allowing the dimensionality d to be large, as in (Bickel & Levina, 2004), so that $\rho/d \to 0$, this fraction goes to 0 which means the lower bound (7.4.2) converges to $\Phi(0) = 1/2$ – in other words random guessing.

7.5 Variability reduction and bounding the deviation of a finite ensemble from its expectation

So far we have demonstrated that our ensemble of RP-FLD classifiers implements a sophisticated regularization scheme at convergence. Moreover, for any choice of k for which $\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R]$ exists, the law of large numbers implies that increasing the ensemble size acts to reduce the variance of the ensemble classifier on average. It would be interesting to know how quickly a finite ensemble approaches the converged ensemble, and this could be achieved by quantifying the rate at which the extreme eigenvalues of the covariance matrix of a finite ensemble approach the extreme eigenvalues of the covariance matrix of the converged ensemble. Furthermore, since the error of the ensemble can be bounded via the condition number of the ensemble covariance matrix, a natural question to ask anyway is how far the condition number of the covariance matrix of a finite ensemble of M RP-FLD classifiers lies from the condition number of the covariance matrix of its expectation, with high probability.

Here we derive some simple high probability guarantees on the scale of the extreme eigenvalues of the ensemble covariance, for the case $k < \rho$, under some boundedness conditions. For the largest eigenvalue of the covariance matrix of a finite ensemble we also prove as a corollary a looser, but somewhat more informative, high probability guarantee upper-bounding this quantity.

Theorem 13 (Large deviation inequalities for sums of symmetric random matrices)

Let $S_1, \ldots S_M$ be a sequence of independent positive semi-definite symmetric random matrices such that $\forall i$ we have, almost surely, that: $0 \leqslant \lambda_{\min}(S_i) \leqslant \lambda_{\max}(S_i) \leqslant L$. Let $t > 0$. Then $\forall t \in (0, \lambda_{\max}(\mathrm{E}[S]))$:

$$\Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right) - \lambda_{\max}\left(\mathrm{E}\left[S\right]\right) \leqslant -t\right\} \leqslant \exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.1}$$

and $\forall t > 0$:

$$\Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right) - \lambda_{\max}\left(\mathrm{E}\left[S\right]\right) \geqslant t + \left(\mathrm{Tr}\left(\mathrm{E}\left[S\right]\right) - \lambda_{\max}\left(\mathrm{E}\left[S\right]\right)\right)\right\} \leqslant \exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.2}$$

Furthermore $\forall t > 0$:

$$\Pr\left\{\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right) - \lambda_{\min}\left(\mathrm{E}\left[S\right]\right) \geqslant t\right\} \leqslant \exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.3}$$

and $\forall t \in (0, \mathrm{E}\left[\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right)\right]]$:

$$\Pr\left\{\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right) - \lambda_{\min}\left(\mathrm{E}\left[S\right]\right) \leqslant -t + \mathrm{E}\left[\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right)\right] - \lambda_{\min}\left(\mathrm{E}\left[S\right]\right)\right\} \leqslant \exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.4}$$

Proof 10 (of Theorem 13)
We begin by confirming the result of Shawe-Taylor et al. (2005) that the extreme eigenvalues of the empirical average:

$$\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right) \text{ and } \lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right)$$

are concentrated, not as one might expect about $\lambda_{\min}(E[S])$ *and* $\lambda_{\max}(E[S])$*, but rather about* $E\left[\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right)\right]$ *and* $E\left[\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right)\right]$*. Furthermore these are not, in general, the same quantities.*
We will employ the following standard tools (which are all given in Chapter 2): Weyl's inequality (Lemma 4), McDiarmid's Inequality (Lemma 15), and Jensen's Inequality (Lemma 14). First note that the functions:

$$\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right) \text{ and } \lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M} S_i\right)$$

satisfy the bounded differences condition on McDiarmid's inequality, lemma 15, in each case with constant terms $\frac{1}{M}L$, $\forall i$*. This is because, by Weyl's inequality, removing any single matrix in the sum changes the upper and lower bounds on the extreme eigenvalues of the sum by at most the difference between the least and greatest possible values that the eigenvalues can take, and these quantities are almost surely bounded in absolute value since they are non-negative and (by assumption) almost surely bounded above. Therefore, as previously shown in (Shawe-Taylor et al., 2005), a direct application of*

McDiarmid gives the following high probability concentration guarantees:

$$Pr\left\{\left|\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-E\left[\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\right|\geqslant t\right\}\leqslant 2\exp\left(\frac{-2t^2M}{L^2}\right)\ \textit{and} \tag{7.5.5}$$

$$Pr\left\{\left|\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-E\left[\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\right|\geqslant t\right\}\leqslant 2\exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.6}$$

where we see that the concentration is indeed about the expectations of the eigenvalues and not about the eigenvalues of the expectation. The following Lemma 22 shows clearly that these quantities are, in general, different:

Lemma 22 (Non-linearity of extreme eigenvalues of symmetric matrices)
Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ be, respectively, the greatest and least eigenvalue of their symmetric (or Hermitian) matrix argument. Then $\lambda_{\max}(\cdot)$ is a convex function and $\lambda_{\min}(\cdot)$ is a concave function.
Proof: Let A, B be Hermitian or symmetric matrices then, by Weyl's inequality, we have:

$$\lambda_{\max}(\alpha A+(1-\alpha)B)\leqslant\lambda_{\max}(\alpha A)+\lambda_{\max}((1-\alpha)B)=\alpha\lambda_{\max}(A)+(1-\alpha)\lambda_{\max}(B),\ \forall\,\alpha\in[0,1]$$

The proof for $\lambda_{\min}(\cdot)$ is much the same, using the other side of Weyl's inequality.

Using the facts that $\lambda_{\max}$ is non-negative by the assumption of positive semi-definiteness, and is a convex function of its matrix argument by lemma 22, by Jensen's inequality we see that:

$$\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-E\left[\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\leqslant\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}(E[S])$$

and therefore applying this to the expression (7.5.5), we obtain the one-sided bound:

$$Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}(E[S])\leqslant -t\right\}\leqslant\exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.8}$$

which holds $\forall t>0$. Similar reasoning, using the fact that $\lambda_{\min}$ is concave, applied to the expression (7.5.6) gives a further one-sided bound:

$$Pr\left\{\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\min}(E[S])\geqslant t\right\}\leqslant\exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.9}$$

These one-sided bounds only prove that the empirical average of our random matrices can be more poorly conditioned than $E[S]$ with high probability. We also want to know how well-conditioned this empirical average is compared to its expectation. In particular we would like to show that the condition number converges quickly as a function of increasing M. To show this we start by rewriting the other side of the first bound, the *expression 7.5.8, above from McDiarmid's inequality:*

$$Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}\left(E[S]\right)\geqslant t-\lambda_{\max}\left(E[S]\right)+E\left[\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\right\}\leqslant\exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.10}$$

Clearly the LHS of this inequality is converging to t as M increases, but this is somewhat unsatisfactory as the RHS is likewise converging (to zero) with M at the same time. We would like to bound the magnitude of the LHS to obtain a high probability large deviation guarantee where the probability of failure will be exponentially decreasing as a function of M.
Taking this route then, for the largest eigenvalue we have the following corollary $\forall t\in(0,\lambda_{\max}(E[S]))$*:*

$$\begin{aligned}
& Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}\left(E[S]\right)\geqslant t-\lambda_{\max}\left(E[S]\right)+E\left[\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\right\}\\
\geqslant\ & Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}\left(E[S]\right)\geqslant t-\lambda_{\max}\left(E[S]\right)+E\left[Tr\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\right\}\\
=\ & Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}\left(E[S]\right)\geqslant t-\lambda_{\max}\left(E[S]\right)+Tr\left(E\left[\frac{1}{M}\sum_{i=1}^{M}S_i\right]\right)\right\}\\
=\ & Pr\left\{\lambda_{\max}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-\lambda_{\max}\left(E[S]\right)\geqslant t+\left(Tr(E[S])-\lambda_{\max}\left(E[S]\right)\right)\right\}\leqslant\exp\left(\frac{-2t^2M}{L^2}\right)
\end{aligned} \tag{7.5.11}$$

That is, with high probability, the largest eigenvalue of the empirical average of the sequence of random matrices is overestimated by no more than the sum of the $d-1$ *smallest eigenvalues of the matrix expectation.*
For the smallest eigenvalue, directly from McDiarmid, we have:

$$Pr\left\{\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)-E\left[\lambda_{\min}\left(\frac{1}{M}\sum_{i=1}^{M}S_i\right)\right]\leqslant -t\right\}\leqslant\exp\left(\frac{-2t^2M}{L^2}\right) \tag{7.5.12}$$

and this completes the proof of the theorem.

Comment

It is worth noting here that, since all matrices in the sum are p.s.d (and therefore all eigenvalues are bounded below by zero with certainty), in order to apply these results in our ensemble setting the crucial quantity over which we would need control is the largest eigenvalue of the inverse covariance estimate of each ensemble member. Specifically, we would need to show that these largest eigenvalues are bounded above almost surely by an absolute constant.
Unfortunately this appears to be very hard to do, and the main technical issues arise from the fact that $\hat{\Sigma}$ is singular. To see this consider any fixed instance of a Gaussian random projection matrix R and observe that in $R^T(R\hat{\Lambda}R^T)^{-1}R$, although the

central matrix $R\hat{\Lambda}R^T$ is almost surely full rank, its smallest eigenvalue which controls $\lambda_{\max}(R^T(R\hat{\Lambda}R^T)^{-1}R)$ can be arbitrarily close to zero. One could bound the smallest eigenvalue of $R\hat{\Lambda}R^T$ away from zero with high probability, but then the probability there exists some matrix in the ensemble violating such a lower bound on $\lambda_{\min}(R\hat{\Lambda}R^T)$ will go to 1 as the ensemble size increases, and the condition on McDiarmid's inequality then fails to hold. Therefore, to make progress in bounding the largest eigenvalue it appears that one has to either work with the full matrix $R^T(R\hat{\Lambda}R^T)^{-1}R$, which has a complicated form and distribution, or with the sum of these matrices which looks equally hard.

On the other hand, our experimental results in the next section 7.6, in particular the stability of the classification performance of the finite ensembles considered there, suggest that theorem 13 characterizes the rate of convergence of the ensemble well.

7.6 Experiments

We now present experimental results which show that our ensemble approach is competitive with the state of the art in terms of prediction performance. We do not claim of course that the choice of FLD as a classifier is optimal for these data sets, rather we demonstrate that the various practical advantages of the RP-FLD approach do not come at a cost in terms of prediction performance. For example, some nice properties of the RP-FLD approach include: Covariance matrices are interpretable in a range of problem settings, the RP-FLD ensemble is simple to implement, fitting k carefully can be done exhaustively if needed, training data can be collected, stored and processed in compressed form, and, (both for training and classification) the ensemble members can run on separate cores. We also note that when d is large and k is small savings in time and space complexity (vs FLD in the data space) are possible with our ensemble approach. For example, inverting a full covariance matrix in the data space using Gauss-Jordan has time complexity $\mathcal{O}(d^3)$ while the corresponding step for the projected ensemble takes $M \cdot \mathcal{O}(k^3)$ on a single core.

7.6.1 Datasets

We used five publicly available high dimensional datasets from the bioinformatics domain (colon, two versions of leukemia, prostate, and duke breast cancer), whose characteristics are as described in Table 7.1. The first two (colon and leukemia) have the smallest dimensionality amongst these and were the highest dimensional data sets used in the empirical RP-classifier study of Fradkin & Madigan (2003a) (although that paper focuses on a single randomly projected classifier vs. the data space equivalent).

7.6.2 Protocol

We standardized each data set to have features with mean 0 and variance 1, and ran experiments on 100 independent splits. In each split we took 12 points for testing and used the remainder for training. For our data space experiments on colon and leukemia we used FLD with ridge regularization and fitted the regularization parameter using 5-fold cross-validation on the first five data splits following Mika et al. (2002). However

Table 7.1: Datasets

Name	Source	#samples	#features
colon	Alon et al. (1999)	62	2000
leukemia	Golub et al. (1999)	72	3571
leukemia large	Golub et al. (1999)	72	7129
prostate	Singh et al. (2002)	102	6033
duke	West et al. (2001)	44	7129

Table 7.2: Mean error rates ± 1 standard error, estimated from 100 independent splits when $k = \rho/2$.

Dataset	$\rho/2$	100 RP-FLD	1000 RP-FLD	SVM
colon	24	13.58 ± 0.89	13.08 ± 0.86	16.58 ± 0.95
leuk.	29	1.83 ± 0.36	1.83 ± 0.37	1.67 ± 0.36
leuk.lge	29	4.91 ± 0.70	3.25 ± 0.60	3.50 ± 0.46
prost.	44	8.00 ± 0.76	8.00 ± 0.72	8.00 ± 0.72
duke	15	17.41 ± 1.27	16.58 ± 1.27	13.50 ± 1.10

on these data this provided no statistically significant improvement over employing a diagonal covariance in the data space, most likely because of the data scarcity. Therefore for the remaining three datasets (which are even higher dimensional) we used diagonal FLD in the data space. Indeed since diagonal FLD is in use for gene array data sets (Dudoit et al., 2002) despite the features being known to be correlated (this constraint acting as a form of regularization) one of the useful benefits of our ensemble is that such a diagonality constraint is no longer necessary.

The randomly projected base learners are FLDs with full covariance and no regularization when $k \leqslant \rho$ (as the projected sample covariances are invertible) and we used pseudo-inversion in the projected space when $k > \rho$ – cf. the setting analyzed in the previous section.

To satisfy ourselves that building on FLDs was a reasonable choice of classifier we also ran experiments using SVM with linear kernel, as was done in Fradkin & Madigan (2003a).

7.6.3 Results

In each case we compare the performance of the RP ensembles with (regularized) FLD in the data space and also with SVM. Summary results for the rule of thumb choice $k = \rho/2$ are listed in Table 7.2. In figure 7.1 we plot the results for the regularized data space FLD, for a single RP-FLD, and for ensembles of 10, 100, and 3000 RP-FLD

classifiers. We see in all cases that our theoretical analysis is well supported, the RP-FLD ensemble outperforms traditional FLD on a range of choices of k, and the rule of thumb choice $k = \rho/2$ is not far from the optimal performance. It is interesting to see that, despite the statistically insignificant difference in performance of full-vs-diagonal covariance models we found for the two lower-dimensional data sets in the data space, for the three higher dimensional data sets (where we used a diagonality constraint for computational tractability) the gap in generalization performance of the data space FLD vs SVM is very large, whereas the gap in performance between the RP-FLD ensembles and SVM is small. Empirically we see, as we might reasonably expect, that capturing the feature covariances via our ensemble approach produces better classification results than working in the data space with a diagonal covariance model.
We ran further experiments on the colon and leukemia data sets to compare the performance of the fast random projections from Achlioptas (2003) to Gaussian random projection matrices, and to compare our decision rule to majority vote. Quite interestingly, the picture is very similar and we find no statistically significant difference in the empirical results in comparison with the ensemble that we have presented and analyzed in detail here. The results of these experiments are plotted in figure 7.2. The performance match between the different choices of random matrix is unsurprising, but the agreement with majority vote is both striking and rather unexpected - we do not yet have an explanation for this behaviour, although it does not appear to arise from the unsigned confidences of the individual ensemble members being concentrated around a particular value.

7.7 Summary and Discussion

We considered a randomly projected (RP) ensemble of FLD classifiers and gave theory which, for a fixed training set, explicitly links this ensemble classifier to its data space analogue. We have shown that the RP ensemble implements an implicit regularization of the corresponding FLD classifier in the data space. We demonstrated experimentally that the ensemble can recover or exceed the performance of a carefully-fitted ridge-regularized data space equivalent but with generally lower computational cost. Our theory guarantees that, for most choices of projection dimension k, the error of a large ensemble remains bounded even when the number of training examples is far lower than the number of data dimensions and we gained a good understanding of the effect of our discrete regularization parameter k. In particular, we argued that the regularization parameter k allows us to finesse the known issue of poor eigenvector estimates in this setting. We also demonstrated empirically that we can obtain good generalization performance even with few training examples, and a rule of thumb choice $k = \rho/2$ appears to work well.
We showed that, for classification, the optimal choice of k depends on the true data parameters (which are unknown) thereby partly answering the question in Marzetta et al. (2011) concerning whether a simple formula for the optimal k exists. It would be interesting to extend this work to obtain similar guarantees for ensembles of generic randomly-projected linear classifiers in convex combination. Furthermore, it would be interesting to derive concentration inequalities to quantify with what probability the

condition number of the sample covariance matrix of a finite ensemble of the form we consider here is far from its expectation – however this appears to be a hard problem and, in particular, the rank deficiency of $\hat{\Sigma}$ is problematic.

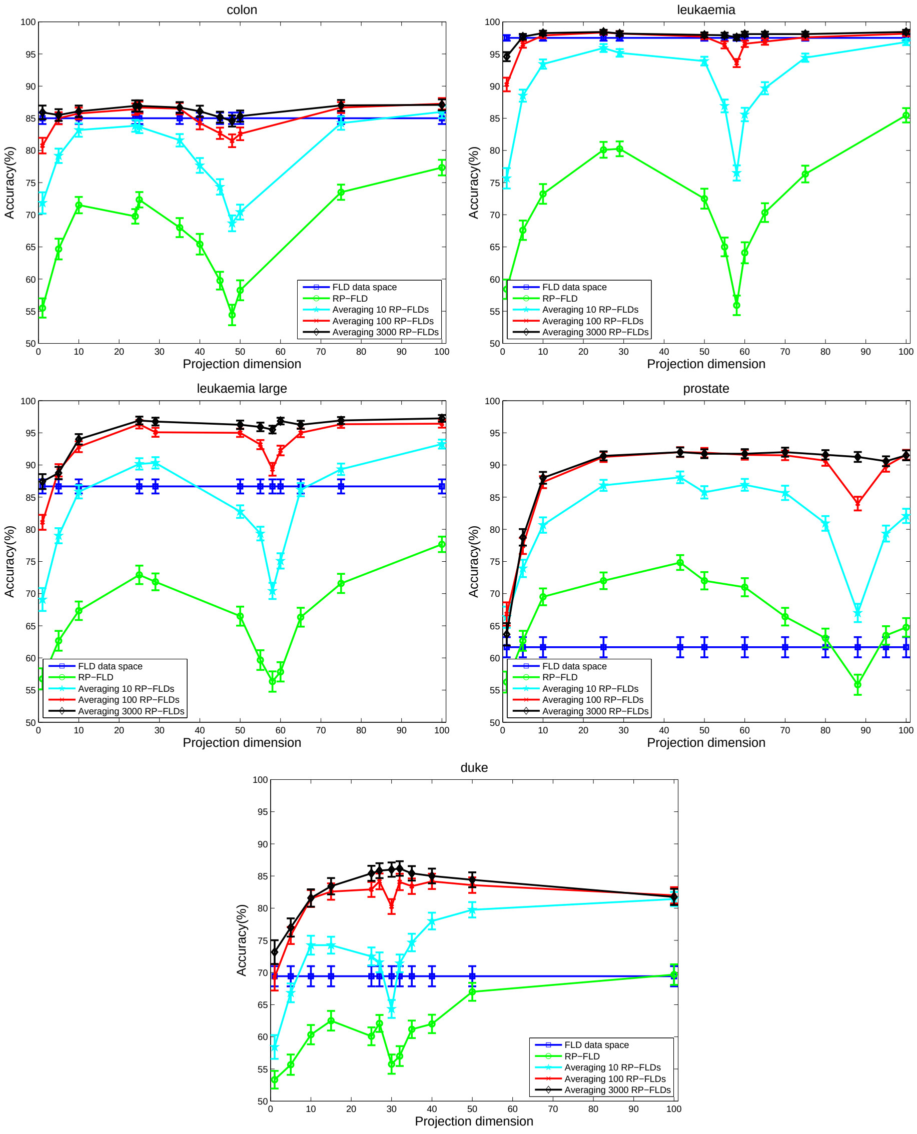

Figure 7.1: Effect of k. Plots show test error rate versus k and error bars mark 1 standard error estimated from 100 runs. In these experiments we used Gaussian random matrices with i.i.d $\mathcal{N}(0, 1)$ entries.

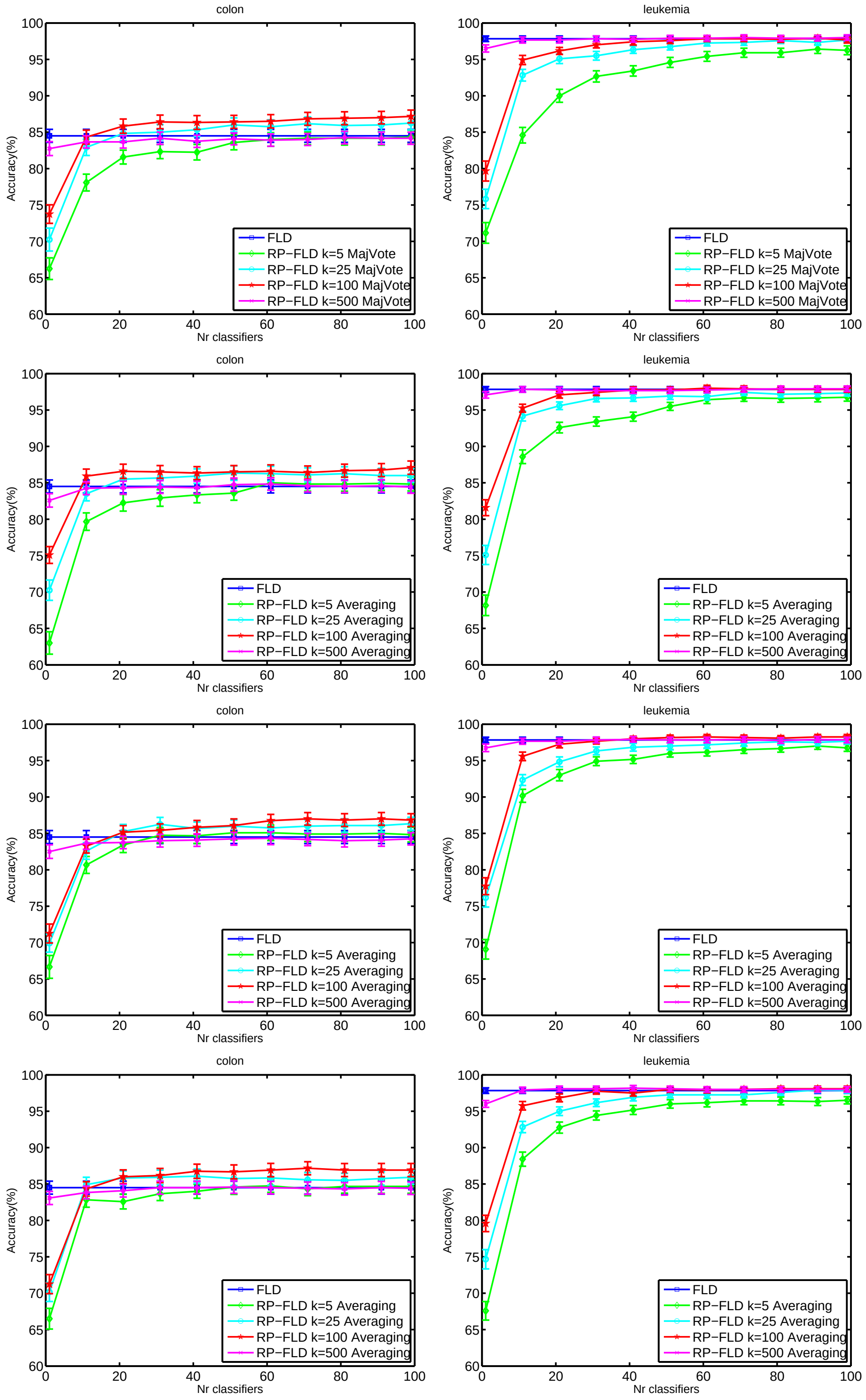

Figure 7.2: Row 1: RP Majority Vote using Gaussian random matrices with i.i.d $\mathcal{N}(0, 1)$ entries; Row 2: RP Voting using Gaussian random matrices with i.i.d $\mathcal{N}(0, 1)$ entries; Row 3: RP Voting using ±1 random matrices with i.i.d entries; Row 4: RP Voting using the sparse $\{-1, 0, +1\}$ random matrices from Achlioptas (2003).

8
Kernel FLD

Summary The previous chapters 5 and 7 have treated high-dimensional classification by analyzing the performance of FLD when working with random projections of the original dataspace. That is, we considered FLD classification in spaces created by non-adaptive linear transformation of the original data via random matrices with i.i.d zero-mean Gaussian entries. A different setting in which randomized dimensionality reduction takes place as an intrinsic part of the algorithm is in the application of kernel methods to classification; for these algorithms one uses linear projection of the data, represented by features in a very high dimensional or infinite dimensional Hilbert space, $\mathcal{H}$, onto the finite dimensional subspace spanned by the training data – called the 'feature space'.

Although this setting is conceptually and mathematically quite different from the randomly projected settings we treated in our earlier chapters, since the principal aim of mapping the data to feature representations in the Hilbert space is to tackle problems where the original data are not linearly separable, we shall see that one can adapt some of the tools and approaches we developed earlier to analyze the generalization performance of the popular Kernel Fisher Linear Discriminant (KFLD) classifier. This is because the KFLD algorithm is mathematically equivalent to a regularized FLD algorithm operating in the feature space, and the role of the random projection matrix is taken by the operator which projects $\mathcal{H}$ onto the feature space.

In this chapter we derive a non-trivial, non-asymptotic upper bound on the classification error of KFLD under the assumption (motivated by theoretical (Dasgupta et al., 2012; Diaconis & Freedman, 1984; Huang et al., 2005) and empirical (Huang et al., 2005) results) that the kernel-induced feature space is a Gaussian Hilbert space – that is, a Hilbert space equipped with a Gaussian probability measure.

8.1 Preliminaries

Kernel Fisher Linear Discriminant (KFLD), first proposed by (Mika et al., 2002), is a generalization to feature space of Fisher's Linear Discriminant (FLD) classifier.
We have discussed the canonical FLD classifier in chapter 4 and the performance of Fisher Linear Discriminant (FLD) in finite-dimensional data space has been well-studied elsewhere in the literature in the asymptotic regime; for example, in the two class setting under the assumption of Gaussian classes the exact error is given in (Bickel & Levina, 2004; Pattison & Gossink, 1999). KFLD has a similar form – when the training observations used to construct the classifier are points in a feature space, then the resulting classifier is KFLD (Mika et al., 2002), however this classifier presents specific technical challenges to deriving generalization bounds that are not present in the data space setting. The work of (Mika et al., 2002) gives a full account of these but, in particular, the kernel induced space in which the classification is carried out need not be finite-dimensional, and even if it is finite dimensional the sample covariance matrix is always singular. Furthermore, since the dimensionality of the feature space in the finite sample setting is of the order of the number of training examples it seems that any bound which accurately reflects the behaviour of the classifier should be dimension-free.

Previous attempts at analyzing KFLD by Mika (2002) approached the analysis from the starting point of the KFLD objective function and its algorithmic solution as an eigenproblem, and tried to quantify the error of the eigenvector estimates. Unfortunately these leave open the question of the generalization error of the KFLD classifier. Diethe et al. (2009) developed a generalization error bound for a sparse version of KFLD. However, their bounds are rather loose and they assume that the induced distribution of the feature-mapped data has bounded support.
In a different vein, theoretical analysis by Huang et al. (2005), which draws on the work of Diaconis & Freedman (1984), focuses on justifying an interesting empirical observation, namely that data mapped in the feature space tend to have a Gaussian distribution. The work of Diaconis & Freedman (1984) showed that, under conditions, high-dimensional data projected onto a line converges in distribution to a univariate Gaussian. More recent theoretical work by Dasgupta et al. (2012) extends those findings to data projections onto spaces of dimension $k > 1$, and shows that the convergence in distribution of high-dimensional data projected to a lower dimensional space to a *multivariate* Gaussian is in fact quite a general phenomenon.
KFLD is a well-performing and popular classifier, yet very little is known about its generalization guarantees – in this chapter we take steps to improve this situation by deriving a bound on the generalization error of KFLD which, under mild assumptions, holds with high probability for any training set of a given size. Our bound is always non-trivial (less than 1), and is given in terms of quantities in the full Hilbert space.
A key term in our bound turns out to be the distance between the class mean functions scaled by the largest eigenvalue of the covariance operator. Since with a suitable kernel choice (any universal kernel, e.g. the radial basis kernel) there is a one-to-one mapping between a data density function in the input space and a mean function in the feature space (Smola et al., 2007), it follows that the classes are always separated and good

generalization can be achieved unless the densities of the two classes coincide in the input space.

Furthermore, although given in the context of functional data, our bound also applies to FLD in finite fixed dimensional settings as a special case, and exhibits the natural properties that it becomes tighter (i) as the number of training examples increases, (ii) when the classes are balanced, (iii) when the sample covariance is a good estimate of the true covariance, and (iv) as the separation of the classes increases.

The structure of the remainder of the chapter is as follows: We briefly recall the classification problem,introduce notation, and describe the problem setting. We then give the generalization error of KFLD when the training set is fixed under the assumption of Gaussian classes in the feature space. Next we give high probability guarantees on the generalization error of KFLD for any training set of size N. Finally we discuss our findings and indicate some possible future directions for this work.

8.1.1 The classification problem

Recall that in a classification problem we observe N examples of labelled training data $\mathcal{T}_N = \{(z_i, y_i)\}_{i=1}^N$ where $(z_i, y_i) \overset{i.i.d}{\sim} \mathcal{D}_{z,y}$. For a given class of functions $\mathcal{F}$, our goal is to learn from $\mathcal{T}_N$ the function $\hat{f} \in \mathcal{F}$ with the lowest possible generalization error in terms of some loss function $\mathcal{L}$. That is, find $\hat{f}$ such that $\mathcal{L}(\hat{f}) = \min_{f \in \mathcal{F}} \mathrm{E}_{z_q, y_q}[\mathcal{L}(f)]$, where $(z_q, y_q) \sim \mathcal{D}_{z,y}$ is a query point drawn from a distribution over some *input space*, I. As in the previous chapters we use the $(0,1)$-loss, $\mathcal{L}_{(0,1)}$, as our measure of performance here.
In the setting we consider here, the class of functions $\mathcal{F}$ consists of instantiations of KFLD learned from points in a feature space. That is, the training observations are functions of the original data mapped to a feature space $\mathcal{H}_N \subseteq \mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space[1], via a kernel mapping ϕ:

$$\phi : I \longrightarrow \mathcal{H} \tag{8.1.1}$$

$$z \longmapsto \phi(z) \tag{8.1.2}$$

Since the space $\mathcal{H} := \phi(I)$ is a separable Hilbert space, there exists an isometric isomorphism (that is, a bijection which preserves norms) between $\mathcal{H}$ and ℓ_2. Hence, without loss of generality we can work in ℓ_2.

8.1.2 Notation

We now introduce some new notation, specific to the setting we consider in this chapter. As noted in the previous section, if (z_i, y_i) is a training observation from the original data domain Z with label y_i, then $(\phi(z_i), y_i)$ is the corresponding training example in the feature space induced by the feature mapping ϕ. To keep our notation compact, we will write x_i for $\phi(z_i)$ from now on, and $\mathcal{D}_{x,y}$ for the induced probability distribution over $\mathcal{H}$. For convenience we assume that the x_i are linearly independent since otherwise very similar arguments to those we will present here still go through when $\dim(\langle x_i \rangle_{i=1}^N) <$

[1] A Hilbert space is a (finite- or infinite-dimensional) vector space equipped with an inner product. A Hilbert space is called separable if and only if it admits a countable orthonormal basis.

N. The feature space $\mathcal{H}_N$ is then the N-dimensional subspace of ℓ_2 spanned by the observations x_i and, as a consequence of the representer theorem, this is where the KFLD algorithm operates. We denote by μ_y the true mean element in $\mathcal{H}$ of the class y, and by $\Sigma = \Sigma_0 = \Sigma_1$ the (shared) true covariance of the classes, where Σ is a positive-definite trace-class covariance operator (i.e. such that the projection of Σ to any subspace of $\mathcal{H}$ is invertible and $\text{Tr}(\Sigma) < \infty$). These properties are a technical requirement if $\mathcal{D}_{x|y}$ is to be a non-degenerate probability distribution over $\mathcal{H}$. We indicate estimated quantities by adding a hat: $\hat{\mu}_y$, $\hat{\Sigma}$.

Note that linear independence of the x_i implies (Schölkopf & Smola, 2002) that we always have $\|\hat{\mu}_1 - \hat{\mu}_0\| > 0$ in our setting.

We use the subscript N to indicate when an object of interest is restricted to $\mathcal{H}_N$; in particular we will denote by x_N the projection of the vector $x \in \mathcal{H}$ onto the subspace $\mathcal{H}_N$ spanned by the observations, that is if $X \in \mathcal{M}_{\infty \times N}$ is the matrix with the x_i as columns and $P = (X^T X)^{-\frac{1}{2}} X^T$ then $x_N = Px$, $\Sigma_N = P \Sigma P^T$, and so on. Note that P is not a canonical projection operator, rather it is a 'projection' in precisely the same sense that the random projection matrices R considered earlier are - i.e. Px is the orthogonal projection of x onto the subspace spanned by the rows of P.

We assume, as the KFLD model implicitly does, that a probability distribution exists over the x_i and we consider the two-class setting only, since upper bounds on the multi-class generalization error can be obtained via lemma 16 of chapter 4.

The set of training observations for KFLD as treated here is therefore: $\mathcal{T}_N = \{(x_i, y_i) : (x_i, y_i) \sim \mathcal{D}_{x,y}\}_{i=1}^N$, and we bound the probability that an arbitrary query point x_q with its true class label y_q unknown is misclassified by the learned classifier. Specifically, with high probability we upper bound the classification error of KFLD under the assumption that $\mathcal{D}_{x|y} \equiv \mathcal{N}(\mu_y, \Sigma)$ in a separable Hilbert space, $\mathcal{H}$, (here taken to be ℓ_2 equipped with Gaussian probability measure over Borel sets) with π_y the prior probability that x_q belongs to class y. We further denote by N_0 and N_1 the number of training observations in the two classes. We will assume throughout this chapter that in $\mathcal{T}_N$ we have N_0 and N_1 both greater than 0, which is the case of practical interest for a classification task.

8.2 Results

We assume functional data (Ramsay, 1997), namely that the original data observations have been mapped into a feature space by some (linear or non-linear) function ϕ, and that this mapping imposes a Gaussian distribution on the features in each class. There are several reasons why we might consider that this assumption is not too restrictive.

Firstly, Huang et al. (2005) have shown that most low-dimensional projections, i.e. from $\mathcal{H}$ onto $\mathcal{H}_N$, are approximately Gaussian when the mapping to the feature space is a proper kernel. This phenomenon is a consequence of central limit like behaviour and is very general.

Furthermore, our assumption allows us to potentially extend our work to a more general setting than is often considered in theoretical treatments of kernel learning, where boundedness of random variables is frequently assumed.

In order to bound the generalization error of KFLD we work with the decision function, assuming access to the feature space. We find this a more convenient formalism in which to derive generalization guarantees than the formulation of this classifier as an optimization problem (the representation via the kernel trick that is required for algorithmic implementation).

Without loss of generality we consider the infinite-dimensional Hilbert space ℓ_2 and we work in the feature space, namely the space spanned by the features from the training set or, equivalently, the orthogonal projection of ℓ_2 on to the span of the training features which we denote $\mathcal{H}_N$. For convenience we will assume the features span the first N dimensions of ℓ_2 (since otherwise we can rotate ℓ_2 so that this is the case).

Our starting point is the decision function for KFLD which, for a training set of size N, is (Herbrich, 2002):

$$\hat{f}(x_q) := \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} \left(x_q - \frac{(\hat{\mu}_0 + \hat{\mu}_1)}{2}\right)_N > 0\right\}$$

where $\hat{\mu}_y = \frac{1}{N_y}\sum_{i=1}^{N_y} x_i$ and the training observations x_i in the summation all have label y, and $\hat{\Sigma}_N$ is a (regularized) sample covariance matrix with its precise form depending on the choice of regularization scheme. Recall that the subscript N indicates that these quantities are orthogonally mapped from $\mathcal{H}$ in $\mathcal{H}_N$. Then the generalization error w.r.t $\mathcal{L}_{(0,1)}$ is given by:

$$\begin{aligned} &\mathrm{E}_{x_q,y_q}\left[\mathcal{L}_{(0,1)}\left(\hat{f}(x_q), y_q\right)\right] \\ &= \mathrm{Pr}_{x_q,y_q}\left\{\hat{f}(x_q) \neq y_q\right\} \end{aligned} \tag{8.2.1}$$

and we upper bound this probability. To achieve this, we first develop a very general bound on sub-exponential random variables that will be one of our main tools, and may also be of independent interest.

8.2.1 Dimension-free bound on (sub)-exponential random variables

Lemma 23

Let $X = (X_1, X_2, X_3, \ldots)$ be a sequence of Gaussian random variables in the Hilbert space $\mathcal{H}$ with mean vector $E[X] = \mu$ and covariance operator Σ, such that the ℓ_2 norm: $\|E[X]\| = \|\mu\| < +\infty$ and Σ is trace-class: $Tr(\Sigma) < +\infty$. Let $\epsilon > 0$. Then:

$$\begin{aligned} &Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(Tr(\Sigma) + \|\mu\|^2\right)\right\} \\ &\leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon} - 1\right)^2\right) \end{aligned} \tag{8.2.2}$$

Furthermore, if $\epsilon \in (0,1)$:

$$\begin{aligned} &Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(Tr(\Sigma)+\|\mu\|^2\right)\right\} \\ &\leqslant \exp\left(-\frac{Tr(\Sigma)+\|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon}-1\right)^2\right) \end{aligned} \tag{8.2.3}$$

Lemma 23 is an extension to Hilbert space of classical finite dimensional results e.g. (Dasgupta, 2000b; Laurent & Massart, 2000). The proof of Lemma 23 uses a combination of elementary techniques and is given in the Appendix.
The proof makes use of the moment generating function (m.g.f.) of (non-central) χ^2 variables, hence the obtained bounds hold for distributions whose m.g.f. is dominated by that of the χ^2 – these are called sub-exponential distributions.

We note that the Bernstein-type bounds we give in lemma 23 are able to exploit variance information and hence avoid the worst-case approach commonly employed in conjunction with bounded random variables. The latter would lead to the data diameter appearing in the bound, e.g. as in (Diethe et al., 2009; Rahimi & Recht, 2008b; Shawe-Taylor & Cristianini, 2003). In particular, our bounds have $\sqrt{\text{Tr}(\Sigma)}$ in this role, which can be considerably smaller than the data diameter, and moreover do not require the boundedness assumption.

8.2.2 Bound on generalization error of KFLD when the training set is fixed

We will use the following bound on the generalization error of KFLD in the feature space $\mathcal{H}_N$.

In the KFLD setting, $\hat{\Sigma}^{-1}$ (and Σ) are operators, so the notation $\hat{\Sigma}^{-1}$ will mean the operator inverse, i.e. inverse on its range. For KFLD it is always the case that the estimated covariance without regularization is singular (it has rank at most $N-2$) and so if we choose to regularize $\hat{\Sigma}$ on the subspace $\mathcal{H}_N$, as is usual in practice, then this regularization ensures that $\hat{\Sigma}_N$ has rank N and $\hat{\Sigma}_N^{-1}$ denotes the usual matrix inverse.

Lemma 24
Let $x_i \sim \sum_{y=0}^{1}\pi_y\mathcal{N}(\mu_y,\Sigma)$, and assume that some suitable regularization scheme ensures that the rank of $\hat{\Sigma}_N$ is N, then the error of KFLD in eq.(8.2.1) is given by:

$$\begin{aligned} &\pi_0\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_0)_N}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}(\hat{\mu}_1-\hat{\mu}_0)_N}}\right)+ \\ &\qquad\pi_1\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_0-\hat{\mu}_1)_N^T\hat{\Sigma}_N^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_1)_N}{\sqrt{(\hat{\mu}_0-\hat{\mu}_1)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}(\hat{\mu}_0-\hat{\mu}_1)_N}}\right) \end{aligned} \tag{8.2.4}$$

Where Φ is the c.d.f of the standard Gaussian distribution.

The proof of lemma 24 is much the same as that for theorem 5 given in chapter 4, noting that for KFLD the decision of which label to assign to a query point x_q is made with respect to the projection of x_q onto $\mathcal{H}_N$; for completeness we give it in the Appendix.

In what follows, we bound the deviation of the quantities appearing in (8.2.4) from their expectations with high probability with respect to the training set, $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$.

8.2.3 Main Result: Tail bound on generalization error of KFLD

We will now estimate the various quantities in (8.2.4) with high probability over all training sets of size $N = N_0 + N_1$. This will ultimately enable us, with confidence $1 - \delta$ (where $\delta(\Sigma, \epsilon_1, \epsilon_2, \epsilon_3, N_0, N_1)$ is an exponentially small quantity), to bound the effect of the parameter estimates with quantities depending on the true parameters and the sample size. We will assume for concreteness that the query point x_q should be assigned the label 0, which entails no loss of generality as similar arguments apply when the label should be 1.

We begin by decomposing the bilinear form $\beta = (\hat{\mu}_1 - \hat{\mu}_0)_N^T \, \hat{\Sigma}_N^{-1} \, (\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)_N$ in the numerator of (8.2.4) as follows:

$$\begin{aligned} \beta = & \quad (\hat{\mu}_1 - \hat{\mu}_0)_N^T \, \hat{\Sigma}_N^{-1} \, (\hat{\mu}_1 - \hat{\mu}_0)_N \\ \ldots & \quad +2 \, (\hat{\mu}_0 - \mu_0)_N^T \, \hat{\Sigma}_N^{-1} \, (\hat{\mu}_1 - \hat{\mu}_0)_N \end{aligned} \tag{8.2.5}$$

Using the decomposition (8.2.5) we can rewrite the first term of lemma 24 in the following form:

$$\Phi\left(-\frac{1}{2}(A - B)\right)$$

Where:

$$A = \frac{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \, \hat{\Sigma}_N^{-1} \, (\hat{\mu}_1 - \hat{\mu}_0)_N}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} \Sigma_N \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}} \tag{8.2.6}$$

is the term responsible for the *estimated error*, and:

$$B = \frac{2 \, (\mu_0 - \hat{\mu}_0)_N^T \, \hat{\Sigma}_N^{-1} \, (\hat{\mu}_1 - \hat{\mu}_0)_N}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} \Sigma_N \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}} \tag{8.2.7}$$

is the term responsible for the *estimation error*. We will lower bound A and upper bound B to bound the whole term from above.

Lower-bounding the term A

We will make use of the Kantorovich inequality, lemma 8, with the choice of positive definite $Q = \hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}}$ and we can then lower bound A with:

$$\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \, \Sigma_N^{-1} \, (\hat{\mu}_1 - \hat{\mu}_0)_N} \cdot \frac{2\sqrt{\lambda_{\min}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}}) \lambda_{\max}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}})}}{\lambda_{\min}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}}) + \lambda_{\max}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}})} \tag{8.2.8}$$

Note that by positive definiteness of $\hat{\Sigma}_N, \Sigma_N$ and the arithmetic-geometric mean inequality we have:

$$1 \geqslant \frac{2\sqrt{\lambda_{\min}(\hat{\Sigma}_N^{-\frac{1}{2}}\Sigma_N\hat{\Sigma}_N^{-\frac{1}{2}})\lambda_{\max}(\hat{\Sigma}_N^{-\frac{1}{2}}\Sigma_N\hat{\Sigma}_N^{-\frac{1}{2}})}}{\lambda_{\min}(\hat{\Sigma}_N^{-\frac{1}{2}}\Sigma_N\hat{\Sigma}_N^{-\frac{1}{2}}) + \lambda_{\max}(\hat{\Sigma}_N^{-\frac{1}{2}}\Sigma_N\hat{\Sigma}_N^{-\frac{1}{2}})} > 0$$

For convenience we now rewrite (8.2.8) in terms of the condition number, κ, of $\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}$ using the identity for square invertible matrices $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \kappa(A^{-1})$ to give:

$$\|\Sigma_N^{-\frac{1}{2}}(\hat{\mu}_1 - \hat{\mu}_0)_N\| \frac{2\sqrt{\kappa(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}})}}{1 + \kappa(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}})} \tag{8.2.9}$$

Now applying Rayleigh quotient, lemma 5, to the norm above we see:

$$\begin{aligned} \|\Sigma_N^{-\frac{1}{2}}(\hat{\mu}_1 - \hat{\mu}_0)_N\| &\geqslant \frac{\|(\hat{\mu}_1 - \hat{\mu}_0)_N\|}{\sqrt{\lambda_{\max}(\Sigma_N)}} = \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|}{\sqrt{\lambda_{\max}(\Sigma_N)}} \\ &\geqslant \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|}{\sqrt{\lambda_{\max}(\Sigma)}} \end{aligned} \tag{8.2.10}$$

where the equality in the chain (8.2.10) follows because the mean estimates lie in the span of the observations $\mathcal{H}_N$, and the final inequality follows from the fact that $\lambda_{\max}(\Sigma_N) = \lambda_{\max}(P\Sigma P^T) = \lambda_{\max}(P^T P\Sigma) \leqslant \lambda_{\max}(P^T P)\lambda_{\max}(\Sigma) = 1 \cdot \lambda_{\max}(\Sigma)$ where the penultimate step uses lemma 9 and the last equality holds since $P^T P$ is a projection operator.

Next, since $\hat{\mu}_1$ and $\hat{\mu}_0$ are independent with $\hat{\mu}_y \sim \mathcal{N}(\mu_y, \Sigma/N_y)$ we have $(\hat{\mu}_1 - \hat{\mu}_0) \sim \mathcal{N}(\mu_1 - \mu_0, \Sigma/N_1 + \Sigma/N_0) = \mathcal{N}(\mu_1 - \mu_0, (N_0 + N_1)\Sigma/N_0N_1) = \mathcal{N}(\mu_1 - \mu_0, N\Sigma/N_0N_1)$. Applying lemma 23 (7.3.2) to $\|\hat{\mu}_1 - \hat{\mu}_0\|$ we lower bound this as:

$$\|\hat{\mu}_1 - \hat{\mu}_0\| \geqslant \sqrt{(1-\epsilon)\left(\frac{N}{N_0N_1}\mathrm{Tr}\,(\Sigma) + \|\mu_1 - \mu_0\|^2\right)} \tag{8.2.11}$$

with probability at least:

$$1 - \exp\left(-\frac{\mathrm{Tr}(\Sigma) + \|\mu_1 - \mu_0\|^2 \frac{N_0N_1}{N}}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon} - 1\right)^2\right) \tag{8.2.12}$$

To complete the bounding of the term A, we denote $g(a) := \frac{\sqrt{a}}{1+a}$, and observe that this is a monotonic decreasing function on $[1, \infty)$. So, replacing a with the condition number $\kappa(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}) \in [1, \infty)$ we see that upper bounding the condition number allows us to lower bound g. Hence, it remains to estimate the least and greatest eigenvalues of $\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}$ – this we do in the next subsection (section 8.2.3), and the resulting upper bound on the condition number of this matrix we denote by $\bar{\kappa}(\epsilon)$ (see eq. (8.2.19)).

Now, replacing, the term A is lower bounded w.h.p by:

$$A \geqslant 2g(\bar{\kappa}(\epsilon))\sqrt{(1-\epsilon)\left(\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)}+\frac{N}{N_0N_1}\frac{\mathrm{Tr}\,(\Sigma)}{\lambda_{\max}(\Sigma)}\right)} \tag{8.2.13}$$

The first summand under the square root in (8.2.13), represents a bound on the negative log of the Bayes error of the classifier. It is governed by the scaled distance between the true mean functions in $\mathcal{H}$ and the larger this distance is the better the performance guarantee. The second summand represents the extent of overestimation of this relative distance – that is the extent to which the estimated error underestimates the true error due to the use of estimated parameters in the place of the true ones. We see this term is largest when the number of training points is smallest and when the 'effective dimension' of the true data density, $\mathrm{Tr}(\Sigma)/\lambda_{\max}(\Sigma)$, is largest. The optimistic misestimation of the true error by the estimated error term will of course be countered by the other terms in the overall error decomposition, namely those that quantify the quality of the parameter estimates κ and B.

Upper-bounding $\kappa(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}})$

Because in KFLD we estimate an $N\times N$ covariance matrix in an N-dimensional subspace of $\mathcal{H}$, and the sample means are linear combinations of the labelled features, the scatter matrices $\sum_{i=1}^{N_y}(x_i-\hat{\mu}_y)(x_i-\hat{\mu}_y)^T$ have rank at most N_y-1 and so the unregularized covariance estimate has rank at most $N-2$. Since the sample covariance matrix is inverted in building the classifier, one must deal with the fact that this matrix is singular. We will hence assume that some suitable form of regularization is employed that ensures that $\hat{\Sigma}_N$ is full rank, and this is indeed what allowed us to write $\kappa(\hat{\Sigma}_N^{-\frac{1}{2}}\Sigma_N\hat{\Sigma}_N^{-\frac{1}{2}}) = \kappa(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}})$ earlier in eq.(8.2.9).

The most common form of regularizing the covariance estimate in the feature space is:

$$X\hat{\Sigma}_{UR}X^T+\alpha C \tag{8.2.14}$$

where α is the regularization parameter, $\hat{\Sigma}_{UR}$ is the unregularized estimate (e.g. the maximum likelihood estimate), which is nothing but the within-class scatter matrix (as defined in e.g. S 4.10 of Duda et al., 2000), normalized by the total number of training points, i.e.:

$$\hat{\Sigma}_{UR} = \frac{1}{N}\sum_{y=0}^{1}\sum_{i=1}^{N_y}(x_i-\hat{\mu}_y)(x_i-\hat{\mu}_y)^T \tag{8.2.15}$$

The regularization term may be chosen as $C=I_N$, or $C=XX^T$. The former is more common, the latter is proposed by Centeno & Lawrence (2006) by drawing a parallel between KFLD and a Bayesian reformulation of it, which was also demonstrated to have superior performance. It is interesting to note that this latter option corresponds to regularizing with αI_N after *orthogonal* projection (i.e. projection by P rather than X) into the N-dimensional linear span of the training points. Indeed, using our earlier notation:

$$\begin{aligned}
&\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}\\
=\ &(P\Sigma P^T)^{-\frac{1}{2}}(P\hat{\Sigma}_{UR}P^T+\alpha I_N)(P\Sigma P^T)^{-\frac{1}{2}}\\
=\ &(X\Sigma X^T)^{-\frac{1}{2}}(X\hat{\Sigma}_{UR}X^T+\alpha XX^T)(X\Sigma X^T)^{-\frac{1}{2}}
\end{aligned}$$

after cancellation of the terms $(XX^T)^{-1/2}$, and we recognise XX^T in place of C. In the following we will employ this regularization choice to have $\hat{\Sigma}_N \equiv P\hat{\Sigma}_{UR}P^T + \alpha I_N$, noting that the alternative $C = I_N$ may be analysed in a similar way.

Then $\lambda_{\max}(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}})$ is equal to:

$$\lambda_{\max}\left(\frac{1}{N}\sum_{y=0}^{1}(P\Sigma P^T)^{-\frac{1}{2}}\sum_{i=1}^{N_y}P(x_i-\hat{\mu}_y)(x_i-\hat{\mu}_y)^TP^T(P\Sigma P^T)^{-\frac{1}{2}}\right.$$
$$\left.+\alpha(P\Sigma P^T)^{-1}\right)$$

Now, observe that for each class:

$$S_y := (P\Sigma P^T)^{-\frac{1}{2}}\sum_{i=1}^{N_y}P(x_i-\hat{\mu}_y)(x_i-\hat{\mu}_y)^TP^T(P\Sigma P^T)^{-\frac{1}{2}}$$

has an N-dimensional singular Wishart distribution (Srivastava, 2003) with $N_y - 1$ degrees of freedom, $\mathcal{W}_N(N_y - 1, I_N)$. Hence $S_0 + S_1$ is Wishart with $N-2$ d.f., $S_0 + S_1 \sim \mathcal{W}_N(N-2, I_N)$. This means that there exists a matrix $Z \in \mathcal{M}_{N\times(N-2)}$ with standard normal entries s.t. ZZ^T has the same distribution as $S_0 + S_1$.

Now, to bound the scatter matrix terms we use the following high probability bound on the singular values of Z:

Lemma 25

Singular values of Gaussian matrices. ((Rudelson & Vershynin, 2010), Eq. (2.3)) *Let A be an $n \times N$ matrix with standard normal entries, and denote by $s_{\min}(A)$, $s_{\max}(A)$ its least and greatest singular values. Then:*

$$\begin{aligned}Pr\{\sqrt{N}-\sqrt{n}-\epsilon \leqslant s_{\min}(A) \leqslant s_{\max}(A) \leqslant \sqrt{N}+\sqrt{n}+\epsilon\}\\ \geqslant 1-2e^{-\epsilon^2/2},\quad \forall\epsilon>0\end{aligned}$$

We can use Weyl's inequality, which gives the crude bound $\lambda_{\max}(A+B) \leqslant \lambda_{\max}(A) + \lambda_{\max}(B)$, to decouple the within class scatters and the regularization term. Then we use the bounds on the extreme singular values of Gaussian matrices given in lemma 25 to bound the eigenvalues of the terms of the unregularized covariance estimate. Hence we have:

$$\begin{aligned}&\lambda_{\max}\left(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}\right) \qquad (8.2.16)\\ \leqslant\ &\left(1+\sqrt{\tfrac{N-2}{N}}+\tfrac{\epsilon}{\sqrt{N}}\right)^2+\alpha/\lambda_{\min}(\Sigma_N)\end{aligned}$$

with probability at least $1 - e^{-\epsilon^2/2}$.

The smallest eigenvalue is governed by the regularization term, and may be lower bounded as:

$$\begin{aligned}\lambda_{\min}(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}) &= \lambda_{\min}(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_{UR}\Sigma_N^{-\frac{1}{2}} + \alpha(\Sigma_N^{-1})) \\ &\geqslant \lambda_{\min}(\alpha(\Sigma_N^{-1})) = \alpha/\lambda_{\max}(\Sigma_N) \qquad (8.2.17) \\ &\geqslant \alpha/\lambda_{\max}(\Sigma) \qquad (8.2.18)\end{aligned}$$

by using the other side of Weyl's inequality with $\lambda_{\min}(A+B) \geqslant \lambda_{\min}(A) + \lambda_{\min}(B)$ and noting that the scatter matrix is singular, $\lambda_{\min}(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_{UR}\Sigma_N^{-\frac{1}{2}}) = 0$.

Putting these together, the condition number is upper bounded with probability at least $1 - e^{-\epsilon^2/2}$ by:

$$\begin{aligned}&\kappa\left(\Sigma_N^{-\frac{1}{2}}\hat{\Sigma}_N\Sigma_N^{-\frac{1}{2}}\right) \\ &\leqslant \tfrac{\lambda_{\max}(\Sigma)}{\alpha}\left(1 + \sqrt{\tfrac{N-2}{N}} + \tfrac{\epsilon}{\sqrt{N}}\right)^2 + \kappa(\Sigma_N) \\ &=: \bar{\kappa}(\epsilon) \qquad (8.2.19)\end{aligned}$$

The first term in eq. (8.2.19) is independent of the data. The last term $\kappa(\Sigma_N)$, however, is the condition number of the projection of the true covariance onto the span of the training points. While $\lambda_{\max}(\Sigma_N) \leqslant \lambda_{\max}(\Sigma)$ for any projection P, removing the data-dependence of $\lambda_{\min}(\Sigma_N)$ seems to be tricky in a general setting — clearly if the condition number of Σ is finite then we can write $\kappa(\Sigma_N) \leqslant \kappa(\Sigma)$ — however finiteness of $\kappa(\Sigma)$ is not necessary for $\kappa(\Sigma_N)$ to be finite. **Comments.** We note that (for either regularizer) there is a trade-off regarding the regularization parameter α: To minimize the condition number α needs to be small to decrease the $\lambda_{\max}$ term, while it has to be large to increase the $\lambda_{\min}$ term. This is indeed how we would expect the classifier error to behave w.r.t the regularization parameter.

We also observe that, if we were to ridge regularize with the choice $C = I_N$ then we would have $\lambda_{\max}$ and $\lambda_{\min}$ of the matrix $X\Sigma X^T$ instead of those of $P\Sigma P^T$ in eq. (8.2.16) and eq. (8.2.17) respectively. These extreme eigenvalues can be more spread out since XX^T is less well-conditioned than $PP^T = I_N$ the identity, which suggests support for the findings of Centeno & Lawrence (2006) that regularization with the kernel matrix can reduce the generalization error of KFLD.

Upper-bounding the term B

To upper bound B, first we multiply $\hat{\Sigma}_N^{-1}$ on the left by the identity to rewrite and bound equation (8.2.7) as:

$$\begin{aligned} B &= \frac{2\,(\mu_0 - \hat{\mu}_0)_N^T\,\Sigma_N^{-\frac{1}{2}}\Sigma_N^{\frac{1}{2}}\hat{\Sigma}_N^{-1}\,(\hat{\mu}_1 - \hat{\mu}_0)_N}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}(\hat{\mu}_1 - \hat{\mu}_0)_N}} \\ &\leqslant \frac{2\|\Sigma_N^{-\frac{1}{2}}(\mu_0 - \hat{\mu}_0)_N\|\|\Sigma_N^{\frac{1}{2}}\hat{\Sigma}_N^{-1}(\hat{\mu}_1 - \hat{\mu}_0)_N\|}{\|\Sigma_N^{\frac{1}{2}}\hat{\Sigma}_N^{-1}(\hat{\mu}_1 - \hat{\mu}_0)_N\|} \\ &= 2\|\Sigma_N^{-\frac{1}{2}}(\mu_0 - \hat{\mu}_0)_N\| \end{aligned} \tag{8.2.20}$$

using Cauchy-Schwarz in the numerator.
Then, using lemma 23 we further upper bound (8.2.20) with:

$$2\sqrt{(1+\epsilon)\cdot \mathrm{Tr}\,(I_N/N_0)} = 2\sqrt{(1+\epsilon)N/N_0} \tag{8.2.21}$$

with probability $\geqslant 1 - \exp\left(-\frac{1}{2}N\cdot(\sqrt{1+\epsilon}-1)^2\right)$.

Putting everything together

Now we collate the results proved so far to arrive at our final bound. Our chain of arguments shows that, $\forall \epsilon_1, \epsilon_2 \in (0,1), \forall \epsilon_3 > 0$ the expression $\Phi\left(-\frac{1}{2}(A-B)\right)$ is bounded above, with probability $1-\delta_0$ by:

$$\Phi\left(-\left[g(\bar{\kappa}(\epsilon_2))\sqrt{(1-\epsilon_1)\left(\tfrac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)} + \tfrac{N}{N_0N_1}\tfrac{\mathrm{Tr}(\Sigma)}{\lambda_{\max}(\Sigma)}\right)} \right.\right. \\ \left.\left. - \sqrt{\bar{\kappa}(\epsilon_2)}\sqrt{(1+\epsilon_3)N/N_0}\right]\right)$$

where $\bar{\kappa}(\epsilon_2)$ is given by eq. (8.2.19), and the risk probability $\delta_0 = \delta_0(\Sigma, \epsilon_1, \epsilon_2, \epsilon_3, N_0, N_1)$ is, by union bound,

$$\delta_0 \leqslant \exp\left(-\frac{1}{2}N\cdot(\sqrt{1+\epsilon_3}-1)^2\right) + \exp\left(-\epsilon_2^2/2\right) \\ \ldots + \exp\left(-\frac{\mathrm{Tr}(\Sigma) + \|\mu_1-\mu_0\|^2\frac{N_0N_1}{N}}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon_1}-1\right)^2\right)$$

Repeating the argument for the case when the query point has label $y_q = 1$ and applying the law of total probability we finally obtain our upper bound on the misclassification error of KFLD. Note that in doing so, the probability bounds employed in bounding the terms A and κ are re-used, so both sides of the final bound will hold simultaneously w.p. at least $1-\delta(\Sigma, \epsilon_1, \epsilon_2, \epsilon_3, N_0, N_1) = 1 - (\delta_0 + \exp(-\frac{1}{2}N\cdot(\sqrt{1+\epsilon_3}-1)^2))$.

For the sake of a better interpretability, we may rearrange this result by suitably choosing $\epsilon_1, \epsilon_2, \epsilon_3$. In particular, putting all four terms of the probability bound $\delta(\Sigma, \epsilon_1, \epsilon_2, \epsilon_3, N_0, N_1)$ to $\delta/4$, solving for $\epsilon_1, \epsilon_2, \epsilon_3$ and replacing, yields after some straightforward algebra the following equivalent formulation:

Theorem 14 *For any $\delta \in (0,1)$, the generalization error of KFLD in a Gaussian Hilbert space, eq.(8.2.4), is upper-bounded w.p. at least $1-\delta$ over the random choice of training set $\mathcal{T}_{N=N_0+N_1}\{(x,y)\}$ with Gaussian class-conditionals $x|y \sim \mathcal{N}(\mu_y,\Sigma)$, by:*

$$Pr_{x_q,y_q}\left\{\hat{f}(x_q) \neq y_q\right\} \leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-\left[\quad g(\bar{\kappa}(\epsilon_2)) \times \ldots\right.\right.$$
$$\left[\sqrt{\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)} + \frac{N_0+N_1}{N_0N_1}\frac{Tr(\Sigma)}{\lambda_{\max}(\Sigma)}} - \sqrt{\frac{2(N_0+N_1)}{N_0N_1}\log\frac{4}{\delta}}\right]_+$$
$$\left.\left.\ldots - \sqrt{\frac{N}{N_y}}\left(1+\sqrt{\frac{2}{N}\log\frac{4}{\delta}}\right)\right]\right)$$

where now $\bar{\kappa}(\epsilon_2)$ is given by replacing $\epsilon_2 := \sqrt{2\log\frac{4}{\delta}}$ in eq. (8.2.19).

We proved this probability bound conditional on any fixed value of $N_0 \in \{1,...,N-1\}$, therefore it also holds for a random N_0 over this set. Hence we can remove the conditioning on the value of N_0 by taking expectation w.r.t N_0 on both sides of the probability bound. Alternatively, we could increase δ in theorem 14 by $\pi_0^N + \pi_1^N$ to bound both N_0 and N_1 away from zero with high probability. We see that a key term in the bound is the scaled distance between the mean functions in the Hilbert space. Using the fact (Smola et al., 2007) that with a suitable kernel choice (any universal kernel) there is an injective mapping between a mean function in the Hilbert space and a class density function in the input space, the distance between the mean functions may be seen as representing a distance between the class-conditional density functions in the input space. This is never zero unless the two class densities coincide – consequently good generalization can be achieved unless the two classes have identical densities in the input space.

It is tempting to attempt to interpret the behaviour of our bound with respect to the sample size. However, we should point out that in a kernel setting the precise relation of the various error terms to the number of training points is much more complex than this level of analysis enables us to see. This is because both μ_y and Σ are functions of the sample size, e.g. due to the fact that the kernel width needs to be decreased as the sample size increases, and their precise relationship is not known. Therefore the bound in Theorem 1 is for a fixed N only.

However, it is instructive to assess this aspect of our bound by noting that it applies to non-kernel FLD as a special case. The only difference is that then $N \neq N_0 + N_1$ but instead N is the fixed dimensionality of the data and $M = M_0 + M_1$ is the sample size that can grow.

Corollary 4 (to Theorem 14)
Let the data be N-dimensional, and having Gaussian class-conditionals $x|y \sim \mathcal{N}(\mu_y,\Sigma)$. Denote by $\hat{h}$ the FLD classifier learned from training data. Then for any $\delta \in (0,1)$, and any training set of size $M = M_0 + M_1$, the generalization error of FLD in $\mathbb{R}^N$ is

upper-bounded w.p. $1-\delta$ *by the following:*

$$Pr_{x_q,y_q}\left\{\hat{h}(x_q) \neq y_q\right\} \leqslant \sum_{y=0}^{1}\pi_y\Phi\Bigg(-\Bigg[\ g(\bar{\kappa}(\epsilon_2)) \times \ldots$$

$$\left[\sqrt{\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)}+\frac{M}{M_0M_1}\frac{Tr(\Sigma)}{\lambda_{\max}(\Sigma)}}-\sqrt{\frac{2M}{M_0M_1}\log\frac{5}{\delta}}\right]_+$$

$$\ldots - \sqrt{\frac{N}{M_y}}\left(1+\sqrt{\frac{2}{N}\log\frac{5}{\delta}}\right)\Bigg]\Bigg) \qquad (8.2.22)$$

where $\bar{\kappa}(\epsilon_2)$ is as in Theorem 1 when the MLE of Σ, $\hat{\Sigma}_{UR}$ is singular (but regularized), and when $\hat{\Sigma}_{UR}$ is non-singular we can bound its minimum eigenvalue away from zero using Lemma 25, which yields the following tighter $\bar{\kappa}(\epsilon_2)$:

$$\kappa\left(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}\right) \leqslant \left(\frac{\sqrt{M-2}+\sqrt{N}+\epsilon}{\sqrt{M-2}-\sqrt{N}-\epsilon}\right)^2 =: \bar{\kappa}(\epsilon) \qquad (8.2.23)$$

with probability at least $1-2e^{-\epsilon^2/2}$. Hence in the latter case we will have $\epsilon_2 := \sqrt{2\log\frac{5}{\delta}}$ in (8.2.22).

More interpretation may be drawn from the bound in the finite dimensional setting in Corollary 1. The first thing to note is that $\text{Tr}(\Sigma)/\lambda_{\max}(\Sigma)$ becomes of the same order as N i.e. the dimensionality of the problem. (In fact it is not difficult to derive a version of the bound that actually contains N in place of $\text{Tr}(\Sigma)/\lambda_{\max}(\Sigma)$ in this setting. This would also have $(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$ in place of $\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)}$.) Then we see clearly how the term of A that is responsible for the optimistic distance estimate, of the form dimension/#points, gets countered by the reverse effect of the same form from B.

More importantly, the consistency of FLD follows from Corollary 1. Indeed, as the sample sizes M_0 and M_1 both increase, the condition number bound (8.2.23) converges to 1, and all the terms other than (an upper bound on) the Bayes error vanish in (8.2.22). Hence we may conclude that our bound behaves in a desirable natural way. We also note in both the kernel and non-kernel settings that, in addition to the good properties already mentioned, class balance makes the bound tighter as it should.

8.3 Summary and Discussion

We derived a dimension-free bound on the generalization error of KFLD which, to the best of our knowledge, is the first non-trivial bound for the standard KFLD model. This puts KFLD on a solid theoretical foundation and improves our understanding of the working of this classifier. In this work we assumed that the kernel-induced space is a Gaussian Hilbert space. Extension to Gaussian classes with different class-conditional covariances appears relatively straightforward using the results and tools already developed in chapters 4, 5 and 7. It also appears, modulo careful checking, that extending these results to give similar guarantees for the much larger family of subgaussian class-conditional distributions should be possible using tools such as those in (Rudelson & Vershynin, 2010; Vershynin, 2012). Further work is required to extend this analysis to a more detailed level, e.g. in order to determine the relationship between kernel parameters and the generalization error. It also seems plausible that, by letting the regularization parameter go to zero as the number of training examples increases to infinity, one could ultimately prove the consistency of KFLD.

9
Conclusions and Open Problems

9.1 Randomly-projected Fisher's Linear Discriminant

In chapter 5 we quantified the cost of dimensionality reduction using random projection on the performance of a single Fisher's Linear Discriminant classifier.
Our main contribution in that chapter was to show that, at least for FLD, uniform approximate preservation of data geometry through the JLL, and the consequent unnatural behaviour w.r.t the number of observations in bounds utilising it, is too strong a requirement in order to give guarantees on classification performance. In particular, on average one can obtain control over the generalization error of FLD even if the data are randomly projected to a subspace with dimensionality, k, which is only logarithmic in the number of classes, rather than logarithmic in the number of data points as was required in earlier JLL-based bounds. More generally we can conclude that using the JLL to uniformly approximately preserve all pairwise distances between data points will not always be necessary when the data are to be used for classification - preserving a subset of *important distances* (the number of which will depend on the classification regime) may be sufficient. For example, consider m-class data which is uniformly 2ϵ-separable (i.e. the margin between any two convex hulls enclosing the classes is at least 2ϵ) and the SVM classifier – intuition says that preserving the m convex hulls bounding the classes would be sufficient to control the generalization error. Preserving these convex hulls requires preserving at most $m \cdot VCdim$ points and this would seem to imply, via the JLL, that $k \in \mathcal{O}(\epsilon^{-2} \log m \cdot VCdim)$ should be enough to control the generalization error of RP-SVM. Whether it is the case that for any classifier, or even for any linear classifier, that random projection to a dimensionality k of the order of the number of classes is sufficient to preserve good generalization performance remains an open problem. A likely first step in examining this question would be to use the observation made in chapter 6, that our flip probability implies an upper bound on the $(0, 1)$-generalization error of any consistent linear classifier trained by empirical risk

minimization (ERM) in a randomly projected space.
Furthermore, although intuitively one feels that carefully constructed deterministic projections should outperform random projections in preserving important distances, obtaining similar guarantees for classification of data following any non-adaptive deterministic projection technique (i.e. one that does not take account of the class labels, such as PCA) without restrictive conditions on the data still looks like a difficult open problem.

9.2 Flip Probability

In chapter 6 we derived the exact probability of 'label flipping' as a result of random projection, for the case where Σ is estimated by a spherical covariance matrix $\hat{\Sigma} = \alpha I$, and proved a simple, yet tight, upper bound on this probability for the general setting when $\hat{\Sigma}$ is allowed to be non-spherical. We conjectured on the basis of this upper bound that for non-spherical $\hat{\Sigma}$ there could be a trade off between reducing k which reduces $\kappa(R\hat{\Sigma}R^T)$ and increasing k which makes the flip probability smaller in the spherical case and, therefore, presumably also does the same in the non-spherical case. While testing this intuition by better quantifying the flip probability for non-spherical $\hat{\Sigma}$ remains for future research, we observe that evaluating the exact flip probability for non-spherical $\hat{\Sigma}$ looks like a particularly difficult open problem - the rotational invariance of the standard Gaussian distribution was key to our approach and we do not know of any tools that would allow us to dispense with it entirely.
Following on from the observation that, via existing VC-type bounds, our flip probability implies an upper bound on the $(0,1)$-generalization error of any (consistent) linear classifier trained by empirical risk minimization (ERM) in a randomly projected space, it would now be straightforward to sharpen the bounds of Garg & Roth (2003); Garg et al. (2002) using the findings in this chapter.

9.3 RP-FLD Ensembles

In chapter 7 we considered an ensemble of RP-FLD classifiers.
Our main contribution here was to show a direct link between the structure of an ensemble of 'weak' randomized classifiers and the 'strong' data space equivalent - in particular the regularization effect of this ensemble on the data space FLD.
The theory we developed here guarantees that, for most choices of projection dimension k, the error of a large ensemble remains bounded even when the number of training examples is far lower than the number of data dimensions, while for pseudo-inverted FLD in the data space it is known that the error is unbounded. As far as we are aware these are the first results for *any* ensemble classifier which explicitly demonstrate that the ensemble can, on average, outperform a data space equivalent.
Furthermore, we gained a good understanding of the effect of our discrete regularization parameter k, and found that a rule of thumb choice $k = \rho/2$ appears to work well in practice. In particular, we argued that the regularization parameter k allows us to finesse the known issue of poor eigenvector estimates in this setting.
We also demonstrated experimentally that this ensemble can recover or exceed the performance of a carefully-fitted ridge-regularized FLD in the data space, but with

generally lower computational cost, and an empirical comparison with linear SVM on the same data suggests that we can obtain good generalization performance with our approach even with few training examples.

We further showed that, for classification, the optimal choice of k depends on the true data parameters (which are unknown) thereby partly answering the question in Marzetta et al. (2011) concerning whether a simple formula for the optimal k exists.

This chapter is new work in a so far sparsely explored area - we could find very few papers using randomly-projected ensembles in the literature - and it would be interesting to extend this work to obtain similar guarantees for ensembles of generic randomly-projected linear classifiers in convex combination. The fact that this ensemble has very similar performance to a majority-voting ensemble is also striking, and we would like to understand better why this is the case.

It would also be very interesting to improve our results on deviations of the extreme eigenvalues of a finite ensemble from the extreme eigenvalues of the converged ensemble. As noted already, this looks particularly challenging: The main obstacle to achieving this is the rank deficiency of $\hat{\Sigma}$, since this implies that, w.p. > 0 there exist matrices R_i for which $\lambda_{\min}(R_i\hat{\Sigma}R_i^T)$ is arbitrarily close to zero (and therefore $\lambda_{\max}(R_i^T(R_i\hat{\Sigma}R_i^T)^{-1}R_i)$ can be arbitrarily large).

9.4 KFLD

In chapter 8 our main contribution is a dimension-free bound on the generalization error of KFLD which, to the best of our knowledge, is the first non-trivial bound for the standard KFLD model, putting KFLD on a solid theoretical foundation and improving our understanding of the working of this classifier. Although the tools used in our analysis are reasonably sharp, our approach does not manage to capture the relationship between kernel parameters and the generalization error, and it could be interesting to attempt a more detailed analysis which addresses this issue. Furthermore, we were unable so far to prove the consistency of KFLD, although it seems plausible that, by letting the regularization parameter go to zero as the number of training examples increases to infinity this could be achieved. For high probability guarantees we assumed that the kernel-induced space is a Gaussian Hilbert space and it appears, modulo careful checking, that extending these results to include Gaussian classes with different class-conditional covariances is possible using the results and tools already developed in chapters 4, 5 and 7. The main difference seems to be that one would need to use recent results for random matrices with subgaussian columns (i.e. the entries need not be i.i.d , but the columns must be i.i.d subgaussian vectors) such as those in (Vershynin, 2012), because the individual covariance matrices Σ_y would not whiten the pooled covariance estimate $\hat{\Sigma}_N$ in the feature space. Likewise similar guarantees for the much larger family of subgaussian class-conditional distributions should be achievable using tools such as those in (Rudelson & Vershynin, 2010; Vershynin, 2012).

Given the similarities between the setting of the previous chapter 7, where we have to deal with a singular covariance matrix embedded in Euclidean space $\mathbb{R}^d$, and the setting considered in this chapter 8 where we have to deal with a singular covariance matrix embedded in $\mathcal{H}_N \subset \ell_2$, it would appear that an ensemble of randomly-projected KFLD classifiers could be used to improve the performance of KFLD in much the same way. It would be interesting to find out if this is indeed the case in practice.

9.5 Some related research directions

Despite their longevity, other randomized approaches for generating ensemble members such as bagging (Breiman, 1996) and random subspaces (Ho, 1998) are still not well understood (Fumera et al., 2008). The scheme we use to link our RP-FLD ensemble to its dataspace FLD equivalent, via the regularization implemented by the random projections, could potentially be extended to consider these other randomized approaches, with the effect of each approach on the FLD ensemble captured by quantifying (using different tools) its effect on the parameters learned by the classifier. This scheme would avoid the weaknesses of bias-variance type decompositions, which leave the question 'when is an ensemble better than a strong classifier?' unanswered, and would allow us to draw comparisons between the ensemble and its dataspace counterpart. This could therefore be an interesting research direction.

Along rather different lines, it seems to me that random projections could be utilized in the area of stochastic local search (SLS) with the potential to provide approximate solutions to very large-scale optimization problems at small computational cost. In particular the approximate geometry preservation properties of random projections, via the JLL, and related results on randomly projected regression would appear to be relevant to the SLS setting. We are currently exploring this possibility as a way to scale up EDA – a state-of-the-art SLS approach though currently only practical in low-dimensional settings.

A
Appendix

A.1 Joint distribution of data

Let $(x, y) \sim \mathcal{D}_{x,y}$ where each instantiation of x is drawn i.i.d from $\mathcal{N}(\mu_y, \Sigma_y)$ with probability π_y. Then for any query point x_q we have $x_q \sim \sum_{y=0}^{1} \pi_y \mathcal{N}(\mu_y, \Sigma_y)$.
Proof:

$$\begin{aligned}
\Pr(x = x_q, y = y_q) &= \Pr(x_q, y_q) \\
&= \Pr(x_q|y_q) \Pr(y_q) \\
&= \mathcal{N}(x_q|\mu_{y_q}, \Sigma_{y_q}) \cdot \pi_0^{1-y_q}(1-\pi_0)^{y_q}
\end{aligned}$$

Then:

$$\begin{aligned}
x_q \sim &= \sum_{y_q=0}^{1} \Pr(x_q, y_q) \\
&= \sum_{y_q=0}^{1} \mathcal{N}(x_q|\mu_{y_q}, \Sigma_{y_q}) \cdot \pi_0^{1-y_q}(1-\pi_0)^{y_q} \\
&= \sum_{y_q=0}^{1} \mathcal{N}(x_q|\mu_{y_q}, \Sigma_{y_q}) \cdot \pi_{y_q} \text{ and so:} \\
x_q &\sim \sum_{y=0}^{1} \pi_y \cdot \mathcal{N}(x_q|\mu_y, \Sigma_y)
\end{aligned}$$

A.2 Projected means and covariances

Let R be any fixed instance of a random projection matrix, then:

(i) The sample mean of the projected data is the projection of the sample mean of the original data set.
Proof: Let $\hat{\mu}_R = \frac{1}{N}\sum_{i=1}^{N} R(x_i)$ (where $x_i \in \mathbb{R}^d$, the data space) be the sample mean of the projected data and $R(\hat{\mu})$ be the projection of the sample mean $\hat{\mu}$ of the data space. Then, by linearity, $\hat{\mu}_R = \frac{1}{N}\sum_{i=1}^{N} R(x_i) = R\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = R\hat{\mu}$ as claimed.

(ii) The mean of the projected data is the projection of the mean of the original data.
Proof: Let $N \to \infty$ in (i).

(iii) If $\Sigma = \mathrm{E}_x\left[(x-\mu)(x-\mu)^T\right]$ is the covariance matrix in the data space, then its projected counterpart Σ_R is $R\Sigma R^T$, and likewise the ML covariance estimate $\hat{\Sigma}_R = R\hat{\Sigma}R^T$:

$$\begin{aligned}\Sigma_R &= \mathrm{E}_x\left[R(x-\mu)(R(x-\mu))^T\right]\\ &= \mathrm{E}_x\left[R(x-\mu)(x-\mu)^T R^T\right]\\ &= R\mathrm{E}_x\left[(x-\mu)(x-\mu)^T\right]R^T\\ &= R\Sigma R^T\end{aligned}$$

Similarly:

$$\begin{aligned}\hat{\Sigma}_R &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left[R(x_i-\mu)(x_i-\mu)^T R^T\right]\\ &= R\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left[(x_i-\mu)(x_i-\mu)^T\right]R^T\\ &= R\hat{\Sigma}R^T\end{aligned}$$

A.3 Moments of quadratic forms involving random projection matrices

In chapter 5 we asserted that the m.g.f. of a quadratic form (i.e. a random variable) involving orthonormalized random projection matrices is bounded above by the corresponding m.g.f involving a normalized random projection matrix, that is by the m.g.f of a χ_k^2 distribution, but we gave no proof. Here we fill that gap by proving such an inequality holds.
Our approach is to prove the following theorem and from this its corollary, which is the result we ultimately want:

Theorem 15 *Let R be a random matrix, $R \in \mathcal{M}_{k\times d}$, $k < d$, with entries $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$. Let v be any (fixed) vector in $\mathbb{R}^d$. Then the moments of $v^T R^T\left(RR^T\right)^{-1}Rv$ are no greater than the moments of $vR^T Rv/\sigma^2 d$. Specifically:*

$$E_R\left[\left(v^T R^T\left(RR^T\right)^{-1}Rv\right)^i\right] \leqslant E_R\left[\left(\frac{1}{\sigma^2 d}\cdot v^T R^T Rv\right)^i\right],\ \forall i \in \mathbb{N} \text{ and } v \in \mathbb{R}^d$$

with equality only when $i = 1$.

Corollary 5
Let R, v be as given in theorem 15. Then:

$$E_R \left[\exp \left(v^T R^T \left(RR^T \right)^{-1} Rv \right) \right] \leqslant E_R \left[\exp \left(\frac{1}{\sigma^2 d} \cdot v^T R^T Rv \right) \right], \ \forall v \in \mathbb{R}^d$$

Note that taking v fixed then Rv is a $k \times 1$ vector with zero-mean Gaussian entries and therefore RHS above is the moment generating function of a χ^2_k distribution.

Proof of Theorem 15

We want to show that:

$$\mathrm{E}_R \left[\left(v^T R^T \left(RR^T \right)^{-1} Rv \right)^i \right] \leqslant \mathrm{E}_R \left[\left(\frac{1}{\sigma^2 d} \cdot v^T R^T Rv \right)^i \right], \ \forall i \in \mathbb{N} \text{ and } v \in \mathbb{R}^d \quad \text{(A.3.1)}$$

where $R \in \mathcal{M}_{k \times d}$ is a $k \times d$ random matrix with entries $r_{ij} \sim \mathcal{N}(0, \sigma^2)$ and so $R/\sigma\sqrt{d}$ is a $k \times d$ random matrix with entries $r_{ij} \sim \mathcal{N}(0, 1/d)$ and normalized rows. The normalization term $(1/\sigma^2 d)$ on RHS of the inequality (A.3.1) is required to make the LHS and RHS comparable, since the matrix $\left(RR^T \right)^{-1/2} R$ already has orthonormal rows[1].

Note that by the method of construction of R its rows are almost surely linearly independent, and hence $\text{rank}\,(R) = k$ with probability 1. In the following we may therefore safely assume that $\text{rank}\,(R) = \text{rank}\left(R^T R \right) = \text{rank}\left(RR^T \right) = k$.

The proof now proceeds via eigendecomposition of the symmetric positive semi-definite matrix $R^T R$.

When $v = 0$ there is nothing to prove, so let v be a non-zero vector in $\mathbb{R}^d$ and let x_j be a unit eigenvector of $R^T R$ with $\lambda(x_j)$ its corresponding eigenvalue. Since $R^T R \in \mathcal{M}_{d \times d}$ is symmetric there exists an orthonormal eigenvector basis $\{x_1, x_2, \ldots, x_d\}$ for $\mathbb{R}^d$, $\mathcal{B} = \{x_1, \ldots, x_d\}$ of eigenvectors of $R^T R$. Since $R^T R$ has rank $k < d$ this basis is, of course, not unique – however it will be enough to choose any suitable orthonormal eigenvector basis and then hold it fixed. Furthermore, since $\text{rank}\left(R^T R \right) = k < d$ we know that k of the eigenvalues $\lambda(x_j)$ are strictly positive and the remaining $d - k$ of

[1] Since $\left(RR^T \right)^{-1/2} RR^T \left(\left(RR^T \right)^{-1/2} \right)^T = \left(RR^T \right)^{-1/2} \left(RR^T \right)^{1/2} \left(RR^T \right)^{1/2} \left(RR^T \right)^{-1/2} = I$

the $\lambda(x_j)$ are zero. Writing $v = \sum_{j=1}^{d} \alpha_j x_j$ we then have:

$$\begin{aligned}
\frac{1}{\sigma^2 d} \cdot v^T R^T R v \quad &= \frac{1}{\sigma^2 d} \cdot \sum_{j=1}^{d} \lambda(x_j)\alpha_j x_j^T x_j \alpha_j = \frac{1}{\sigma^2 d} \cdot \sum_{j=1}^{d} \lambda(x_j)\alpha_j^2 \|x_j\|^2 \\
&= \frac{1}{\sigma^2 d} \left(\sum_{\{j:\lambda(x_j)\neq 0\}} \lambda(x_j)\alpha_j^2 \|x_j\|^2 + \sum_{\{j:\lambda(x_j)=0\}} 0 \cdot \alpha_j^2 \|x_j\|^2 \right) \\
&= \frac{1}{\sigma^2 d} \cdot \sum_{\{j:\lambda(x_j)\neq 0\}} \lambda(x_j)\alpha_j^2 \|x_j\|^2 \\
&= \frac{1}{\sigma^2 d} \cdot \sum_{\{j:\lambda(x_j)\neq 0\}} \lambda(x_j)\alpha_j^2 \qquad \text{(A.3.2)}
\end{aligned}$$

Next, note that if x_j is an eigenvector of $R^T R$ with non-zero eigenvalue $\lambda(x_j)$, then Rx_j is an eigenvector of RR^T with the same non-zero eigenvalue, since:

$$R^T R x_j = \lambda(x_j) x_j \Longrightarrow R R^T R x_j = \lambda(x_j) R x_j$$

There are k such non-zero eigenvalues, and as $\operatorname{rank}(RR^T) = k$ the non-zero eigenvalues of $R^T R$ are the eigenvalues of RR^T. Furthermore, since $RR^T \in \mathcal{M}_{k\times k}$ and has rank k, RR^T is invertible. It now follows that if x_j is an eigenvector of $R^T R$ with non-zero eigenvalue $\lambda(x_j)$, then Rx_j is an eigenvector of $\left(RR^T\right)^{-1}$ with non-zero eigenvalue $1/\lambda(x_j)$. Hence:

$$\begin{aligned}
v^T R^T \left(RR^T\right)^{-1} R v \quad &= \sum_{j=1}^{k} \frac{1}{\lambda(x_j)} \cdot \lambda(x_j)\alpha_j^2 \|x_j\|^2 \\
&= \sum_{\{j:\lambda(x_j)\neq 0\}} \alpha_j^2 \qquad \text{(A.3.3)}
\end{aligned}$$

We can now rewrite the inequality (A.3.1) to be proved as the following equivalent problem. For all $i \in \mathbb{N}$:

$$\begin{aligned}
\mathrm{E}_R\left[\left(\sum_{\{j:\lambda(x_j)\neq 0\}} \alpha_j^2\right)^i\right] \quad &\leqslant \mathrm{E}_R\left[\left(\frac{1}{\sigma^2 d}\sum_{\{j:\lambda(x_j)\neq 0\}} \lambda(x_j)\alpha_j^2\right)^i\right] \\
\Longleftrightarrow \ \mathrm{E}_{\lambda,\alpha}\left[\left(\sum_{\{j:\lambda(x_j)\neq 0\}} \alpha_j^2\right)^i\right] &\leqslant \mathrm{E}_{\lambda,\alpha}\left[\left(\frac{1}{\sigma^2 d}\sum_{\{j:\lambda(x_j)\neq 0\}} \lambda(x_j)\alpha_j^2\right)^i\right] \\
\Longleftrightarrow \ \mathrm{E}_{\lambda,\alpha}\left[\left(\sum_{\{j:\lambda(x_j)\neq 0\}} \alpha_j^2\right)^i\right] &\leqslant \mathrm{E}_{\alpha}\left[\mathrm{E}_{\lambda|\alpha}\left[\left(\sum_{\{j:\lambda(x_j)\neq 0\}} \frac{1}{\sigma^2 d}\lambda(x_j)\alpha_j^2\right)^i\right]\right]
\end{aligned}$$
$$\text{(A.3.4)}$$

Now, in RHS of (A.3.4) x_j and $\lambda(x_j)$ are independent of one another (e.g. Tulino & Verdú (2004) Lemma 2.6, Artin (2010) Proposition 4.18) and v is arbitrary but fixed. Since the α_j depend only on v and the x_j, we see that λ and α are independent of one another, so we can rewrite this term as:

$$\mathrm{E}_\alpha\left[\mathrm{E}_\lambda\left[\left(\sum_{\{j:\lambda(x_j)\neq 0\}}\frac{1}{\sigma^2 d}\lambda(x_j)\alpha_j^2\right)^i\right]\right] \tag{A.3.5}$$

and since (A.3.5) is the expectation of a convex function, applying Jensen's inequality to the inner term we see that:

$$\mathrm{E}_\alpha\left[\left(\mathrm{E}_\lambda\left[\sum_{\{j:\lambda(x_j)\neq 0\}}\frac{1}{\sigma^2 d}\lambda(x_j)\alpha_j^2\right]\right)^i\right] \leqslant \mathrm{E}_\alpha\left[\mathrm{E}_\lambda\left[\left(\sum_{\{j:\lambda(x_j)\neq 0\}}\frac{1}{\sigma^2 d}\lambda(x_j)\alpha_j^2\right)^i\right]\right] \tag{A.3.6}$$

Note that when $i = 1$ we have equality in (A.3.6) and strict inequality when $i > 1$. If we can show that the LHS of (A.3.6) above is no less than the LHS of (A.3.1) then we are done. Now, equation (A.3.3) implies that all terms in LHS of (A.3.1) are positive, and so in order to prove the theorem it is enough to show that $\mathrm{E}_R\left[\lambda(x_j)\right]/\sigma^2 d \geqslant 1$. But this is certainly so since:

$$\mathrm{E}_R\left[\lambda(x_j)\right] = \frac{1}{k}\sum_{\{j:\lambda(x_j)\neq 0\}}\mathrm{E}_R\left[\lambda(x_j)\right] = \frac{1}{k}\mathrm{E}_R\left[\mathrm{Tr}\left(RR^T\right)\right] = \frac{1}{k}\sum_{j=1}^{k}\mathrm{E}_R\left[r_j^T r_j\right] \tag{A.3.7}$$

where r_j is the j-th row of R. Then, since the $r_j \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \mathrm{diag}(\sigma^2)\right)$ we have $r_j^T r_j/\sigma^2 \overset{\text{i.i.d}}{\sim} \chi_d^2$ and so $\mathrm{E}_R\left[r_j^T r_j/\sigma^2\right] = d$. Finally, it then follows that $\frac{1}{k}\sum_{j=1}^{k}\mathrm{E}_R\left[\lambda(x_j)\right]/\sigma^2 d = 1$ and this completes the proof. □

Proof of Corollary 5

To prove the corollary we rewrite the inequality (A.3.1) using the Taylor series expansion for exp to see that:

$$\begin{aligned}
&\mathrm{E}_R\left[\sum_{i=0}^{\infty}\frac{\left(v^T R^T\left(RR^T\right)^{-1}Rv\right)^i}{i!}\right] &&\leqslant \mathrm{E}_R\left[\sum_{i=0}^{\infty}\frac{\left(\frac{1}{\sigma^2 d}\cdot v^T R^T R v\right)^i}{i!}\right] \\
\Longrightarrow\quad &\mathrm{E}_R\left[\exp\left(v^T R^T\left(RR^T\right)^{-1}Rv\right)\right] &&\leqslant \mathrm{E}_R\left[\exp\left(\frac{1}{\sigma^2 d}\cdot v^T R^T R v\right)\right]
\end{aligned} \tag{A.3.8}$$

Since by theorem 15 we have the required inequality for each of the i-th powers in the summations in equation (A.3.8), the result follows immediately. □

A.4 Proof of Theorem 11

There are five terms to simultaneously bound with high probability, namely the two B_y, A, and the two extreme eigenvalues involved in the condition number bound. We use the standard approach of setting each of the confidence probabilities no greater than $\delta/5$ and solving for ϵ (or a function of ϵ appearing in the bound) then back-substituting and applying the union bound to derive a guarantee which holds with probability $1-\delta$. Firstly, for the extreme eigenvalues we have (twice):

$$\begin{aligned} \exp\left(-\epsilon_3^2/2\right) &\leqslant \delta/5 \\ \implies \sqrt{2\log(5/\delta)} &\leqslant \epsilon_3 \end{aligned} \tag{A.4.1}$$

For the upper bounds on the B_y we have:

$$\exp\left(-\frac{d}{2}\left(\sqrt{1+\epsilon_y}-1\right)^2\right) \leqslant \delta/5$$

and solving for $\sqrt{1+\epsilon_y}$ we obtain:

$$\begin{aligned} \sqrt{\frac{2\log(5/\delta)}{d}} &\leqslant \pm\left(\sqrt{1+\epsilon_y}-1\right) \\ \implies 1+\sqrt{\frac{2\log(5/\delta)}{d}} &\geqslant \sqrt{1+\epsilon_y} \end{aligned} \tag{A.4.2}$$

Finally, for the lower bound on A (which holds for both classes simultaneously) we solve for $\sqrt{1-\epsilon_2}$ to obtain:

$$\exp\left(-\left(\frac{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}{2N/N_0N_1}\right)\left(\sqrt{1-\epsilon_2}-1\right)^2\right) \leqslant \delta/5$$

$$\begin{aligned} \iff \frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2} &\leqslant \left(\sqrt{1-\epsilon_2}-1\right)^2 \\ \iff \sqrt{\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}} &\leqslant \pm\left(\sqrt{1-\epsilon_2}-1\right) \\ \implies 1-\sqrt{\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}} &\geqslant \sqrt{1-\epsilon_2} \end{aligned} \tag{A.4.3}$$

Plugging the left hand sides of the inequalities (A.4.1), (A.4.2) and (A.4.3) into the bounds on κ, B_0, B_1 and A for ϵ_3, $\sqrt{1+\epsilon_0}$, $\sqrt{1+\epsilon_1}$ and $\sqrt{1-\epsilon_2}$ respectively gives, after some algebra, the stated Theorem 11.

A.5 Proof of Lemma 23

We prove the statement of eq. (8.2.2) fully, and outline the proof of (8.2.3) which is very similar. Let $t > 0$ be a positive real constant (to be optimized later), then:

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\}\\
&= \Pr\left\{\exp\left(t\|X\|^2\right) \geqslant \exp\left(t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\right\}\\
&\leqslant \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\mathrm{E}[\exp\left(t\|X\|^2\right)]
\end{aligned} \tag{A.5.1}$$

Where (A.5.1) follows by Markov's inequality. Now, $X \sim \mathcal{N}(\mu, \Sigma)$ and so $\|X\|^2 = \sum_{i=1}^{\infty} X_i^2$ has a non-central χ^2 distribution, and therefore $\mathrm{E}\left[\exp\left(t\|X\|^2\right)\right]$ is the moment generating function of a non-central χ^2 distribution. Hence (e.g. (Maniglia & Rhandi, 2004) proposition 1.2.8) for all $t \in (0, 1/2\lambda_{\max}(\Sigma))$ we have (A.5.1) is equal to:

$$\begin{aligned}
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\prod_{i=1}^{\infty}\left(1-2t\lambda_i\right)^{-\frac{1}{2}}\exp\left(\tfrac{t\mu_i^2}{1-2t\lambda_i}\right)\\
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\prod_{i=1}^{\infty}\left(1+\tfrac{2t\lambda_i}{1-2t\lambda_i}\right)^{\frac{1}{2}}\exp\left(\tfrac{t\mu_i^2}{1-2t\lambda_i}\right)\\
&\leqslant \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\prod_{i=1}^{\infty}\exp\left(\tfrac{1}{2}\tfrac{2t\lambda_i}{1-2t\lambda_{\max}(\Sigma)}\right)\exp\left(\tfrac{t\mu_i^2}{1-2t\lambda_{\max}(\Sigma)}\right)\\
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)+\tfrac{t\left(\sum_{i=1}^{\infty}\lambda_i+\mu_i^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right)\\
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)+\tfrac{t\left(\mathrm{Tr}(\Sigma)+\|\mu\|^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right)
\end{aligned} \tag{A.5.2}$$

Now taking $t = \frac{1-(1+\epsilon)^{-\frac{1}{2}}}{2\lambda_{\max}(\Sigma)} \in (0, 1/2\lambda_{\max}(\Sigma))$ and substituting this value of t into (A.5.2) yields, after some algebra, (8.2.2):

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\}\\
&\leqslant \exp\left(-\tfrac{\mathrm{Tr}(\Sigma)+\|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon}-1\right)^2\right)
\end{aligned}$$

The second inequality (8.2.3) is proved similarly. We begin by noting:

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\}\\
= \;&\Pr\left\{\exp\left(-t\|X\|^2\right) \geqslant \exp\left(-t\left(1-\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\right\}\\
\leqslant \;&\exp\left(t(1-\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)-t\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)/1+2t\lambda_{\max}(\Sigma)\right)
\end{aligned}$$

and then complete the proof as before, substituting in the optimal $t = \frac{1+(1-\epsilon)^{-\frac{1}{2}}}{2\lambda_{\max}(\Sigma)}$ to give the bound.

Comment: Bound on sub-exponential random variables.

Our probability bound uses the moment generating function of a non-central chi-square and therefore also holds for probability distributions whose m.g.f is dominated by that of the chi-square or, with the appropriate changes, by a constant multiple of it. Such distributions are called sub-exponential distributions (Vershynin, 2012).

A.6 Proof of Lemma 24

Without loss of generality let x_q have label 0, and note that for KFLD the decision of which label to assign to a query point x_q is made with respect to the projection of x_q onto $\mathcal{H}_N$. The probability that x_q is misclassified is therefore given by:

$$\Pr_{x_q|y_q=0}\left\{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N>0\right\} \tag{A.6.1}$$

Define $a_N^T := (\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}$ and observe that if $x_q \sim \mathcal{N}(\mu_0,\Sigma)$ then:

$$\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N \sim \mathcal{N}\left(\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N,\Sigma_N\right)$$

and so:

$$a_N^T\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N \sim \mathcal{N}\left(a_N^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N,a_N^T\Sigma_N a_N\right)$$

which is a univariate Gaussian. Therefore:

$$\frac{a_N^T\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N-a_N^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N}{\sqrt{a_N^T\Sigma_N a_N}} \sim \mathcal{N}(0,1)$$

Hence, for the query point x_q we have the probability (A.6.1) is given by:

$$\begin{aligned}\Pr_{x_q}&\left\{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\left(x_q-\tfrac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N>0\middle|\, y=0\right\}\\ &=\Phi\left(\frac{a_N^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N}{\sqrt{a_N^T\Sigma_N a_N}}\right)\\ &=\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_0)_N}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}(\hat{\mu}_1-\hat{\mu}_0)_N}}\right)\end{aligned}$$

where Φ is the c.d.f of the standard Gaussian.
A similar argument deals with the case when x_q belongs to class 1, and applying the law of total probability gives the lemma.

References

Abramowitz, M. and Stegun, I.A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover, New York, 10th edition, 1972. 67

Achlioptas, D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. xx, 8, 14, 17, 49, 76, 100, 103

Aha, D.W., Kibler, D., and Albert, M.K. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991. 23

Ailon, N. and Chazelle, B. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 557–563. ACM, 2006. 22

Alon, N. Problems and results in extremal combinatorics, Part I. *Discrete Math*, 273: 31–53, 2003. 51

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 20–29. ACM, 1996. 22

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745, 1999. 99

Anthony, M. and Bartlett, P.L. *Neural Network Learning: Theoretical Foundations.* Cambridge University press, 1999. 18

Arriaga, R. and Vempala, S. An algorithmic theory of learning. *Machine Learning*, 63 (2):161–182, 2006. 22, 24, 30

Arriaga, R.I. and Vempala, S. An algorithmic theory of learning: Robust concepts and random projection. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 616–623. IEEE, 1999. 3, 14, 50, 51, 62

Artin, M. *Algebra.* Pearson Education, 2010. ISBN 9780132413770. URL http://books.google.co.uk/books?id=QsOfPwAACAAJ. 8, 129

Avogadri, R. and Valentini, G. Fuzzy ensemble clustering based on random projections for dna microarray data analysis. *Artificial Intelligence in Medicine*, 45(2):173–183, 2009. 22

Balcan, M.-F., Blum, A., and Vempala, S. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65:79–94, 2006. 25

Ball, K. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997. 58, 70

Baraniuk, R.G. and Wakin, M.B. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009. 25

Bellman, R.E. *Methods of nonlinear analysis.* Academic Press, 1970. 1

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. When is nearest neighbor meaningful? In *Proc. Int. Conf. Database Theory*, pp. 217–235, 1999. 2

Bickel, P. and Levina, E. Some theory for Fisher's linear discriminant function, 'naïve Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004. 29, 31, 36, 81, 83, 92, 94, 106

Bingham, E. and Mannila, H. Random projection in dimensionality reduction: applications to image and text data. In F. Provost and R. Srikant (ed.), *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pp. 245–250, 2001. 3, 14, 22, 23, 24

Blum, A. Random Projection, Margins, Kernels, and Feature-Selection. In et al., Saunders (ed.), *SLSFS 2005*, number 3940 in LNCS, pp. 55–68, 2006. 22

Boutsidis, C. and Drineas, P. Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications*, 431(5-7):760–771, 2009. 22

Boyali, A. and Kavakli, M. A robust gesture recognition algorithm based on sparse representation, random projections and compressed sensing. In *7th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2012*, pp. 243–249, july 2012. 22, 24

Breiman, L. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 124

Brown, G. Ensemble Learning. In Sammut, C. and Webb, G.I. (eds.), *Encyclopedia of Machine Learning.* Springer, 2009. 76

Buhler, J. and Tompa, M. Finding motifs using random projections. *Journal of computational biology*, 9(2):225–242, 2002. 22

Calderbank, R., Jafarpour, S., and Schapire, R. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Rice University, 2009. 3, 22, 25, 30

Candes, E. and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007. 75

Candes, E.J. and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12): 5406–5425, 2006. ISSN 0018-9448. 22, 24

Candes, E.J., Wakin, M.B., and Boyd, S.P. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008. 3

Centeno, T.P. and Lawrence, N.D. Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. *The Journal of Machine Learning Research*, 7:455–491, 2006. 113, 115

Chen, Y., Nasrabadi, N.M., and Tran, T.D. Random projection as dimensionality reduction and its effect on classical target recognition and anomaly detection techniques. In *3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2011)*, pp. 1–4. IEEE, 2011. 22, 24

Cukier, K. Data, data, everywhere, February 2010. URL `http://www.economist.com/node/15557443`. The Economist - Special Report on Managing Information. 41

Dasgupta, S. Learning Mixtures of Gaussians. In *Annual Symposium on Foundations of Computer Science*, volume 40, pp. 634–644, 1999. 22, 24, 25, 43, 51, 57, 71

Dasgupta, S. Experiments with random projection. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pp. 143–151, 2000a. 14, 23, 24, 51, 57, 71

Dasgupta, S. *Learning Probability Distributions*. PhD thesis, Berkeley, 2000b. 24, 110

Dasgupta, S. and Freund, Y. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pp. 537–546. ACM, 2008. 22

Dasgupta, S. and Gupta, A. An Elementary Proof of the Johnson-Lindenstrauss Lemma. *Random Struct. Alg.*, 22:60–65, 2002. 14, 17, 51, 93

Dasgupta, S., Papadimitriou, C.H., and Vazirani, U. *Algorithms*. McGraw Hill, 2008. 3

Dasgupta, S., Hsu, D., and Verma, N. A concentration theorem for projections. *arXiv preprint arXiv:1206.6813*, 2012. 105, 106

Davenport, M.A., Boufounos, P.T., Wakin, M.B., and Baraniuk, R.G. Signal Processing with Compressive Measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, April 2010. 43

De Bruijn, N.G. Inequalities concerning minors and eigenvalues. *Nieuw Archief voor Wiskunde*, 3:18–35, 1956. 14

Diaconis, P. and Freedman, D. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984. ISSN 0090-5364. 105, 106

Diethe, T., Hussain, Z., Hardoon, D., and Shawe-Taylor, J. Matching pursuit kernel fisher discriminant analysis. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pp. 121–128, 2009. 25, 106, 110

Donoho, D.L. Compressed Sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, 2006. 3, 22, 24

Dubhashi, D.P. and Panconesi, A. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2012. 3

Duda, R.O., Hart, P.E., and Stork, D.G. *Pattern Classification*. Wiley, 2 edition, 2000. 113

Dudoit, S., Fridlyand, J., and Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002. 30, 99

Durrant, R.J. and Kabán, A. When is 'nearest neighbour' meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397, 2009. 2

Durrant, R.J. and Kabán, A. A bound on the performance of LDA in randomly projected data spaces. In *Proceedings 20th International Conference on Pattern Recognition (ICPR 2010)*, pp. 4044–4047, 2010a. 43

Durrant, R.J. and Kabán, A. Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data. In *Proceedings16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 2010b. 13, 43, 84

Durrant, R.J. and Kabán, A. A tight bound on the performance of Fisher's linear discriminant in randomly projected data spaces. *Pattern Recognition Letters*, 2011. doi: http://dx.doi.org/10.1016/j.patrec.2011.09.008. 15, 43

Durrant, R.J. and Kabán, A. Random Projections for Machine Learning and Data Mining: Theory & Applications, 2012a. URL https://sites.google.com/site/rpforml/. ECML-PKDD 2012 Tutorial, 28th September 2012. 25

Durrant, R.J. and Kabán, A. Random Projections as Regularizers: Learning a Linear Discriminant Ensemble from Fewer Observations than Dimensions. Technical Report CSR-12-01, University of Birmingham, School of Computer Science, 2012b. URL http://www.cs.bham.ac.uk/~durranrj/RP_FLD_Ensembles.pdf. 22

Eisenberg, B. and Sullivan, R. Random triangles in n dimensions. *American Mathematical Monthly*, pp. 308–318, 1996. 62

Eltoft, T. and deFigueiredo, R.J.P. Pattern classification of non-sparse data using optimal interpolative nets. *Neurocomputing*, 10(4):385 – 403, 1996. ISSN 0925-2312. doi: DOI:10.1016/0925-2312(95)00045-3. URL http://www.sciencedirect.com/science/article/B6V10-3VS5JP3-X/2/dce3f877fa892f2d998c235699541d3f. Financial Applications. 3

Fard, M., Grinberg, Y., Pineau, J., and Precup, D. Compressed least-squares regression on sparse spaces. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. 24

Fern, X.Z. and Brodley, C.E. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings 20th International Conference on Machine Learning (ICML 2003)*, volume 20(1), pp. 186, 2003. 22, 23

Fodor, I.K. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494, US Dept. of Energy, Lawrence Livermore National Laboratory, 2002. 2

Fradkin, D. and Madigan, D. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 522–529. ACM, 2003a. 3, 14, 22, 23, 98, 99

Fradkin, D. and Madigan, D. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 522. ACM, 2003b. 51

Friedman, J.H. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989. 31, 32, 81

Fumera, G., Roli, F., and Serrau, A. A theoretical analysis of bagging as a linear combination of classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1293–1299, 2008. 124

Garg, A. and Roth, D. Margin distribution and learning algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pp. 210–217, 2003. 72, 122

Garg, A., Har-Peled, S., and Roth, D. On generalization bounds, projection profile, and margin distribution. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, pp. 171–178, 2002. 25, 72, 122

Goel, N., Bebis, G., and Nefian, A. Face recognition experiments with random projection. In *Proceedings of SPIE*, volume 5779, pp. 426, 2005. 22, 24

Goemans, M.X. and Williamson, D.P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1145, 1995. 63, 68, 69

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531, 1999. 99

Guo, Y., Hastie, T., and Tibshirani, R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007. 30

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning; data mining, inference, and prediction.* Springer, 2001. 2, 31, 81

Haupt, J., Castro, R., Nowak, R., Fudge, G., and Yeh, A. Compressive sampling for signal classification. In *Proc. 40th Asilomar Conf. on Signals, Systems, and Computers*, pp. 1430–1434, 2006. 43

Hegde, C., Wakin, M.B., and Baraniuk, R.G. Random projections for manifold learningproofs and analysis. In *Neural Information Processing Systems*, 2007. 22

Herbrich, R. *Learning kernel classifiers: theory and algorithms.* The MIT Press, 2002. 109

Ho, T.K. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998. 124

Horn, R.A. and Johnson, C.R. *Matrix Analysis.* Cambridge University Press, 1985. 8, 11, 12, 13, 14, 66, 80, 86

Hoyle, D. Accuracy of Pseudo-Inverse Covariance Learning – A Random Matrix Theory Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(7):1470 – 81, 2011. 83

Huang, S.Y., Hwang, C.R., and Lin, M.H. Kernel fishers discriminant analysis in gaussian reproducing kernel hilbert space. Technical report, Institute of Statistical Science, Academia Sinica, 2005. 105, 106, 108

Indyk, P. and Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613. ACM New York, NY, USA, 1998. 3, 22, 25

Indyk, P. and Naor, A. Nearest neighbor preserving embeddings. *ACM Trans. Algorithms*, 3(3):31, 2007. 22

Johnson, N.L., Kotz, S., and Balakrishnan, N. *Continuous Univariate Distributions*, volume 1. Wiley, 2 edition, 1994. 37

Johnstone, I.M. and Lu, A.Y. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104 (486):682–693, 2009. 86

Kalai, A.T., Moitra, A., and Valiant, G. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012. 22, 25

Kaski, S. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 1, pp. 413–418. IEEE, 1998. 23

Kendall, MG. *A Course in the Geometry of n Dimensions*. Dover Pubns, New York, 2004. 69

Khor, L.C., Woo, W.L., and Dlay, S.S. Non-sparse approach to underdetermined blind signal estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP05)*, volume 5, pp. 309–312, 2005. 3

Kim, T.K. and Kittler, J. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005. 30

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002. 84

Kurimo, M. and IDIAP, M. Fast latent semantic indexing of spoken documents by using self-organizing maps. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings*, volume 6, 2000. 23

Lachenbruch, P.A. and Goldstein, M. Discriminant analysis. *Biometrics*, pp. 69–85, 1979. 32

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. 110

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004. 31, 81, 86

Ledoux, M. *The concentration of measure phenomenon*, volume 89. American Mathematical Society, 2001. 2

Liu, Y. and Liu, Y. Reducing the run-time complexity of support vector data descriptions. In *IJCNN'09: Proceedings of the 2009 international joint conference on Neural Networks*, pp. 3075–3082, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-3549-4. 22

Lu, J., Plataniotis, K.N., and Venetsanopoulos, A.N. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26(2):181–191, 2005. 30

Ma, J., Kulesza, A., Dredze, M., Crammer, K., Saul, L.K., and Pereira, F. Exploiting feature covariance in high-dimensional online learning. In *Proc. 13th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2010)*, volume 9 of *JMLR W&CP*, pp. 493–500, 2010. 31

Maillard, O. and Munos, R. Compressed Least-Squares Regression. In *NIPS*, 2009. 22, 25

Maniglia, S. and Rhandi, A. Gaussian measures on separable Hilbert spaces and applications. *Quaderni del Dipartimento di Matematica dell'Università del Salento*, 1 (1):1–24, 2004. 131

Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate analysis.* Academic Press, London, 1979. 12, 66, 89, 90, 93

Marzetta, T.L., Tucci, G.H., and Simon, S.H. A Random Matrix–Theoretic Approach to Handling Singular Covariance Estimates. *IEEE Trans. Information Theory*, 57 (9):6256–71, September 2011. 77, 78, 91, 100, 123

Matoušek, J. On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008. 17

McDiarmid, C. On the method of bounded differences. *Surveys in combinatorics*, 141 (1):148–188, 1989. 18

McLachlan, G.J. *Discriminant analysis and statistical pattern recognition*, volume 544. Wiley-Interscience, 2004. 30, 31

Mika, S. *Kernel Fisher Discriminants.* PhD thesis, Technical University of Berlin, 2002. 25, 106

Mika, S., Ratsch, G., Weston, J., Schölkopf, B., and Mullers, KR. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 41–48. IEEE, 2002. ISBN 078035673X. 98, 106

Mu, Y., Dong, J., Yuan, X., and Yan, S. Accelerated low-rank visual recovery by random projection. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pp. 2609–2616. IEEE, 2011. 22

Panconesi, A. Randomized Algorithms: Looking for certain uncertainty, January 2012. URL `http://www.cirm.univ-mrs.fr/liste_rencontre/programmes/Renc704/Panconesi2.pdf`. CIRM Workshop on Concentration Inequalities and their Applications. 3

Pang, S., Ozawa, S., and Kasabov, N. Incremental linear discriminant analysis for classification of data streams. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 35(5):905–914, 2005. 30

Pattison, T. and Gossink, D. Misclassification Probability Bounds for Multivariate Gaussian Classes. *Digital Signal Processing*, 9:280–296, 1999. 29, 36, 106

Paul, B., Athithan, G., and Murty, M.N. Speeding up adaboost classifier with random projection. In *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*, pp. 251–254. IEEE, 2009. 23

Paul, D. and Johnstone, I.M. Augmented sparse principal component analysis for high dimensional data. *arXiv preprint arXiv:1202.1242*, 2012. 86

Paul, S., Boutsidis, C., Magdon-Ismail, M., and Drineas, P. Random projections for support vector machines. *arXiv preprint arXiv:1211.6085*, 2012. 25

Penrose, R. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955. 82

Pillai, J.K., Patel, V.M., Chellappa, R., and Ratha, N.K. Secure and robust iris recognition using random projections and sparse representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1877–1893, 2011. 22, 24

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2008a. 22

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems (NIPS)*, 2008b. 110

Ramsay, J.O. Functional data analysis. *Encyclopedia of Statistical Sciences*, 1997. 108

Raudys, S. and Duin, R.P.W. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19(5):385–392, 1998. 31, 76, 83

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011. 18, 22

Rosenthal, J.S. *A first look at rigorous probability theory.* World Scientific Pub Co Inc, 2006. ISBN 9812703705. 15

Rudelson, M. and Vershynin, R. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians, Hyderabad, India, 2010*, 2010. 114, 119, 123

Sarlos, T. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 143–152. IEEE, 2006. 22

Schclar, A. and Rokach, L. Random projection ensemble classifiers. In Filipe, Joaquim, Cordeiro, Jos, Aalst, Wil, Mylopoulos, John, Rosemann, Michael, Shaw, Michael J., and Szyperski, Clemens (eds.), *Enterprise Information Systems*, volume 24 of *Lecture Notes in Business Information Processing*, pp. 309–316. Springer, 2009. 22, 23

Schölkopf, B. and Smola, A.J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002. ISBN 0262194759. 108

Shawe-Taylor, J. and Cristianini, N. Estimating the moments of a random vector with applications. In *Proceedings of GRETSI 2003 conference*, volume 1, pp. 47–52, 2003. 110

Shawe-Taylor, J., Williams, C.K.I, Cristianini, N., and Kandola, J. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005. 95

Shi, Q., Shen, C., Hill, R., and van den Hengel, A. Is margin preserved after random projection? In *29th International Conference on Machine Learning (ICML 2012)*, 2012. 25

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., and Sellers, W.S. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002. 99

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In Hutter, Marcus, Servedio, Rocco, and Takimoto, Eiji (eds.), *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pp. 13–31. Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-75224-0. 106, 117

Srivastava, M.S. Singular wishart and multivariate beta distributions. *Annals of Statistics*, pp. 1537–1560, 2003. 114

Tulino, A.M. and Verdú, S. *Random matrix theory and wireless communications*. Now Publishers Inc, 2004. ISBN 193301900X. 129

Vershynin, R. Introduction to Non-asymptotic Random Matrix Theory. In Eldar, Y. and Kutyniok, G. (eds.), *Compressed Sensing, Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. 16, 35, 44, 49, 89, 90, 119, 123, 132

Vu, V. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4):526–538, 2011. ISSN 1098-2418. doi: 10.1002/rsa.20367. URL `http://dx.doi.org/10.1002/rsa.20367`. 86

Vu, V. and Lei, J. Minimax rates of estimation for sparse pca in high dimensions. In *Proceedings 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, volume 22, pp. 1278–1286. JMLR W&CP, 2012. 86

Watanabe, T., Takimoto, E., Amano, K., and Maruoka, A. Random projection and its application to learning. In *Randomness and Computation Joint Workshop 'New Horizons in Computing' and 'Statistical Mechanical Approach to Probabilistic Information Processing'*, July 2005. URL http://www.smapip.is.tohoku.ac.jp/~smapip/2005/NHC+SMAPIP/ExtendedAbstracts/KazuyukiAmano.pdf. 50

Weisstein, Eric W. Chi-squared distribution. From MathWorld—A Wolfram Web Resource. URL http://mathworld.wolfram.com/Chi-SquaredDistribution.html. Retrieved 12/07/2012. 48

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R., and Nevins, J.R. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462, 2001. 99

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., and Ma, Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 22, 24

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, DL, Heuerding, S., and Erni, F. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to nir data. *Analytica chimica acta*, 329(3): 257–265, 1996. 30, 32

Zhang, L., Jin, R., and Yang, T. Recovering optimal solution by dual random projection. *arXiv preprint arXiv:1211.3046*, 2012. 25

Zhou, S., Lafferty, J., and Wasserman, L. Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009. 22, 25