A tight bound on the performance of Fisher's linear discriminant in randomly projected data spaces

Robert John Durrant*, Ata Kabán*

School of Computer Science, University of Birmingham, Edgbaston, UK, B15 2TT

Abstract

We consider the problem of classification in non-adaptive dimensionality reduction. Specifically, we give an average-case bound on the classification error of Fisher's Linear Discriminant classifier when the classifier only has access to randomly projected versions of a given training set. By considering the system of random projection and classifier together as a whole, we are able to take advantage of the simple class structure inherent in the problem, and so derive a non-trivial performance bound without imposing any sparsity or underlying low-dimensional structure restrictions on the data. Our analysis also reveals and quantifies the effect of class 'flipping' – a potential issue when randomly projecting a finite sample. Our bound is reasonably tight, and unlike existing bounds on learning from randomly projected data, it becomes tighter as the quantity of training data increases. A preliminary version of this work received an IBM Best Student Paper Award at the 20th International Conference on Pattern Recognition.

1. Introduction

The application of pattern recognition and machine learning techniques to very high dimensional data sets presents unique challenges, often described by the term 'the curse of dimensionality'. These include issues concerning the collection and storage of such high dimensional data, as well as time- and space-complexity issues arising from working with the data. The analysis of learning from non-adaptive data projections has therefore received increasing interest in recent years [1, 2, 3, 4].

Here we consider the supervised learning problem of classifying a query point $\mathbf{x}_q \in \mathbb{R}^d$ as belonging to one of two Gaussian or sub-Gaussian classes using Fisher's Linear Discriminant (FLD) and the misclassification error arising if,

*Corresponding author

Email addresses: R.J.Durrant@cs.bham.ac.uk (Robert John Durrant), A.Kaban@cs.bham.ac.uk (Ata Kabán)

URL: http://www.cs.bham.ac.uk/∼durranrj (Robert John Durrant), http://www.cs.bham.ac.uk/∼axk (Ata Kabán)

instead of learning the classifier in the data space $\mathbb{R}^d$, we instead learn it in some low dimensional random projection of the data space $R(\mathbb{R}^d) \equiv \mathbb{R}^k$, where $R \in \mathcal{M}_{k\times d}$ is an orthonormalised random projection matrix with entries drawn i.i.d from a zero-mean finite variance Gaussian. Such bounds on the classification error for FLD in the data space are already known, for example those in [5, 6], but in neither of these papers is classification error in the projected domain considered; indeed in [2] it is stated that establishing the probability of error for a classifier in the projected domain is, in general, a difficult problem.

Unlike the bounds in [7], where the authors' use of the Johnson-Lindenstrauss Lemma on the set of training points has the unwanted side-effect that their bound loosens as the number of training examples increases, our bound tightens with more training data. Moreover, we do not require any sparsity structure from the data, as the Compressed Sensing based analysis in [1] does. Starting from first principles, and using standard techniques, we are able to exploit the class structure implied by the problem and bypass the need to preserve all pairwise distances from the data space. We derive a non-trivial bound on the average error of the classifier learned in a random projection of the data space that decays exponentially as the projection dimension increases. Our results could be seen, in some respects, as a generalization of work by [3] that considers m-ary hypothesis testing to identify a signal from a few measurements against a known collection of prototypes.

1.1. The supervised learning problem

In a supervised learning problem we observe N examples of labelled training data $\mathcal{T}_N = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $(\mathbf{x}_i, y_i) \overset{i.i.d}{\sim} \mathcal{D}$ some (usually unknown) distribution with $\mathbf{x}_i \sim \mathcal{D}_x \subseteq \mathbb{R}^d$ and $y_i \sim \mathcal{C}$, where $\mathcal{C}$ is a finite collection of class labels partitioning $\mathcal{D}$. 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 $\mathcal{L}(\hat{h}) = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathbf{x}_q}[\mathcal{L}(h)]$, where $\mathbf{x}_q \sim \mathcal{D}_x$ is a query point. Here we use the $(0,1)$-loss $\mathcal{L}_{(0,1)}$ as our measure of performance.

In the setting we consider here, the class of functions $\mathcal{H}$ consists of instantiations of FLD learned on randomly-projected data, $\mathcal{T}_N = \{(R\mathbf{x}_i, y_i) : R\mathbf{x}_i \in \mathbb{R}^k, \mathbf{x}_i \sim \mathcal{D}_x; y_i \in \{0,1\}\}_{i=1}^N$, and we bound the probability that the projection of a previously unseen query point $R\mathbf{x}_q$ with its true class label y_q unknown is misclassified by the learned classifier.

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. If $\Sigma = \Sigma_0 = \Sigma_1$ and μ_0 and μ_1

are known, then the optimal classifier is given by Bayes' rule [5]:

$$h(\mathbf{x}_q) = \mathbf{1}\left\{\log\frac{f_1(\mathbf{x}_q)}{f_0(\mathbf{x}_q)} > 0\right\}$$
$$= \mathbf{1}\left\{(\mu_1-\mu_0)^T\Sigma^{-1}\left(\mathbf{x}_q - \frac{(\mu_0+\mu_1)}{2}\right) > 0\right\}$$

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:

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

For the true data distribution, we will allow different class covariance matrices which we shall denote by Σ_y, although the FLD model constrains the covariance estimates to be shared across the classes, i.e. $\hat{\Sigma}_y = \hat{\Sigma}, \forall y$. Furthermore we will not constrain the priors π_y to be identical in the true data distribution, although the model priors $\hat{\pi}_y$ will be taken to be equal. The parameters μ_y and Σ are estimated from a given set of training data.

1.3. Random Projection

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 sub-Gaussian distribution [7, 8, 9]. Note that therefore R almost surely (or for the sub-Gaussians given in [8], with high probability) has rank k.
Theoretical treatments of RP frequently assume that the rows of R have been orthonormalised, but in practice if the original data dimensionality d is very high this may not be necessary [10, 11, 12] as the rows of R, treated as random vectors, will almost surely 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 1.1. *Let* $\mathbf{s}$ *and* $\mathbf{t}$ *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* $\langle \mathbf{s}, \mathbf{t}\rangle$ *denote the inner product of* $\mathbf{s}$ *and* $\mathbf{t}$*. Then:*

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\left\langle\frac{\mathbf{s}}{\|\mathbf{s}\|},\frac{\mathbf{t}}{\|\mathbf{t}\|}\right\rangle = 0\right\} = 1 \tag{1.1}$$

and

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\|\mathbf{s}\|}{\|\mathbf{t}\|} = 1\right\} = 1 \tag{1.2}$$

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

$$Pr_{\mathbf{s}}\left\{\lim_{d\to\infty}\frac{\sum_{i=1}^{d}s_i^2}{d}=\sigma^2\right\}=Pr_{\mathbf{s}}\left\{\lim_{d\to\infty}\sqrt{\frac{\sum_{i=1}^{d}s_i^2}{d}}=\sigma\right\} \tag{1.3}$$

and this probability is equal to 1 by applying the strong law of large numbers for i.i.d random variables (e.g. [13] Thm. 5.4.4 Pg 62) to the LHS of (1.3). A similar argument shows that $\|\mathbf{t}\|/\sqrt{d}$ 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:

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{d}=0\right\}=Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\sum_{i=1}^{d}s_it_i}{d}=0\right\}=1 \tag{1.4}$$

We may rewrite (1.4) as:

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\sum_{i=1}^{d}s_it_i}{d}=0\right\}=Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{\|\mathbf{s}\|\|\mathbf{t}\|}\cdot\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d}=0\right\} \tag{1.5}$$

we will prove (1.1) by showing that $\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d} \xrightarrow{a.s.} \sigma^2 \in (0,\infty)$ and hence conclude that $\frac{\langle\mathbf{s},\mathbf{t}\rangle}{\|\mathbf{s}\|\|\mathbf{t}\|} \xrightarrow{a.s.} 0$.
Utilising the independence of $\mathbf{s}$ and $\mathbf{t}$ we see, via the strong law and by applying the product rule for limits of continuous functions to (1.3), that:

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d}=\sigma^2\right\}=1 \tag{1.6}$$

Indeed, negating and combining (1.4) and (1.6), via the union bound we observe:

$$\begin{aligned}&Pr_{\mathbf{s},\mathbf{t}}\left\{\left(\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{d}\neq 0\right)\vee\left(\lim_{d\to\infty}\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d}\neq\sigma^2\right)\right\}\\ \leqslant\ &Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{d}\neq 0\right\}+Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d}\neq\sigma^2\right\}=0+0=0\end{aligned} \tag{1.7}$$

and so:

$$\begin{aligned}&Pr_{\mathbf{s},\mathbf{t}}\left\{\left(\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{d}=0\right)\wedge\left(\lim_{d\to\infty}\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d}=\sigma^2\right)\right\}\\ \geqslant\ &1-\left(Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{d}\neq 0\right\}+Pr_{\mathbf{s},\mathbf{t}}\lim_{d\to\infty}\left\{\frac{\|\mathbf{s}\|\|\mathbf{t}\|}{d}\neq\sigma^2\right\}\right)=1\end{aligned} \tag{1.8}$$

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

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\langle\mathbf{s},\mathbf{t}\rangle}{\|\mathbf{s}\|\|\mathbf{t}\|}\cdot\sigma^2=0\right\}=Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\left\langle\frac{\mathbf{s}}{\|\mathbf{s}\|},\frac{\mathbf{t}}{\|\mathbf{t}\|}\right\rangle=0\right\}=1 \tag{1.9}$$

as required.
To prove the almost sure convergence of norms (1.2) *we again use equation* (1.3) *and the fact that* $\|\mathbf{s}\|/\sqrt{d}$ *and* $\|\mathbf{t}\|/\sqrt{d}$ *converge almost surely to* σ*. Then applying the quotient rule for limits, we have (since* $\sigma\neq 0$*):*

$$Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\left(\frac{\sum_{i=1}^{d}s_i^2}{d}\right)^{1/2}}{\left(\frac{\sum_{i=1}^{d}t_i^2}{d}\right)^{1/2}}=1\right\}=Pr_{\mathbf{s},\mathbf{t}}\left\{\lim_{d\to\infty}\frac{\|\mathbf{s}\|}{\|\mathbf{t}\|}=1\right\}=1 \tag{1.10}$$

as required. □

Table 1: Notation used in this paper

Random vector	$\mathbf{x}$
Observation/class label pair	$(\mathbf{x}_i, y_i)$
Query point (unlabelled observation)	$\mathbf{x}_q$
Random projection matrix	R
'Data space' - real vector space of d dimensions	$\mathbb{R}^d$
'Projected space' - real vector space of $k \leqslant d$ dimensions	$\mathbb{R}^k$
True class label of the query point $\mathbf{x}_q$	y_q
Mean of class $y \in \mathcal{C}$	μ_y
Sample mean of class $y \in \mathcal{C}$	$\hat{\mu}_y$
Covariance matrix of the sub-Gaussian distribution $\mathcal{D}_{\mathbf{x}\mid y}$	Σ_y
Assumed model covariance matrix of $\mathcal{D}_{\mathbf{x}\mid y}$	$\hat{\Sigma}$

2. Results

The following theorem bounds the estimated probability of misclassification error of FLD in a random projection of the data space, on average over the random choices of the projection matrix.
Throughout we assume that $\mathbf{x} \sim \mathcal{D}_x$ where $\mathcal{D}_x = \sum_y \pi_y \cdot \mathcal{D}_{x|y}$, $\pi_y = \Pr(y = y_q)$ where y_q denotes the unknown (true) class to which $\mathbf{x}_q$ belongs and $\mathcal{D}_{x|y}$ is any sub-Gaussian distribution i.e. any distribution whose moment generating function is dominated by that of a Gaussian. This class of distributions is quite rich and includes, for example, Gaussian distributions and any distribution with bounded support [14].
We bound the expected error of the classifier learned in the randomly projected space in terms of quantities in the original data space making use of the fact that, by linearity of the expectation operator and of R, the sample mean and the true mean of a projected data class coincide with the projection of the corresponding means of the original data.

In the statement of theorem 2.1, the condition $\alpha_y^R > 0$ simply says that in the projected space the estimated and true means for class y both lie on the same side of the decision hyperplane as one another.[1] Note that there is no assumption made regarding sparsity of the data.

Theorem 2.1. *Let $\mathbf{x}_q \sim \mathcal{D}_{\mathbf{x}}$ and let $R \in \mathcal{M}_{k\times d}$, $k \leqslant d$, be a random projection matrix with entries drawn i.i.d from the univariate Gaussian $\mathcal{N}(0, 1/d)$.*
Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}^R$ be the instance learned from the randomly projected training data $\mathcal{T}_N = \{(R\mathbf{x}_i, y_i) \in \mathbb{R}^k \times \mathcal{C}\}_{i=1}^N$, where $\mathcal{C} = \{0, 1\}$.
Assume that $\alpha_y^R = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_{\neg y} - \hat{\mu}_y) > 0$.
Then the estimated misclassification error $Pr_{R,\mathbf{x}_q}[\hat{h}^R(R\mathbf{x}_q) \neq y_q | \alpha_y^R > 0, \forall y \in \mathcal{C}]$ is bounded above by:

$$Pr_{R,\mathbf{x}_q}[\hat{h}^R(R\mathbf{x}_q) \neq y | \alpha_y^R > 0, \forall y \in \mathcal{C}] \leqslant$$

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{k}{2}\log\left(1 + \frac{1}{4d}\cdot\|\hat{\mu}_y - \hat{\mu}_{\neg y}\|^2 \cdot \frac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right)\right) \tag{2.1}$$

Here, μ_y is the (true) mean of the class from which $\mathbf{x}_q$ was drawn with (true) covariance matrix Σ_y, $\mu_{\neg y}$ is the mean of the class from which $\mathbf{x}_q$ was not drawn, and $\hat{\mu}_y, \hat{\mu}_{\neg y}$ are the corresponding estimated class means with model covariance $\hat{\Sigma}$.

2.0.1. Remark

Note that in theorem 2.1, and from now on in this paper, we consider finite random projection matrices with entries drawn i.i.d from the Gaussian $\mathcal{N}(0, 1/d)$, i.e. with a fixed value of d. This choice for the variance is simply in order to ensure that the rows of the matrix have expected norm 1 which simplifies our exposition slightly. In practice, any zero-mean Gaussian with fixed finite variance may be used to generate the entries of R without affecting any of the conclusions drawn here and, in particular, without affecting the form of the bound given in theorem 2.1.

2.1. Structure of paper

The remainder of this paper is structured as below. First we prove our main theorem 2.1 as follows:
We commence by bounding the error probability of FLD in the data space. The error of FLD in the data space has, of course, been studied previously and the exact probability of misclassification error can be found in [5, 6] for example. However, the exact expression of error relies on Gaussianity of the classes

[1] In section 2.5 of this paper we will relax this condition for a specific case of interest, namely when $\hat{\Sigma}$ is chosen to be spherical.

and, furthermore, it would make our subsequent derivation of a bound for the average-case analysis in the projected space (w.r.t. all random choices of R) analytically intractable. We therefore derive an upper bound on the error in the data space which addresses both of these issues and has the same qualitative properties as the exact error. Our data space bound has the further advantage that it is more general than those in [5, 6], in the sense that it is also good for both Gaussian and sub-Gaussian distributions (i.e. distributions where the tails decay more quickly than in the Gaussian case). Finally we use our data space bound on the error probability to derive a bound in any fixed randomly projected space, followed by computing the expected error over every projected space w.r.t the random choices of R. The proof of our main result, the above theorem 2.1, then follows.

A setting of interest is when $\hat{\Sigma}$ is spherical, since previous studies have shown that the covariance in the projected space becomes more spherical than the original data space covariance [15, 12, 16]. Relating the bound given in theorem 2.1 with our bound in theorem 4.8 of [16] we see that the bound presented here is of simpler form, yet when $\hat{\Sigma}$ is chosen to be spherical it is tighter than our other bound. We therefore use theorem 2.1 to analyse in more depth the case of spherical model covariance in this paper.

We next consider the likelihood that the condition on our theorem 2.1 is violated. This is a particular finite sample effect due to the random projection of the given finite sample, namely the possibility of the true mean being 'flipped' by random projection across the decision boundary in the projected space with a corresponding increase in classifier error. In the spherical model covariance setting the probability of such a 'flip', namely $\Pr[\alpha_y^R \leqslant 0 | \alpha_y > 0]$ (where $\alpha_y > 0$ is the condition corresponding to $\alpha_y^R > 0$ in the data space) can be evaluated exactly, and the probability of this event happening turns out to be both independent of the original data dimensionality and exponentially decreasing w.r.t k. In section 2.5 we incorporate the 'flip' probability into our estimated error bound, yielding the unconditional probability of error for FLD in the randomly projected domain.

Finally, we summarise our findings, and discuss possible future directions and some open problems.

We now commence the sequence of arguments leading to the proof of our main result:

2.2. Bound on two-class FLD in the data space

Lemma 2.2. *(Bound on two-class FLD in the data space) Let $\mathbf{x}_q \sim \mathcal{D}_x$. 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 there is sufficient training data so that $\alpha_y = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T \hat{\Sigma}^{-1} (\hat{\mu}_{\neg y} - \hat{\mu}_y)$ is positive.*[2] *Then the probability that $\mathbf{x}_q$ is misclassified*

[2] This simply means that the estimated and true means for class y both lie on the same side of the decision hyperplane as one another.

is given by $Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y_q | \alpha_y > 0,\ \forall y \in \mathcal{C}] \leqslant$

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T \hat{\Sigma}^{-1} (\hat{\mu}_{\neg y} - \hat{\mu}_y) \right]^2}{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^{-1} \Sigma_y \hat{\Sigma}^{-1} (\hat{\mu}_{\neg y} - \hat{\mu}_y)} \right)$$

with $\pi_y = Pr(y = y_q)$, μ_y *the mean of the class from which* $\mathbf{x}_q$ *was drawn, estimated class means* $\hat{\mu}_y$ *and* $\hat{\mu}_{\neg y}$, *and model covariance* $\hat{\Sigma}$.

The proof of the data space bound uses standard Chernoff-bounding techniques and is given in the appendix 3.1. In the case of Gaussian classes we can do better and directly bound the exact error of FLD derived in [5, 6] which is:

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

where $\Phi(\cdot)$ is the Gaussian CDF. Then using equation (13.48) of [17] which bounds this quantity we get:

$$\Phi(-x) = 1 - \Phi(x) \leqslant 1 - \frac{1}{2}\left[1 + \sqrt{1 - e^{-x^2/2}}\right] \leqslant \frac{1}{2}\exp(-x^2/2) \tag{2.2}$$

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 lemma 2.2.
Returning to the sub-Gaussian case, for interpretability we now identify the contributions to the misclassification error from the estimated error and from the effect of finite samples separately. It is therefore useful to reformulate our bound by decomposing it into two terms, one of which will go to zero as the number of training examples increases.

Lemma 2.3. *(Decomposition of the two-class bound) Let* $\mathbf{x}_q \sim \mathcal{D}_x$. *Let* $\mathcal{H}$ *be the class of FLD functions and let* $\hat{h}$ *be the instance learned from the training data* $\mathcal{T}_N$. *Write for the estimated error:*

$$\hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma_0, \Sigma_1) = \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma_y \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)} \right) \tag{2.3}$$

and define:

$$B_y(\hat{\mu}_0, \hat{\mu}_1, \mu_0, \mu_1, \hat{\Sigma}, \Sigma_y) = \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T \hat{\Sigma}^{-1} (\hat{\mu}_{\neg y} - \hat{\mu}_y) \right]^2}{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^{-1} \Sigma_y \hat{\Sigma}^{-1} (\hat{\mu}_{\neg y} - \hat{\mu}_y)} \right)$$

taken from the right hand side of lemma 2.2. Then, the whole of the right hand side of lemma 2.2, i.e. $\pi_0 B_0 + \pi_1 B_1$ *is bounded above by:*

$$\leqslant \hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma_0, \Sigma_1) + \max_{y,i} \sup \left\{ \left| \frac{\partial B_y}{\partial \mu_{yi}} \right| \right\} \cdot \sum_{y \in \{0,1\}} \pi_y \sum_i |\hat{\mu}_{yi} - \mu_{yi}| \tag{2.4}$$

with μ_y the mean of the class from which $\mathbf{x}_q$ was drawn, estimated class means $\hat{\mu}_y$ with $\hat{\mu}_{yi}$ the i-th component, and model covariance $\hat{\Sigma}$.

Proof. We will use the mean value theorem (see appendix, lemma 3.5) so we start by differentiating B_y with respect to μ_y to find $\nabla_{\mu_y} B_y = B_y \cdot \frac{1}{2}\kappa_y \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} - \hat{\mu}_y)$, where $\kappa_y = \alpha_y/(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^{-1} \Sigma_y \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} - \hat{\mu}_y)$, is the optimal parameter minimizing lemma 2.2 [16]. Since the exponential term is bounded between zero and one, the supremum of the i-th component of this gradient exists provided that $\|\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y\| < \infty$ and $\|\hat{\mu}_{\neg y} - \hat{\mu}_y\| < \infty$. So we have that

$$B_y \leqslant \hat{B}_y(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma_0, \Sigma_1) + \max_i \sup_{\tilde{\mu}_y \in \gamma} \left\{ \left| \frac{\partial B_y}{\partial \mu_{yi}}_{|_{\mu_y = \tilde{\mu}_y}} \right| \right\} \sum_i |\hat{\mu}_{yi} - \mu_{yi}|$$

where γ is the line between $\hat{\mu}_y$ and μ_y and $\tilde{\mu}_y = t_y \hat{\mu}_y + (1 - t_y)\mu_y$, $t_y \in [0, 1]$. Plugging this back in to the right hand side of lemma 2.2, i.e. $\pi_0 B_0 + \pi_1 B_1$, and then taking the maximum over both classes yields the desired result. □

We call the two terms obtained in (2.4) the 'estimated error' and 'estimation error' respectively. Note that the estimation error decays exponentially with the size of the training set (which can be seen using standard Chernoff bounding techniques) and so, as the number of observations becomes large, the estimated error of the classifier approximates the Bayes' error.

We now have the framework in place to bound the estimated misclassification probability if we choose to work with a k-dimensional random projection of the original data. We first obtain a bound that holds for any fixed random projection matrix R, and finally on average over all R.

2.3. Bound on two-class FLD in the projected space

Proof. (of Theorem 2.1) Denote the sample mean and the true mean of a projected data class by $\hat{\mu}^R$ and μ^R respectively. From the linearity of the expectation operator and of R, these coincide with the projection of the corresponding means of the original data: $\hat{\mu}^R = \frac{1}{N}\sum_{i=1}^N R(\mathbf{x}_i)$, and $\mu^R = R\mu$. Using these, if Σ_y is the covariance matrix of the y-th class in the data space, then its projected counterpart $(\Sigma_y)_R$ is $R\Sigma_y R^T$, and likewise $\hat{\Sigma}_R = R\hat{\Sigma}R^T$.

By lemma 2.2, the estimated error in the projected space defined by any given R is now upper bounded by:

$$\begin{aligned} &\Pr_{\mathbf{x}_q}\left\{\hat{h}(R\mathbf{x}_q) \neq y_q | \alpha_y^R > 0,\ \forall y \in \mathcal{C}\right\} \\ &\leqslant \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \cdot \frac{\left[(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)^T \hat{\Sigma}_R^{-1}(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)\right]^2}{(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)^T \hat{\Sigma}_R^{-1} (\Sigma_y)_R \hat{\Sigma}_R^{-1}(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)} \right) \end{aligned} \quad (2.5)$$

We would like to analyse the expectation of (2.5) 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 (2.5) using majorization of the numerator by the Rayleigh quotient (lemma 3.1 in the appendix), where we take $\mathbf{v} = \left(\hat{\Sigma}_R\right)^{-1/2}(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)$ and take, for the y-th class, our positive definite Q_y to be $Q_y = \hat{\Sigma}_R^{-1/2}(\Sigma_y)_R\hat{\Sigma}_R^{-1/2}$ and we use the fact that since $\hat{\Sigma}_R^{-1}$ is symmetric positive definite it has a unique symmetric positive semi-definite square root $\hat{\Sigma}_R^{-1/2} = \left(\hat{\Sigma}_R^{-1}\right)^{1/2} = \left(\hat{\Sigma}_R^{1/2}\right)^{-1} = \left(\hat{\Sigma}_R^{-1/2}\right)^T$ ([18], Theorem 7.2.6, pg. 406). Then, we have (2.5) is less than or equal to:

$$\sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\frac{\left[(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)\right]^2}{\lambda_{\max}(Q_y)(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)}\right) \tag{2.6}$$

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 ([19], Theorem A.6.2, pg. 468), for each term in the summation we may write $\lambda_{\max}(Q_y) = \lambda_{\max}(\hat{\Sigma}_R^{-1/2}(\Sigma_y)_R\hat{\Sigma}_R^{-1/2}) = \lambda_{\max}(\hat{\Sigma}_R^{-1}(\Sigma_y)_R)$ and we may now bound equation (2.5) from above with:

$$\sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\frac{(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R)}{\lambda_{\max}((\Sigma_y)_R\hat{\Sigma}_R^{-1})}\right) \tag{2.7}$$

$$\leqslant \sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\frac{\|\hat{\mu}_{\neg y}^R - \hat{\mu}_y^R\|^2}{\lambda_{\max}(\hat{\Sigma})}\frac{1}{\lambda_{\max}((\Sigma_y)_R\hat{\Sigma}_R^{-1})}\right) \tag{2.8}$$

where in the last line we used minorization by Rayleigh quotient of the numerator and applied Poincaré separation theorem to $\hat{\Sigma}_R^{-1}$ (see Appendix lemma 3.3).

It now remains to deal with the term $\lambda_{\max}((\Sigma_y)_R\hat{\Sigma}_R^{-1})$. We see this encodes a measure of how well the form of the model covariance matches the true covariance(s), and the bound is tightest when the match is closest. Note that this term is not simply a function of the training set size, but also of the (diagonal, or spherical) constraints that it is often convenient to impose on the model covariance.

Continuing the proof, we now use lemma (2.4) (which we give shortly) to upper bound equation (2.8) by:

$$\sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\frac{\|R(\hat{\mu}_{\neg y} - \hat{\mu}_y)\|^2}{\lambda_{\max}(\hat{\Sigma})}\frac{1}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right) \tag{2.9}$$

$$= \sum_{y=0}^{1}\pi_y \exp\left(-\frac{1}{8}\cdot\|R(\hat{\mu}_{\neg y} - \hat{\mu}_y)\|^2\cdot\frac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})}\right) \tag{2.10}$$

The change of term in the denominator uses the fact that $\lambda_{\max}((\Sigma_y)_R\hat{\Sigma}_R^{-1}) \leqslant \lambda_{\max}(\Sigma_y\hat{\Sigma}^{-1})$, which we now prove in the following lemma:

Lemma 2.4 (Eigenvalues of projected matrix products)**.** *Let $\hat{\Sigma}$ and Σ be symmetric positive definite, and let R be a $k \times d$ matrix with rank k. Then:*

$$\lambda_{\max}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}) \leqslant \lambda_{\max}(\hat{\Sigma}^{-1}\Sigma) = \lambda_{\max}(\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2})$$

First, by lemma 3.1:

$$\lambda_{\max}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{2.11}$$

$$= \max_{\mathbf{u}\in\mathbb{R}^k}\left\{\frac{\mathbf{u}^T[R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}\right\} \tag{2.12}$$

Writing $\mathbf{v} = [R\hat{\Sigma}R^T]^{-1/2}\mathbf{u}$ so that $\mathbf{u} = [R\hat{\Sigma}R^T]^{1/2}\mathbf{v}$ then we may rewrite the expression (2.12)*, as the following:*

$$= \max_{\mathbf{v}\in\mathbb{R}^k}\left\{\frac{\mathbf{v}^TR\Sigma R^T\mathbf{v}}{\mathbf{v}^TR\hat{\Sigma}R^T\mathbf{v}}\right\} \tag{2.13}$$

Writing $\mathbf{w} = R^T\mathbf{v}$, and noting that the span of all possible vectors $\mathbf{w}$ is a k-dimensional subspace of $\mathbb{R}^d$, we can bound the expression 2.13 by allowing the maximal vector $\mathbf{w} \in \mathbb{R}^d$ not to lie in this subspace:

$$\leqslant \max_{\mathbf{w}\in\mathbb{R}^d}\left\{\frac{\mathbf{w}^T\Sigma\mathbf{w}}{\mathbf{w}^T\hat{\Sigma}\mathbf{w}}\right\} \tag{2.14}$$

Now put $\mathbf{y} = \hat{\Sigma}^{1/2}\mathbf{w}$, with $\mathbf{y} \in \mathbb{R}^d$. This $\mathbf{y}$ exists uniquely since $\hat{\Sigma}^{1/2}$ is invertible, and we may rewrite (2.14) as the following:

$$= \max_{\mathbf{y}\in\mathbb{R}^d}\left\{\frac{\mathbf{y}^T\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right\} \tag{2.15}$$

$$= \lambda_{\max}(\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}) = \lambda_{\max}(\hat{\Sigma}^{-1}\Sigma) \tag{2.16}$$

This completes the proof of the lemma.

The bound in (2.10) holds deterministically, for any fixed projection matrix R. We can also see from (2.10) 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, this may imply considerable variability with the random choice of R, and we are more interested in the misclassification probability on average over all random choices of R.

Since the entries of R where drawn i.i.d from $\mathcal{N}(0, 1/d)$ the term $\|R(\hat{\mu}_1 - \hat{\mu}_0)\|^2$ is χ^2_k distributed, and we can compute this expectation via the moment generating function of independent χ^2 variables as follows:

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

which, noting that under the $(0,1)$-loss the probability of an error coincides with the expected error, finally yields the Theorem 2.1. $\square$

2.3.1. Remark

In making the step from (2.7) to (2.8) we used Poincaré inequality, and this is in fact the only point in our proof where we use the condition that the rows of R are orthogonal. In fact, here we could equally well have employed the theorem of De Bruijn given in the appendix (3.4) and instead shown that:

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \cdot \frac{(\hat{\mu}^R_{\neg y} - \hat{\mu}^R_y)^T \hat{\Sigma}_R^{-1} (\hat{\mu}^R_{\neg y} - \hat{\mu}^R_y)}{\lambda_{\max}((\Sigma_y)_R \hat{\Sigma}_R^{-1})}\right) \tag{2.17}$$

$$\leqslant \sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \cdot \|\hat{\mu}^R_{\neg y} - \hat{\mu}^R_y\|^2 \cdot \frac{\lambda_{\min}(\hat{\Sigma}^{-1}) \lambda_{\min}([RR^T]^{-1})}{\lambda_{\max}((\Sigma_y)_R \hat{\Sigma}_R^{-1})}\right) \tag{2.18}$$

Now, we see when the rows of R have been orthonormalised then $RR^T = \mathbf{I}$ is the identity, and we recover (2.8). In turn, when the rows of R are *not* orthonormalised, then lemma 1.1 suggests that for large d equation (2.8) still holds approximately. This intuition can be made more concrete using existing results on the deviation of the eigenvalues of RR^T from their expectation. For example, equation (2.3) of [20] gives the following high probability bound on the least and greatest singular values $s_{\min}$ and $s_{\max}$ of the random projection matrix P with entries drawn from the standard Gaussian $\mathcal{N}(0,1)$. For all $\epsilon > 0$:

$$\Pr_R\left(\sqrt{d} - \sqrt{k} - \epsilon \leqslant s_{\min}(P) \leqslant s_{\max}(P) \leqslant \sqrt{d} + \sqrt{k} + \epsilon\right) \geqslant 1 - 2e^{-\epsilon^2/2} \tag{2.19}$$

Using the facts that $\left[RR^T\right]^{-1}$ is invertible with smallest eigenvalue $\lambda_{\min}\left(\left[RR^T\right]^{-1}\right) = 1/\lambda_{\max}\left(RR^T\right)$ and that our random projection matrix R has entries drawn from $\mathcal{N}(0, 1/d)$ we see that:

$$\Pr_R\left(\lambda_{\max}(RR^T) \leqslant (1 + \sqrt{k/d} + \epsilon)^2\right) \geqslant 1 - e^{-\epsilon^2 d/2} \tag{2.20}$$

In particular, with probability at least $1 - e^{-\epsilon^2 d/2}$ we have that the estimated error of randomly projected FLD is no more than:

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \cdot \|\hat{\mu}^R_{\neg y} - \hat{\mu}^R_y\|^2 \cdot \frac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}((\Sigma_y)_R \hat{\Sigma}_R^{-1})(1 + \sqrt{k/d} + \epsilon)^2}\right). \tag{2.21}$$

2.4. Relation to theorem 4.8 in [16]

We find it interesting to briefly compare the upper bound we give here on the average estimated error of randomly projected FLD with the bound we give in theorem 4.8 of our KDD 2010 paper [16]. Our KDD bound has the same

preconditions as our theorem (2.1) presented here, but has (with the appropriate notational changes) the following different form:

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{k}{2}\log\left(1+\frac{1}{4d}\cdot\|\mu_{\neg y}-\mu_y\|^2\cdot\frac{g(\Sigma_y\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma_y)}\right)\right) \tag{2.22}$$

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 equations (2.22) and (2.1) that the two bounds differ in that (2.22) has the function $f_1(\hat{\Sigma},\Sigma) \stackrel{\text{def}}{=} g(\Sigma\hat{\Sigma}^{-1})/\lambda_{\max}(\Sigma)$ in the bound whereas here in (2.1) we have $f_2(\hat{\Sigma},\Sigma) \stackrel{\text{def}}{=} \lambda_{\min}(\hat{\Sigma}^{-1})/\lambda_{\max}(\Sigma\hat{\Sigma}^{-1})$. 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 f_1 and f_2 take their maximum value when $\hat{\Sigma} = \Sigma$ and then we have:

$$f_1^* = f_1(\hat{\Sigma}=\Sigma,\Sigma) = \frac{1}{\lambda_{\max}(\Sigma)} = f_2(\hat{\Sigma}=\Sigma,\Sigma) = f_2^* \tag{2.23}$$

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 most useful for studying the effects of covariance misspecification in the projected space, as we demonstrated in [16].

On the other hand, the bound given here in theorem 2.1 is quantitatively sharper in particular covariance settings, most notably it also takes its best value when $\hat{\Sigma}$ is chosen to be spherical. Indeed in this case the $\lambda_{\max}$ term in the denominator of (2.1) factorises and we have

$$f_2(\hat{\Sigma}=\mathbf{I},\Sigma) = \frac{1}{\lambda_{\max}(\Sigma)} = f_1^* \tag{2.24}$$

since the $\lambda_{\min}(\hat{\Sigma}^{-1})$ in the numerator cancels with the $\lambda_{\max}(\hat{\Sigma}^{-1})$ in the denominator. Hence, the bound presented here is better suited to studying the spherical model setting in more detail.

Furthermore, this setting is one of particular interest since our analysis in [16] 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 paper will consider randomly projected FLD in the spherical model setting in more depth. In the following section we will show that in this setting we can bound the average estimated error tightly even if we relax the condition required on theorem 2.1. Figure 1 gives a comparison of the bounds of theorem 2.1 and theorem 4.8 in [16] 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 identical Gaussian classes and averaged over 2500 random projections. The data dimensionality is $d = 100$ in each case, the projected dimensionality $k \in \{1, 10, 20, \ldots, 100\}$ and the training set size in each case is 40 observations (i.e. fewer observations than the data dimensionality d). The constant $c \stackrel{\text{def}}{=} \frac{\|\mu_0 - \mu_1\|}{\sqrt{d \cdot \lambda_{\max}(\Sigma)}}$ is the class separation metric used by Dasgupta in [15, 12].

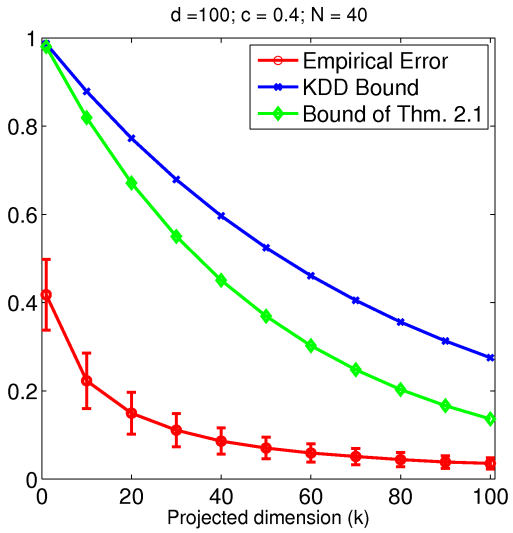

Figure 1: Comparison of our bound from theorem 2.1 given here to our bound given in theorem 4.8 of [16] against empirical estimates of misclassification error for two identical $c = 0.4$-separated Gaussian classes with $\lambda_{\max}(\Sigma) \simeq 4$. We ran 2500 trials, fixing the data dimensionality at $d = 100$ while k varies. Error bars mark one standard deviation. In each trial the training set size was only 40 and we estimated the empirical error from 2000 randomly drawn query points each time.

2.5. The mean flipping problem

In theorem 2.1 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 $\hat{\Sigma}$ is spherical, however in simulations (which we do not show due to space constraints) 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 $\mathbf{n} = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)$ and $\mathbf{m} = (\hat{\mu}_{\neg y} - \hat{\mu}_y) \in \mathbb{R}^d$ 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 2.6. For now, we state the theorem:

Theorem 2.5 (Flip Probability)**.** *Let* $\mathbf{n},\ \mathbf{m} \in \mathbb{R}^d$ *with angular separation* $\theta \in [0, \pi/2]$.
Let $R \in \mathcal{M}_{k\times d}$ *be a random projection matrix with entries* $r_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/d)$ *and let* $R\mathbf{n},\ R\mathbf{m} \in \mathbb{R}^k$ *be the projections of* $\mathbf{n},\ \mathbf{m}$ *into* $\mathbb{R}^k$ *with angular separation* θ_R.
Then the 'flip probability' $Pr_R[\theta_R > \pi/2|\theta] = Pr_R[(R\mathbf{n})^T R\mathbf{m} < 0|\mathbf{n}^T\mathbf{m} \geqslant 0] = Pr_R[\alpha_y^R < 0|\alpha_y \geqslant 0]$ *is given by:*

$$Pr_R[\theta_R > \pi/2|\theta] = \frac{\int_0^\theta \sin^{k-1}(\phi)\ d\phi}{\int_0^\pi \sin^{k-1}(\phi)\ d\phi} \tag{2.25}$$

The proof of theorem (2.5) is technical and is given elsewhere [21]. Note particularly the surprising fact that the flip probability in theorem 2.5 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 (2.25) 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 [21], and this quantity is bounded above by $\exp\left(-\frac{1}{2}k\cos^2(\theta)\right)$ ([22], Lemma 2.2, Pg 11).
This result seems quite counterintuitive, especially the fact that the flip probability is independent of the original data dimensionality. 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 $\mathbf{m}, \mathbf{n}$. For each (k, d, θ) tuple, we generated 5000 $k\times d$ random projection matrices R with which we randomly projected $\mathbf{m}$ and $\mathbf{n}$. Finally we counted the number of times, N, that the dot product $(R(\mathbf{m}))^T R(\mathbf{n}) < 0$ and estimated the flip probability by $N/5000$.
We give plots of the results: Figure 2 shows the close match between our theoretical values and empirical estimates of the flip probabilities, while figure 3

gives empirical validation of the fact that the flip probability is independent of d.

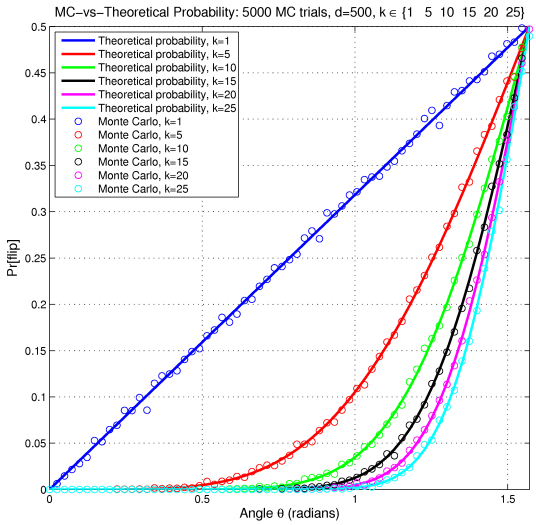

Figure 2: Experiment illustrating match between probability calculated in theorem 2.5 and empirical trials. We fixed $d = 500$ and allowed k to vary. For each trial we generated two random unit-length vectors in $\mathbb{R}^{500}$ with angular separation θ and for each (θ, k) pair we randomly projected 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.

2.6. Corollary to theorems (2.1) *and* (2.5)

Taking into account the flip probability, we may now give the following bound on the unconditional probability of the estimated error.

Corollary 2.6. *Let* $\mathbf{x}_q \sim \mathcal{D}_{\mathbf{x}}$*, and let* y_q *be its class label. 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 2.1 and 2.5 we have:*

$$Pr_{\mathbf{x}_q,R}[\hat{h}(R\mathbf{x}_q) \neq y_q] \leqslant$$
$$\sum_{y=0}^{1} \pi_y Pr_R[\alpha_y^R > 0] \cdot \exp\left(-\frac{k}{2}\log\left(1 + \frac{1}{4d} \cdot \|\hat{\mu}_y - \hat{\mu}_{\neg y}\|^2 \cdot \frac{1}{\lambda_{\max}(\Sigma_y)}\right)\right)$$
$$+ \sum_{y=0}^{1} \pi_y (1 - Pr_R[\alpha_y^R > 0]) \tag{2.26}$$

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

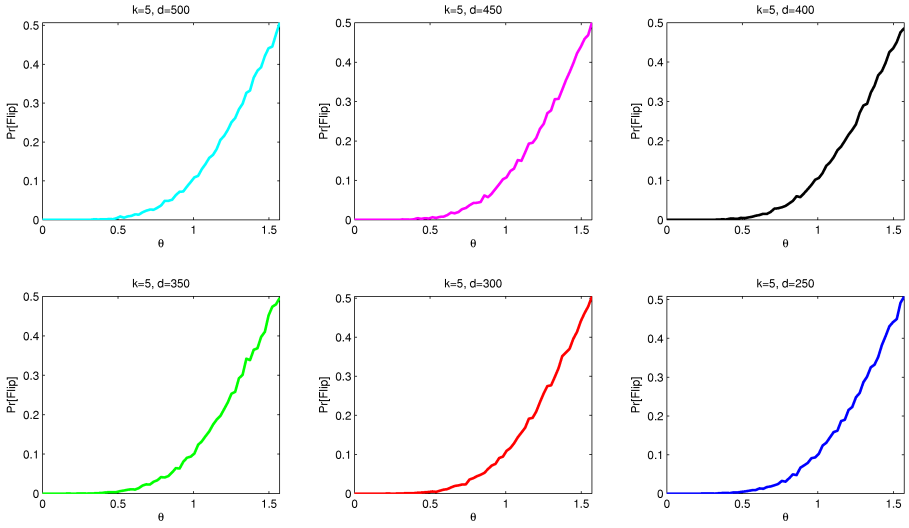

Figure 3: Experiment illustrating d-invariance of the flip probability of theorem 2.5. We fixed $k = 5$ and allowed d to vary, estimating the empirical flip probability with 5000 different random projections from $\mathbb{R}^d$ into $\mathbb{R}^5$ for each (θ, d) pair. The results for each of six choices of d are plotted on separate graphs, highlighting the similarity of the outcomes.

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

$$\begin{aligned}\Pr_{\mathbf{x}_q,R}[\hat{h}(R\mathbf{x}_q) \neq y_q] = \quad & \textstyle\sum_{y=0}^{1} \pi_y \Big(\Pr_R[\alpha_y^R > 0] \cdot \Pr_{\mathbf{x}_q,R}[\hat{h}(R\mathbf{x}_q) \neq y_q | y_q = y, \alpha_y^R > 0] \\ & +(1 - \Pr_R[\alpha_y^R > 0]) \cdot \Pr_{\mathbf{x}_q,R}[\hat{h}(R\mathbf{x}_q) \neq y_q | y_q = y, \alpha_y^R \leqslant 0]\Big)\end{aligned} \tag{2.27}$$

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 2.1 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, $k = 1$).

3. Discussion and future work

This paper presents initial findings of our ongoing work concerning the effects of dimensionality reduction on classifier performance.
Our work was motivated by 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. 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 those important distances. In the case of randomly projected 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; however 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 these more complex scenarios.
In the case of FLD we have left open the problems of data distributions where the data are not Gaussian or sub-Gaussian, or where they are not approximately linearly separable. These are the subject of our current research and we hope to present findings regarding these settings in the near future.

[1] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Rice University, 2009.

[2] M.A. Davenport, P.T. Boufounos, M.B. Wakin and R.G. Baraniuk. Signal Processing with Compressive Measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, April 2010.

[3] J. Haupt, R. Castro, R. Nowak, G. Fudge and A. Yeh. Compressive sampling for signal classification. In *Proc. 40th Asilomar Conf. on Signals, Systems, and Computers*, pages 1430–1434, 2006.

[4] O-A. Maillard and R. Munos. Compressed Least-Squares Regression. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 1213–1221, 2009.

[5] P. Bickel and E. Levina. 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.

[6] T. Pattison and D. Gossink. Misclassification Probability Bounds for Multivariate Gaussian Classes. *Digital Signal Processing*, 9:280–296, 1999.

[7] R. Arriaga and S. Vempala. An algorithmic theory of learning. *Machine Learning*, 63(2):161–182, 2006.

[8] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

[9] S. Dasgupta and A. Gupta. An Elementary Proof of the Johnson-Lindenstrauss Lemma. *Random Struct. Alg.*, 22:60–65, 2002.

[10] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In F. Provost and R. Srikant, editor, *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pages 245–250, 2001.

[11] D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 522–529. ACM, 2003.

[12] S. Dasgupta. Experiments with random projection. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pages 143–151, 2000.

[13] J.S. Rosenthal. *A first look at rigorous probability theory*. World Scientific Pub Co Inc, 2006.

[14] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027v4 [math.PR], March 2011.

[15] S. Dasgupta. Learning Mixtures of Gaussians. In *Annual Symposium on Foundations of Computer Science*, volume 40, pages 634–644, 1999.

[16] R.J. 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.

[17] N.L. Johnson, S. Kotz and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, 2 edition, 1994.

[18] R.A. Horn and C.R. Johnson. *Matrix Analysis*. CUP, 1985.

[19] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Academic Press, London, 1979.

[20] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians , Hyderabad, India, 2010*, 2010.

[21] R.J. Durrant and A. Kabán. Flip Probabilities for Random Projections of θ-separated Vectors. Technical Report CSR-10-10, School of Computer Science, University of Birmingham, 2010.

[22] K. Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

[23] N.G. De Bruijn. Inequalities concerning minors and eigenvalues. *Nieuw Archief voor Wiskunde*, 3:18–35, 1956.

Appendix

3.1. Proof of dataspace bound

Proof. (of lemma 2.2) We prove one term of the bound using standard techniques, the other term being proved similarly.
Without loss of generality let $\mathbf{x}_q$ have label $y = 0$. If $\mathbf{x}_q$ was drawn from a multivariate Gaussian and $\hat{\pi}_0 = \hat{\pi}_1 = \frac{1}{2}$ then the probability that $\mathbf{x}_q$ is misclassified is given by:

$$\begin{aligned}
&\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y | y = 0] = \Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0] \\
=&\Pr_{\mathbf{x}_q}\left[(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right] \\
=&\Pr_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)\right) > 1\right] \text{ for all } \kappa_0 > 0 \\
\leqslant&\mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)\right)\right] \text{ by Markov inequality} \\
=&\exp\left(-\frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1)\right) \mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\mathbf{x}_q\right)\right]
\end{aligned}$$

This expectation is the moment generating function of a multivariate Gaussian and so:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\mathbf{x}_q\right)\right] \\
&= \exp\left(\mu_0^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0^2 \hat{\Sigma}^{-1}\Sigma_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\right) \quad (3.1)
\end{aligned}$$

where μ_0 is the true mean, and Σ_0 is the true covariance matrix, of $\mathcal{D}_{x|0}$. Hence the probability of misclassification is bounded above by the following:

$$\begin{aligned}
\exp\Bigg(&-\frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1) + \mu_0^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \ldots \\
&\ldots + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0^2 \hat{\Sigma}^{-1}\Sigma_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\Bigg)
\end{aligned}$$

Optimising $\kappa_0 > 0$ gives:

$$\kappa_0 = \frac{(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}{2(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\Sigma_0\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)} \tag{3.2}$$

which is strictly positive as required, since the denominator is always positive (Σ_0 is positive definite, then so is $\hat{\Sigma}^{-1}\Sigma_0\hat{\Sigma}^{-1}$), and the numerator is taken to be positive as a precondition in the theorem.
Substituting this κ_0 back into the bound then yields, after some algebra:

$$\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0] \leqslant \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\Sigma_0\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right)$$

The second term, for when $\mathbf{x}_q$ belongs to the class with label 1, is derived similarly and gives:

$$\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 1] \leqslant \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_1)^T \hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1)\right]^2}{(\hat{\mu}_0 - \hat{\mu}_1)^T \hat{\Sigma}^{-1}\Sigma_1\hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}\right)$$

Finally, putting these two terms together and applying the law of total probability gives the lemma. □

3.1.1. Remarks

1. Note that in equation (3.1) it is in fact sufficient to have inequality. Therefore our bound also holds when the true distributions $\mathcal{D}_{x|y}$ of the data classes are such that they have a moment generating function dominated by that of the Gaussian i.e. sub-Gaussian distributions (distributions whose tail decays faster than that of the Gaussian).
2. To see that the requirement $\alpha_y > 0$ in lemma 2.2 is a mild one note that, because the denominator in equation (3.2) is always positive, the condition $\kappa_y > 0$ holds when:

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

   One can see that $\alpha_y > 0,\ \forall y$ is therefore equivalent to requiring that for each class the true and estimated means are both on the same side of the decision hyperplane.

Lemma 3.1 (Rayleigh quotient ([18], Theorem 4.2.2 Pg 176))**.** *If* $\mathbf{Q}$ *is a real symmetric matrix then its eigenvalues* λ *satisfy:*

$$\lambda_{\min}(\mathbf{Q}) \leqslant \frac{\mathbf{v}^T\mathbf{Q}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \leqslant \lambda_{\max}(\mathbf{Q}) \tag{3.3}$$

Lemma 3.2 (Poincaré Separation Theorem ([18], Corollary 4.3.16 Pg 190))**.** *Let* $\mathbf{S}$ *be a symmetric matrix* $\mathbf{S} \in \mathcal{M}_d$*, let* k *be an integer,* $1 \leqslant k \leqslant d$*, and let* $\mathbf{r}_1, \ldots, \mathbf{r}_k \in \mathbb{R}^d$ *be* k *orthonormal vectors. Let* $\mathbf{T} = \mathbf{r}_i^T \mathbf{S} \mathbf{r}_j = R\mathbf{S}R^T \in \mathcal{M}_k$ *(that is, in our setting the* $\mathbf{r}_i^T$ *are the rows, and the* $\mathbf{r}_j$ *the columns, of the random projection matrix* $R \in \mathcal{M}_{k \times d}$*). Arrange the eigenvalues* λ_i *of* $\mathbf{S}$ *and* $\mathbf{T}$ *in increasing magnitude, then:*

$$\lambda_i(\mathbf{S}) \leqslant \lambda_i(\mathbf{T}) \leqslant \lambda_{i+n-k}(\mathbf{S}), \quad i \in \{1, \ldots, k\} \tag{3.4}$$

and in particular:

$$\lambda_{\min}(\mathbf{S}) \leqslant \lambda_{\min}(\mathbf{T}) \text{ and } \lambda_{\max}(\mathbf{T}) \leqslant \lambda_{\max}(\mathbf{S}) \tag{3.5}$$

Lemma 3.3 (Corollary to lemmata 3.1 and 3.2.)**.** *Let* $\mathbf{Q}$ *be symmetric positive definite, such that* $\lambda_{\min}(\mathbf{Q}) > 0$ *and so* $\mathbf{Q}$ *is invertible. Let* $\mathbf{u} = R\mathbf{v},\ \mathbf{v} \in \mathbb{R}^d,\ \mathbf{u} \neq 0 \in \mathbb{R}^k$*. Then:*

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

Proof: We use the eigenvalue identity $\lambda_{\min}(\mathbf{Q}^{-1}) = 1/\lambda_{\max}(\mathbf{Q})$*. Combining this identity with lemma 3.1 and lemma 3.2 we have:*

$$\lambda_{\min}([R\mathbf{Q}R^T]^{-1}) = 1/\lambda_{\max}(R\mathbf{Q}R^T) \text{ since } R\mathbf{Q}R^T \text{ is symmetric positive definite,} \tag{3.6}$$

$$0 < \lambda_{\max}(R\mathbf{Q}R^T) \leqslant \lambda_{\max}(\mathbf{Q}) \text{ by positive definiteness and lemma 3.2} \tag{3.7}$$

$$\iff \quad 1/\lambda_{\max}(R\mathbf{Q}R^T) \geqslant 1/\lambda_{\max}(\mathbf{Q}) > 0 \tag{3.8}$$

$$\iff \quad \lambda_{\min}([R\mathbf{Q}R^T]^{-1}) \geqslant \lambda_{\min}(\mathbf{Q}^{-1}) > 0 \text{ and so:} \tag{3.9}$$

$$\mathbf{u}^T \left[R\mathbf{Q}R^T\right]^{-1} \mathbf{u} \geqslant \lambda_{\min}([R\mathbf{Q}R^T]^{-1}) \mathbf{u}^T \mathbf{u} \geqslant \lambda_{\min}(\mathbf{Q}^{-1}) \mathbf{u}^T \mathbf{u} > 0 \text{ by lemma 3.1.} \tag{3.10}$$

Lemma 3.4 (De Bruijn ([23], Theorem 14.2))**.** *Let* $\mathbf{S}$ *be a symmetric positive definite matrix* $\mathbf{S} \in \mathcal{M}_d$*, let* k *be an integer,* $1 \leqslant k \leqslant d$*, and let* R *be an arbitrary* $k \times d$ *matrix then:*

$$\lambda_{\min}(\mathbf{R}\mathbf{S}\mathbf{R}^{\mathbf{T}}) \geqslant \lambda_{\min}(\mathbf{S}) \cdot \lambda_{\min}(\mathbf{R}\mathbf{R}^{\mathbf{T}}) \tag{3.11}$$

Lemma 3.5 (Mean value theorem in several variables)**.** *Let* S *be an open subset of* $\mathbb{R}^d$ *and let* $\mathbf{x}, \mathbf{y} \in S$ *such that the line between* $\mathbf{x}$ *and* $\mathbf{y}$*,* $\gamma = \{t \cdot \mathbf{x} + (1-t) \cdot \mathbf{y}\}$*,* $t \in [0,1]$ *also lies in* S*, i.e.* $\gamma \subseteq S$*. If* $f : \mathbb{R}^d \to \mathbb{R}$ *is a continuous function which is differentiable on* S *then:*

$$f(\mathbf{y}) - f(\mathbf{x}) = (\nabla f((1-t)\mathbf{x} + t\mathbf{y}))^T (\mathbf{y} - \mathbf{x})$$

for some $t \in (0,1)$*.*
Proof: Define the function $g(t) = f((1-t)\mathbf{x} + t\mathbf{y})$*. Then, on the one hand,*

g is the restriction of f to $\gamma \subseteq S$ hence differentiable on γ, while on the other g is a real valued differentiable function of a single real variable, t, and therefore the univariate MVT asserts that $g(1) - g(0) = g'(a)$ for some $a \in (0, 1)$. Then, since $g(1) = f(\mathbf{y})$ and $g(0) = f(\mathbf{x})$, by applying the chain rule we have $g(1) - g(0) = g'(a) = (\nabla f((1-a)\mathbf{x} + a\mathbf{y}))^T(\mathbf{y} - \mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x})$ as required.