A bound on the performance of LDA in randomly projected data spaces

Robert J. Durrant and Ata Kabán
School of Computer Science, The University of Birmingham, UK
E-mail: {R.J.Durrant, A.Kaban}@cs.bham.ac.uk

Abstract

We consider the problem of classification in non-adaptive dimensionality reduction. Specifically, we bound the increase in classification error of Fisher's Linear Discriminant classifier resulting from randomly projecting the high dimensional data into a lower dimensional space and both learning the classifier and performing the classification in the projected space. 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 without requiring any sparsity structure from the data.

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 complexity issues arising from working with the data. Therefore the analysis of learning from non-adaptive data projections has received increasing interest in recent years [3],[7],[5],[8].

Here we consider the supervised learning problem of classifying a query point $\mathbf{x}_q \in \mathbb{R}^d$ as belonging to one of several Gaussian classes using Fisher's Linear Discriminant (FLD) and the misclassification error arising if, 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 $R(\mathbb{R}^d) = \mathbb{R}^k$, where $R \in \mathcal{M}_{k \times d}$ is an orthonormalized random projection matrix with entries drawn i.i.d from the Gaussian $\mathcal{N}(0, 1/d)$. Such bounds on the classification error for FLD in the data space are already known, for example those in [2, 9], but in neither of these papers is classification error in the projected domain considered; indeed in [7] 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 [1], where the authors' use of the Johnson-Lindenstrauss Lemma 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 Compressive Sensing based analysis in [3] does. Starting from first principles, and using standard techniques, we are able to exploit the class structure implied by the problem, bypassing the need to preserve all pairwise distances from the data space. Our results could be seen, in some respects, as a generalization of work by [5] 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 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 $\mathcal{L}(\hat{h}) = \underset{h \in \mathcal{H}}{\arg\min}\, \mathrm{E}_{\mathbf{x}_q}[\mathcal{L}(h)]$, where $\mathbf{x}_q \sim \mathcal{D}$ 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, $\mathcal{T}_N = \{(R(\mathbf{x}_i), y_i) : R(\mathbf{x}) \in \mathbb{R}^k, \mathbf{x} \sim \mathcal{N}(\mu_\mathbf{y}, \Sigma_y); y \in \{0,1\}\}_{i=1}^N$, and we bound the probability that the projection of an unseen query point $R(\mathbf{x}_q) : \mathbf{x}_q \sim \mathcal{D}_x = \mathcal{N}(\mu_\mathbf{y}, \Sigma_y)$ 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 π_0, $\Sigma = \Sigma_0 = \Sigma_1$ and μ_0 and μ_1 are known then the optimal classifier is given by Bayes' rule [2]:

$$h(\mathbf{x}_q) = \mathbf{1}\left\{\log\frac{(1-\pi_0)f_1(\mathbf{x}_q)}{\pi_0 f_0(\mathbf{x}_q)} > 0\right\}$$

$$= \mathbf{1}\left\{\log\left(\frac{1-\pi_0}{\pi_0}\right) + (\mu_1-\mu_0)^T\Sigma^{-1}\left(\mathbf{x}_q - \frac{(\mu_0+\mu_1)}{2}\right) > 0\right\}$$

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.

We shall assume that the observations $\mathbf{x}$ are drawn with equal probability from one of two multivariate Gaussian classes $\mathcal{D}_x = \mathcal{N}(\mu_y, \Sigma)$, for simplicity, but we must estimate μ_y and Σ from training data.

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 $k \leqslant d$ dim.	$\mathbb{R}^k$
Mean of class $y \in \mathcal{C}$	μ_y
Sample mean of class $y \in \mathcal{C}$	$\hat{\mu}_y$
Covariance matrix of the Gaussian distribution $\mathcal{D}_{x_y}$	Σ
Assumed model covariance matrix of $\mathcal{D}_{x_y}$	$\hat{\Sigma}$

2 Results

The following theorem bounds the *estimated* probability of misclassification error of LDA in the random projection space of the data, on average over the random choices of the projection matrix. Notice that this bound does not depend on the number of training points, and there is no sparsity assumption made.

Theorem 2.1. *Let $\mathbf{x_q} \sim \mathcal{D}_x = \mathcal{N}(\mu_y, \Sigma)$ and $y \in \{0,1\}$. Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with entries drawn i.i.d from the univariate Gaussian $\mathcal{N}(0, 1/d)$. Then the estimated misclassification error $\hat{Pr}_{R,\mathbf{x}_q}[\hat{h}(R\mathbf{x}_q) \neq y]$ is bounded above by:*

$$\exp\left(-\frac{k}{2}\log\left(1 + \frac{1}{4d}\cdot\|\hat{\mu}_1 - \hat{\mu}_0\|^2\cdot\frac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma\hat{\Sigma}^{-1})}\right)\right) \tag{2.1}$$

with μ_y the mean of the class from which $\mathbf{x}_q$ was drawn with covariance matrix Σ, estimated class means $\hat{\mu}_0$ and $\hat{\mu}_1$ with model covariance $\hat{\Sigma}$, and $\lambda_{\min}(\cdot)$, $\lambda_{\max}(\cdot)$ respectively the least and greatest eigenvalues of their argument.

The structure of the proof is as follows. We commence by bounding the error probability of FLD in the data space. Although this has been studied long before, the exact expression of error would make our subsequent derivation of a deterministic bound for the analysis of average behaviour (w.r.t. all random choices of R) analytically intractable. The proof of our main result, the above theorem, then follows.

Lemma 2.2. *(Bound on two-class FLD in the data space) Let $\mathbf{x_q} \sim \mathcal{D}_x = \mathcal{N}(\mu_y, \Sigma)$ with equal probability $\forall y$. 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 there is sufficient training data that $\kappa_y = \frac{(\hat{\mu}_{\neg y}+\hat{\mu}_y-2\mu_y)^T\hat{\Sigma}^{-1}(\hat{\mu}_{\neg y}-\hat{\mu}_y)}{2(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_{\neg y}-\hat{\mu}_y)}$ is positive[1], where $y, \neg y \in \mathcal{C} = \{0,1\}, y \neq \neg y$. Then the probability that $\mathbf{x}_q$ is misclassified is given by $Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y] \leqslant$*

$$\frac{1}{2}\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\right]^2}{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}\right)\dots$$

$$\dots+\frac{1}{2}\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_0+\hat{\mu}_1-2\mu_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_0-\hat{\mu}_1)\right]^2}{(\hat{\mu}_0-\hat{\mu}_1)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_0-\hat{\mu}_1)}\right)$$

with μ_y the mean of the class from which $\mathbf{x}_q$ was drawn, estimated class means $\hat{\mu}_0$ and $\hat{\mu}_1$, model covariance $\hat{\Sigma}$.

We omit the proof of the data space bound, which uses standard Chernoff-bounding techniques and can be found in the appendix to our technical report [10].

Let us assume that we have sufficient training examples, we will decompose the bound 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 $\mathbf{x_q} \sim \mathcal{D}_x = \mathcal{N}(\mu_y, \Sigma)$ with equal probability. 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:*

$$\hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma) = \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\right]^2}{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}\right) \tag{2.2}$$

and $B_{y\in\mathcal{C}}(\hat{\mu}_0, \hat{\mu}_1, \mu_0, \mu_1, \hat{\Sigma}, \Sigma)$ for the right hand side of lemma 2.2. Then, $B_{y\in\mathcal{C}}(\hat{\mu}_0, \hat{\mu}_1, \mu_0, \mu_1, \hat{\Sigma}, \Sigma)$

$$\leqslant \hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma) + \max_{y,i}\sup\left\{\left|\frac{\partial B_{y\in\mathcal{C}}}{\partial\mu_{yi}}\right|\right\}\cdot\sum_{y,i}|\hat{\mu}_{yi} - \mu_{yi}| \tag{2.3}$$

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, model covariance $\hat{\Sigma}$, and uniform class priors.

[1] 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.

Proof. (Sketch) We will use the mean value theorem, so we start by differentiating $B_{y\in\mathcal{C}}$ with respect to μ_0 to find $\nabla_{\mu_0} B_{y\in\mathcal{C}} = \frac{1}{2}\exp\left(-\frac{1}{8}\cdot\frac{\left[(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\right]^2}{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}\right)\cdot\frac{1}{2}\kappa_0\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_0)$. Since the exponential term is bounded between zero and one, the supremum of the i-th component of this gradient exists provided that $|\hat{\mu}_1+\hat{\mu}_0-2\mu_0| < \infty$ and $|\hat{\mu}_1-\hat{\mu}_0| < \infty$. So we have[2] that $B_{y\in\mathcal{C}} \leqslant \frac{1}{2}\hat{B}(\hat{\mu}_0,\hat{\mu}_1,\hat{\Sigma},\Sigma) + \max_i \sup\left\{\left|\frac{\partial B_{y\in\mathcal{C}}}{\partial\mu_{0i}}\right|\right\}\sum_i|\hat{\mu}_{0i}-\mu_{0i}| + \frac{1}{2}\mathrm{Pr}_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q)\neq 1|y=1]$. Now applying the mean value theorem again w.r.t. μ_1 decomposes the latter term similarly, then taking the maximum over both classes yields the desired result. We call the obtained two terms in (2.3) the 'estimated error' and 'estimation error' respectively. The estimation error can be bounded using Chernoff bounding techniques, and converges to zero with increasing number of training examples. □

We now have the framework in place to bound 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.

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 $\Sigma = \mathrm{E}_x\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\right]$ is the covariance matrix in the data space, then its projected counterpart Σ_R is $R\Sigma R^T$, and likewise $\hat{\Sigma}_R = R\hat{\Sigma}R^T$.

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

$$\exp\left(-\frac{1}{8}\cdot\frac{\left[(\hat{\mu}_1^R-\hat{\mu}_0^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_1^R-\hat{\mu}_0^R)\right]^2}{(\hat{\mu}_1^R-\hat{\mu}_0^R)^T\hat{\Sigma}_R^{-1}\Sigma_R\hat{\Sigma}_R^{-1}(\hat{\mu}_1^R-\hat{\mu}_0^R)}\right) \tag{2.4}$$

We would like to analyse the expectation of this in terms of the quantities of the original space. We proceed by majorization of the numerator by the Rayleigh quotient (lemma 3.1), where we take $\mathbf{v} = \left(\hat{\Sigma}_R\right)^{-1/2}(\hat{\mu}_1^R-\hat{\mu}_0^R)$ and take our positive definite Q to be $Q = \hat{\Sigma}_R^{-1/2}\Sigma_R\hat{\Sigma}_R^{-1/2}$ 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 $\hat{\Sigma}_R^{-1/2} = \left(\hat{\Sigma}_R^{-1}\right)^{1/2} = \left(\hat{\Sigma}_R^{1/2}\right)^{-1} = \left(\hat{\Sigma}_R^{-1/2}\right)^T$ ([4], Theorem 7.2.6, pg. 406). Then, we have (2.4) is less than or equal to:

$$\exp\left(-\frac{1}{8}\cdot\frac{\left[(\hat{\mu}_1^R-\hat{\mu}_0^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_1^R-\hat{\mu}_0^R)\right]^2}{\lambda_{\max}(Q)(\hat{\mu}_1^R-\hat{\mu}_0^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_1^R-\hat{\mu}_0^R)}\right) \tag{2.5}$$

Simplifying, and using the identity eigval(AB) = eigval(BA) ([6], pg. 29), we may write $\lambda_{\max}(Q) = \lambda_{\max}(\hat{\Sigma}_R^{-1/2}\Sigma_R\hat{\Sigma}_R^{-1/2}) = \lambda_{\max}(\hat{\Sigma}_R^{-1}\Sigma_R)$ and we may now bound equation (2.4) from above with:

$$\exp\left(-\frac{1}{8}\cdot\frac{(\hat{\mu}_1^R-\hat{\mu}_0^R)^T\hat{\Sigma}_R^{-1}(\hat{\mu}_1^R-\hat{\mu}_0^R)}{\lambda_{\max}(\Sigma_R\hat{\Sigma}_R^{-1})}\right) \tag{2.6}$$

$$\leqslant \exp\left(-\frac{1}{8}\cdot\frac{\|\hat{\mu}_1^R-\hat{\mu}_0^R\|^2}{\lambda_{\max}(\hat{\Sigma})}\frac{1}{\lambda_{\max}(\Sigma_R\hat{\Sigma}_R^{-1})}\right) \tag{2.7}$$

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 $\lambda_{\max}(\Sigma_R\hat{\Sigma}_R^{-1})$. We see this encodes a measure of how well the form of the model covariance matches the true covariance, and the bound is tightest when the match is closest. This is not just a function of the training set size, but rather of the (diagonal, or spherical) constraints that it is often convenient to impose on the model covariance. Although a more in-depth analysis of the effect of this in the projection space would be interesting, the following bound will be tight when $\hat{\Sigma}$ (or Σ) is spherical or when $\hat{\Sigma} = \Sigma$:

$$\exp\left(-\frac{1}{8}\cdot\frac{\|R(\hat{\mu}_1-\hat{\mu}_0)\|^2}{\lambda_{\max}(\hat{\Sigma})}\frac{1}{\lambda_{\max}(\Sigma\hat{\Sigma}^{-1})}\right) \tag{2.8}$$

$$= \exp\left(-\frac{1}{8}\cdot\|R(\hat{\mu}_1-\hat{\mu}_0)\|^2\cdot\frac{\lambda_{\min}(\hat{\Sigma}^{-1})}{\lambda_{\max}(\Sigma\hat{\Sigma}^{-1})}\right) \tag{2.9}$$

The change of term in the denominator uses the fact that $\lambda_{\max}(\Sigma_R\hat{\Sigma}_R^{-1}) \leqslant \lambda_{\max}(\Sigma\hat{\Sigma}^{-1})$, which follows from the second step of lemma 4.9 in [10].

This bound holds deterministically, for any fixed projection matrix R. We can also see from (2.9) 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 implies 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.

Observing that the expected value of the squared Euclidean norm $\|R(\hat{\mu}_1-\hat{\mu}_0)\|^2$ is $\frac{k}{d}\cdot\|(\hat{\mu}_1-\hat{\mu}_0)\|^2$ this implies, via Jensen's inequality, that the expectation of (2.9) w.r.t. R remains just above that of a similar exponential form. We can compute this expectation using the moment generating function of independent χ^2

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

variables, which yields the expression given in Theorem 2.1.[3] □

3 Discussion and future work

This paper presents initial findings of our ongoing work concerning the effects of dimensionality reduction on classifier performance. Due to space constraints we have not been able to demonstrate how to extend this result to multiclass LDA. Lemma 4.2 of [10] gives the details.

Interesting open problems include analysing the behaviour of the estimation error and finding the probability of getting a 'bad' random projection, namely one that projects the sample mean from the correct to the wrong side of the decision boundary. In particular the latter potentially has implications for classification when the sample sizes are imbalanced and the cost of misclassification is not symmetric over the two classes (e.g. in medical diagnosis where we might prefer false positives to false negatives).

References

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

[2] 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.

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

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

[5] 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.

[6] K.B.Petersen and M.S.Pedersen. *The Matrix Cookbook*. Technical University of Denmark, November 2008.

[7] 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.

[8] O.-A. Maillard and R. Munos. Compressed Least-Squares Regression. In *NIPS*, 2009.

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

[10] Robert J. Durrant and Ata Kabán. Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data. Technical Report CSR-10-03, School of Computer Science, University of Birmingham, 2010.

Appendix

Lemma 3.1 (Rayleigh quotient ([4], Theorem 4.2.2 Pg 176))**.** *If $\mathbf{Q}$ is a real symmetric matrix then its eigenvalues λ satisfy:*

$$\lambda_{\min}(\mathbf{Q}) \leqslant \frac{\mathbf{v}^T\mathbf{Q}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \leqslant \lambda_{\max}(\mathbf{Q}) \tag{3.1}$$

Lemma 3.2 (Poincaré Separation Theorem ([4], 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, 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:*

$$\lambda_i(\mathbf{S}) \leqslant \lambda_i(\mathbf{T}) \leqslant \lambda_{i+n-k}(\mathbf{S}), \quad i \in \{1, \ldots, k\} \tag{3.2}$$

and in particular:

$$\lambda_{\min}(\mathbf{S}) \leqslant \lambda_{\min}(\mathbf{T}) \textit{ and } \lambda_{\max}(\mathbf{T}) \leqslant \lambda_{\max}(\mathbf{S}) \tag{3.3}$$

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:*

$$\mathbf{u}^T\left[R\mathbf{Q}R^T\right]^{-1}\mathbf{u} \geqslant \lambda_{\min}(\mathbf{Q}^{-1})\mathbf{u}^T\mathbf{u} = \mathbf{u}^T\mathbf{u}/\lambda_{\max}(\mathbf{Q}) \tag{3.4}$$

Proof (Sketch): Use the eigenvalue identity $\lambda_{\min}(\mathbf{Q}^{-1}) = 1/\lambda_{\max}(\mathbf{Q})$ combined with lemma 3.1 and lemma 3.2.

[3] We note that the bound given here is numerically tighter than the one proved using more sophisticated tools in our technical report. However, at the cost of some tightness, the bound given in [10] reflects the behaviour of the estimated error more faithfully.