Error bounds for Kernel Fisher Linear Discriminant in Gaussian Hilbert space

Robert J. Durrant **Ata Kabán**
School of Computer Science, University of Birmingham, Edgbaston, UK, B15 2TT

Abstract

We give a non-trivial, non-asymptotic upper bound on the classification error of the popular Kernel Fisher Linear Discriminant classifier under the assumption that the kernel-induced space is a Gaussian Hilbert space.

1 Introduction

The performance of Fisher Linear Discriminant (FLD) in finite-dimensional data space has been well-studied in the asymptotic regime; for example, in the two class setting under the assumption of Gaussian classes the exact error is given in [1, 2]. However, Kernel FLD (KFLD) [3] presents specific technical challenges to deriving generalization bounds that are not present in the data space setting. 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 [3] have approached 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 error of the eigenvector estimates. Unfortunately this leaves open the question of the generalization error of the KFLD classifier. In [4] a generalization error bound is developed for a sparse version of KFLD, however with the weight vector and bias term taken to be fixed across different training sets. In a different vein, the theoretical analysis of [5], which draws on the work of [6], focuses on justifying an interesting empirical observation, namely that data mapped in the feature space tend to have a Gaussian distribution. KFLD is a well-performing and popular classifier, yet very little is known about its generalization guarantees.

Here 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. 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 the 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 [7], 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 paper is as follows: We briefly describe the classification problem, the KFLD classifier, and 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.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

1.1 The classification problem

In classification we observe N examples of labelled training data $\mathcal{T}_N = \{(z_i, y_i)\}_{i=1}^N$ where $z_i \overset{i.i.d}{\sim} \mathcal{D}_z$ and each z_i has an associated label $y_i \in \{0, 1\}$. 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 $\mathcal{L}(\hat{f}) = \arg\min_{f \in \mathcal{F}} \mathrm{E}_{z_q}[\mathcal{L}(f)]$, where $z_q \sim \mathcal{D}_z$ 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 $\mathcal{F}$ 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, via a kernel mapping ϕ:

$$\phi : \mathcal{D}_z \longrightarrow \mathcal{H} \tag{1.1}$$

$$z \longmapsto \phi(z) \tag{1.2}$$

1.2 Notation

To keep our notation compact, we will write x_i for $\phi(z_i)$ from now on and, since $\mathcal{H}$ is separable, without loss of generality we will work in ℓ_2. The feature space $\mathcal{H}_N$ is 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 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 $\mathrm{Tr}(\Sigma) < \infty$). These properties are a technical requirement if $\mathcal{D}$ is to be a non-degenerate probability distribution over $\mathcal{H}$. We indicate estimated quantities by adding a hat: $\hat{\mu}_y$, $\hat{\Sigma}$.

For convenience we assume that the x_i are linearly independent since otherwise very similar arguments to those we present still go through when $\dim(\langle x_i \rangle_{i=1}^N) < N$. Consequently [8] 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, i.e 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.

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 an extension to multi-class is relatively straightforward (e.g. [2],[9]).

The set of training observations for KFLD as treated here is therefore: $\mathcal{T}_N = \{(x_i, y_i) : x_i \sim \mathcal{D}_x, y \in \{0, 1\}\}_{i=1}^N$, and we bound the probability that a previously unseen 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 $\mathcal{D}_x \equiv \sum_{y \in \{0,1\}} \pi_y \mathcal{N}(\mu_y, \Sigma)$ in a separable Hilbert space, $\mathcal{H}$, (here taken to be ℓ_2 equipped with Gaussian probability measure over Borel sets) where π_y is 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 paper that in $\mathcal{T}_N$ we have N_0 and N_1 both greater than 0, which is the case of practical interest for the classification task.

1.3 Kernel Fisher Linear Discriminant

KFLD, first proposed by [3], is a generalization to feature space of the popular FLD classifier.

The canonical FLD is a generative classifier that seeks to model, given a set of training observations $\mathcal{T}_N$, the optimal decision boundary between classes. In the two-class setting, if $\Sigma = \Sigma_0 = \Sigma_1$ and μ_0 and μ_1 are known, the optimal classifier is given by Bayes' rule [1, 10]:

$$\begin{aligned} h(x_q) &= \mathbf{1}\left\{\log \frac{f_1(x_q)}{f_0(x_q)} > 0\right\} \\ &= \mathbf{1}\left\{(\mu_1 - \mu_0)^T \Sigma^{-1} \left(x_q - \frac{\mu_0 + \mu_1}{2}\right) > 0\right\} \end{aligned}$$

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 function μ_y and trace class covariance operator Σ. When the training observations used to construct the classifier are points in a feature space, then the resulting classifier is KFLD [3].

2 Results

We assume functional data [11], 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, in [5] it is 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 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 is [12]:

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

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:

$$\begin{aligned}&\mathrm{E}_{x_q}\left[\mathbf{1}\left\{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\left(x_q-\frac{(\hat{\mu}_0+\hat{\mu}_1)}{2}\right)_N>0\right\}\right] \quad (2.1)\\ &=\Pr_{x_q}\left\{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\left(x_q-\frac{(\hat{\mu}_0+\hat{\mu}_1)}{2}\right)_N>0\right\}\end{aligned}$$

where the estimates are obtained from a training set of size N 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.

2.1 Dimension-free bound on (sub)-exponential random variables

Lemma 1 *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:*

$$\begin{aligned}&Pr\left\{\|X\|^2\geqslant(1+\epsilon)\left(Tr(\Sigma)+\|\mu\|^2\right)\right\}\\ &\leqslant\exp\left(-\frac{Tr(\Sigma)+\|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon}-1\right)^2\right) \quad (2.2)\end{aligned}$$

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

$$\begin{aligned}&Pr\left\{\|X\|^2\leqslant(1-\epsilon)\left(Tr(\Sigma)+\|\mu\|^2\right)\right\}\\ &\leqslant\exp\left(-\frac{Tr(\Sigma)+\|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon}-1\right)^2\right) \quad (2.3)\end{aligned}$$

Lemma 1 is an extension to Hilbert space of classical finite dimensional results e.g. [13, 14]. The proof of Lemma 1 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 1 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 [4, 15, 16]. In particular, our bounds have $\sqrt{\mathrm{Tr}(\Sigma)}$ in this role, which can be considerably smaller than the data diameter, and do not require the boundedness assumption.

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 2 *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.(2.1) is given by:*

$$\begin{aligned}&\pi_0\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_0)_N}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}(\hat{\mu}_1-\hat{\mu}_0)_N}}\right)+\\ &\pi_1\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_0-\hat{\mu}_1)_N^T\hat{\Sigma}_N^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_1)_N}{\sqrt{(\hat{\mu}_0-\hat{\mu}_1)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}(\hat{\mu}_0-\hat{\mu}_1)_N}}\right)\end{aligned} \quad (2.4)$$

Where Φ is the c.d.f of the standard Gaussian distribution.

The proof of lemma 2 is much the same as that given in [1]; for completeness we give it in the Appendix.

In what follows, we bound the deviation of the quantities appearing in (2.4) from their expectations with high probability with respect to the training set, $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$.

2.3 Main Result: Tail bound on generalization error of KFLD

We will now estimate the various quantities in (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 (2.4) as follows:

$$\begin{aligned} \beta = \quad & (\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N \\ \ldots \quad & +2 (\hat{\mu}_0 - \mu_0)_N^T \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N \end{aligned} \tag{2.5}$$

Using the decomposition (2.5) we can rewrite the first term of lemma 2 in the following form:

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

Where:

$$A = \frac{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} \Sigma_N \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}} \tag{2.6}$$

is the term responsible for the *estimated error*, and:

$$B = \frac{2 (\mu_0 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} \Sigma_N \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}} \tag{2.7}$$

is the term responsible for the *estimation error*. We will lower bound A and upper bound B to bound the whole term from above.

2.3.1 Lower-bounding the term A

We will make use of the Kantorovich inequality:

Lemma 3 Kantorovich Inequality. ([17], Theorem 7.4.41 Pg 444) *Let Q be a symmetric positive definite matrix $Q \in \mathcal{M}_N$ with eigenvalues $0 < \lambda_{\min}(Q) \leqslant \ldots \leqslant \lambda_{\max}(Q)$. Then, for all $\mathbf{v} \in \mathcal{H}_N$:*

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

With equality holding for some unit vector $\mathbf{v}$.

Applying Kantorovich inequality with the choice of positive definite $Q = \hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}}$ we can lower bound A with:

$$\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \Sigma_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N} \cdot \frac{2\sqrt{\lambda_{\min}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}}) \lambda_{\max}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}})}}{\lambda_{\min}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}}) + \lambda_{\max}(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma_N \hat{\Sigma}_N^{-\frac{1}{2}})} \tag{2.8}$$

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 (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:

$$\| \Sigma_N^{-\frac{1}{2}} (\hat{\mu}_1 - \hat{\mu}_0)_N \| \frac{2\sqrt{\kappa(\Sigma_N^{-\frac{1}{2}} \hat{\Sigma}_N \Sigma_N^{-\frac{1}{2}})}}{1 + \kappa(\Sigma_N^{-\frac{1}{2}} \hat{\Sigma}_N \Sigma_N^{-\frac{1}{2}})} \tag{2.9}$$

Now, applying Rayleigh quotient to the norm above we see:

$$\begin{aligned} \| \Sigma_N^{-\frac{1}{2}} (\hat{\mu}_1 - \hat{\mu}_0)_N \| \quad &\geqslant \quad \frac{\| (\hat{\mu}_1 - \hat{\mu}_0)_N \|}{\sqrt{\lambda_{\max}(\Sigma_N)}} = \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|}{\sqrt{\lambda_{\max}(\Sigma_N)}} \\ &\geqslant \quad \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|}{\sqrt{\lambda_{\max}(\Sigma)}} \end{aligned} \tag{2.10}$$

where the equality in the chain (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 $\lambda_{\max}(\Sigma_N) = \lambda_{\max}(P\Sigma P^T) = \lambda_{\max}(P^T P \Sigma) \leqslant \lambda_{\max}(P^T P)\lambda_{\max}(\Sigma) = 1 \cdot \lambda_{\max}(\Sigma)$ with this last equality holding 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 $(\hat{\mu}_1 - \hat{\mu}_0) \sim \mathcal{N}(\mu_1 - \mu_0, \Sigma/N_1 + \Sigma/N_0) = \mathcal{N}(\mu_1 - \mu_0, (N_0 + N_1)\Sigma/N_0 N_1) = \mathcal{N}(\mu_1 - \mu_0, N\Sigma/N_0 N_1)$. Applying lemma 1 (2.3) to $\|\hat{\mu}_1 - \hat{\mu}_0\|$ we lower bound this as:

$$\|\hat{\mu}_1 - \hat{\mu}_0\| \geqslant \sqrt{(1-\epsilon)\left(\frac{N}{N_0 N_1} \mathrm{Tr}(\Sigma) + \|\mu_1 - \mu_0\|^2\right)} \tag{2.11}$$

with probability at least:

$$1-\exp\left(-\frac{\mathrm{Tr}(\Sigma)+\|\mu_1-\mu_0\|^2\frac{N_0N_1}{N}}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon}-1\right)^2\right) \tag{2.12}$$

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 2.3.2), and the resulting upper bound on the condition number of this matrix we denote by $\bar{\kappa}(\epsilon)$ (see eq. (2.19)).

Now, replacing, the term A is lower bounded w.h.p by:

$$A \geqslant 2g(\bar{\kappa}(\epsilon))\sqrt{(1-\epsilon)\left(\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)}+\frac{N}{N_0N_1}\frac{\mathrm{Tr}(\Sigma)}{\lambda_{\max}(\Sigma)}\right)} \tag{2.13}$$

The first summand under the square root in (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 $\mathcal{H}$ — the larger this distance 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.

2.3.2 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.(2.9).

The most common form of regularizing the covariance estimate in the feature space is:

$$X\hat{\Sigma}_{UR}X^T + \alpha C \tag{2.14}$$

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 [18]), normalized by the total number of training points , i.e.:

$$\hat{\Sigma}_{UR} = \frac{1}{N}\sum_{y=0}^{1}\sum_{i=1}^{N_y}(x_i-\hat{\mu}_y)(x_i-\hat{\mu}_y)^T \tag{2.15}$$

The regularization term may be chosen as $C = I_N$, or $C = XX^T$. The former is more common, the latter is proposed in [19] 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:

$$\lambda_{\max}\left(\frac{1}{N}\sum_{y=0}^{1}(P\Sigma P^T)^{-\frac{1}{2}}\sum_{i=1}^{N_y}P(x_i-\hat{\mu}_y)(x_i-\hat{\mu}_y)^TP^T(P\Sigma P^T)^{-\frac{1}{2}} + \alpha(P\Sigma P^T)^{-1}\right)$$

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 [20] 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 4 Singular values of Gaussian matrices. ([21], 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:*

$$Pr\{\sqrt{N} - \sqrt{n} - \epsilon \leqslant s_{\min}(A) \leqslant s_{\max}(A) \leqslant \sqrt{N} + \sqrt{n} + \epsilon\} \geqslant 1 - 2e^{-\epsilon^2/2}, \quad \forall \epsilon > 0$$

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 4 to bound the eigenvalues of the terms of the unregularized covariance estimate. Hence we have:

$$\lambda_{\max}\left(\Sigma_N^{-\frac{1}{2}} \hat{\Sigma}_N \Sigma_N^{-\frac{1}{2}}\right) \tag{2.16}$$
$$\leqslant \left(1 + \sqrt{\tfrac{N-2}{N}} + \tfrac{\epsilon}{\sqrt{N}}\right)^2 + \alpha/\lambda_{\min}(\Sigma_N)$$

with probability at least $1 - e^{-\epsilon^2/2}$.

The smallest eigenvalue is governed by the regularization term, and may be lower bounded as:

$$\lambda_{\min}(\Sigma_N^{-\frac{1}{2}} \hat{\Sigma}_N \Sigma_N^{-\frac{1}{2}}) = \lambda_{\min}(\Sigma_N^{-\frac{1}{2}} \hat{\Sigma}_{UR} \Sigma_N^{-\frac{1}{2}} + \alpha(\Sigma_N^{-1}))$$
$$\geqslant \lambda_{\min}(\alpha(\Sigma_N^{-1})) = \alpha/\lambda_{\max}(\Sigma_N) \tag{2.17}$$
$$\geqslant \alpha/\lambda_{\max}(\Sigma) \tag{2.18}$$

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:

$$\kappa\left(\Sigma_N^{-\frac{1}{2}} \hat{\Sigma}_N \Sigma_N^{-\frac{1}{2}}\right)$$
$$\leqslant \tfrac{\lambda_{\max}(\Sigma)}{\alpha}\left(1 + \sqrt{\tfrac{N-2}{N}} + \tfrac{\epsilon}{\sqrt{N}}\right)^2 + \kappa(\Sigma_N)$$
$$=: \bar{\kappa}(\epsilon) \tag{2.19}$$

The first term in eq. (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. (2.16) and eq. (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 in [19] that regularization with the kernel matrix can reduce the generalization error of KFLD.

2.3.3 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 (2.7) as:

$$\begin{aligned} B &= \frac{2\,(\mu_0 - \hat{\mu}_0)_N^T\, \Sigma_N^{-\frac{1}{2}} \Sigma_N^{\frac{1}{2}} \hat{\Sigma}_N^{-1}\, (\hat{\mu}_1 - \hat{\mu}_0)_N}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)_N^T \hat{\Sigma}_N^{-1} \Sigma_N \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N}} \\ &\leqslant \frac{2\|\Sigma_N^{-\frac{1}{2}}(\mu_0 - \hat{\mu}_0)_N\| \|\Sigma_N^{\frac{1}{2}} \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N\|}{\|\Sigma_N^{\frac{1}{2}} \hat{\Sigma}_N^{-1} (\hat{\mu}_1 - \hat{\mu}_0)_N\|} \\ &= 2\|\Sigma_N^{-\frac{1}{2}}(\mu_0 - \hat{\mu}_0)_N\| \end{aligned} \tag{2.20}$$

using Cauchy-Schwarz in the numerator.
Then, using lemma 1 we further upper bound (2.20) with:

$$2\sqrt{(1+\epsilon) \cdot \mathrm{Tr}\,(I_N/N_0)} = 2\sqrt{(1+\epsilon)N/N_0} \tag{2.21}$$

with probability $\geqslant 1 - \exp\left(-\frac{1}{2}N \cdot (\sqrt{1+\epsilon} - 1)^2\right)$.

2.3.4 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:

$$\Phi\left(-2\left[g(\bar{\kappa}(\epsilon_2))\sqrt{(1-\epsilon_1)\left(\tfrac{\|\mu_1 - \mu_0\|^2}{\lambda_{\max}(\Sigma)} + \tfrac{N}{N_0 N_1}\tfrac{\mathrm{Tr}(\Sigma)}{\lambda_{\max}(\Sigma)}\right)} - \sqrt{\bar{\kappa}(\epsilon_2)}\sqrt{(1+\epsilon_3)N/N_0}\right]\right)$$

where $\bar{\kappa}(\epsilon_2)$ is given by eq. (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,

$$\delta_0 \leqslant \exp\left(-\frac{1}{2}N \cdot (\sqrt{1+\epsilon_3} - 1)^2\right) + \exp\left(-\epsilon_2^2/2\right)$$
$$\ldots + \exp\left(-\frac{\mathrm{Tr}(\Sigma) + \|\mu_1 - \mu_0\|^2 \frac{N_0 N_1}{N}}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon_1} - 1\right)^2\right)$$

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. $1-\delta(\Sigma,\epsilon_1,\epsilon_2,\epsilon_3,N_0,N_1) = 1-(\delta_0+\exp(-\frac{1}{2}N\cdot(\sqrt{1+\epsilon_3}-1)^2))$.

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 1 *For any $\delta\in(0,1)$, the generalization error of KFLD in a Gaussian Hilbert space, eq.(2.4), is upper-bounded w.p. at least $1-\delta$ over the random choice of training set $\mathcal{T}_{N=N_0+N_1}$, by:*

$$\sum_{y=0}^{1}\pi_y\Phi\left(-2\Bigg[\ g(\bar{\kappa}(\epsilon_2))\times\ldots \left[\sqrt{\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)}+\frac{N_0+N_1}{N_0N_1}\frac{Tr(\Sigma)}{\lambda_{\max}(\Sigma)}}-\sqrt{\frac{2(N_0+N_1)}{N_0N_1}\log\frac{4}{\delta}}\right]_+ \ldots-\sqrt{\frac{N}{N_y}}\left(1+\sqrt{\frac{2}{N}\log\frac{4}{\delta}}\right)\Bigg]\right)$$

where now $\bar{\kappa}(\epsilon_2)$ is given by replacing $\epsilon_2:=\sqrt{2\log\frac{4}{\delta}}$ in eq. (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.

We see that a key term in the bound is the scaled distance between the mean functions in the Hilbert space. Using the fact [7] that with a suitable kernel choice (universal kernels) 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 the 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 1 (to theorem 1) *Let the data be N-dimensional, and having Gaussian class-conditionals $x|y\sim\mathcal{N}(\mu_y,\Sigma)$. 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:*

$$\sum_{y=0}^{1}\pi_y\Phi\left(-2\Bigg[\ g(\bar{\kappa}(\epsilon_2))\times\ldots \left[\sqrt{\frac{\|\mu_1-\mu_0\|^2}{\lambda_{\max}(\Sigma)}+\frac{M}{M_0M_1}\frac{Tr(\Sigma)}{\lambda_{\max}(\Sigma)}}-\sqrt{\frac{2M}{M_0M_1}\log\frac{5}{\delta}}\right]_+ \ldots-\sqrt{\frac{N}{M_y}}\left(1+\sqrt{\frac{2}{N}\log\frac{5}{\delta}}\right)\Bigg]\right) \tag{2.22}$$

where $\bar{\kappa}(\epsilon_2)$ is as in Theorem 1 when the MLE of Σ, $\hat{\Sigma}_{UR}$ is singular, and when $\hat{\Sigma}_{UR}$ is non-singular we can bound its minimum eigenvalue away from zero using Lemma 4, which yields the following tighter $\bar{\kappa}(\epsilon_2)$:

$$\kappa\left(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}\right)\leqslant\left(\frac{\sqrt{M-2}+\sqrt{N}+\epsilon}{\sqrt{M-2}-\sqrt{N}-\epsilon}\right)^2=:\bar{\kappa}(\epsilon)$$

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 (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 $\mathrm{Tr}(\Sigma)/\lambda_{\max}(\Sigma)$ 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 $\mathrm{Tr}(\Sigma)/\lambda_{\max}(\Sigma)$ 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 (2.23) converges to 1, and all the terms other than (an upper bound on) the Bayes error vanish in eq.(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.

3 Conclusions

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 or to the much wider family of sub-Gaussian class-conditional distributions appears relatively straightforward using the results and tools in [21]. 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.

Appendix

Proof of Lemma 1

We prove the statement of eq. (2.2) fully, and outline the proof of (2.3) which is very similar. Let $t > 0$ be a positive real constant (to be optimized later), then:

$$\begin{aligned}&\Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\}\\ &= \Pr\left\{\exp\left(t\|X\|^2\right) \geqslant \exp\left(t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\right\}\\ &\leqslant \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\mathrm{E}\left[\exp\left(t\|X\|^2\right)\right] \quad (3.1)\end{aligned}$$

Where (3.1) 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. [22] proposition 1.2.8) for all $t \in (0, 1/2\lambda_{\max}(\Sigma))$ we have (3.1) is equal to:

$$\begin{aligned}&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\\ &\quad \cdots \prod_{i=1}^{\infty}\left(1-2t\lambda_i\right)^{-\frac{1}{2}}\exp\left(\frac{t\mu_i^2}{1-2t\lambda_i}\right)\\ &= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\\ &\quad \cdots \prod_{i=1}^{\infty}\left(1+\frac{2t\lambda_i}{1-2t\lambda_i}\right)^{\frac{1}{2}}\exp\left(\frac{t\mu_i^2}{1-2t\lambda_i}\right)\\ &\leqslant \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\\ &\quad \cdots \prod_{i=1}^{\infty}\exp\left(\frac{1}{2}\frac{2t\lambda_i}{1-2t\lambda_{\max}(\Sigma)}\right)\exp\left(\frac{t\mu_i^2}{1-2t\lambda_{\max}(\Sigma)}\right)\\ &= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)+\frac{t\left(\sum_{i=1}^{\infty}\lambda_i+\mu_i^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right)\\ &= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)+\frac{t\left(\mathrm{Tr}(\Sigma)+\|\mu\|^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right) \quad (3.2)\end{aligned}$$

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 (3.2) yields, after some algebra, (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 (2.3) is proved similarly. We begin by noting:

$$\begin{aligned}&\Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\}\\ =\ &\Pr\left\{\exp\left(-t\|X\|^2\right) \geqslant \exp\left(-t\left(1-\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\right\}\\ \leqslant\ &\exp\left(t(1-\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right.\\ -\ &\left.t\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)/1+2t\lambda_{\max}(\Sigma)\right)\end{aligned}$$

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.

Proof of Lemma 2

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:

$$\Pr_{x_q}\left\{\left(\hat{\mu}_1-\hat{\mu}_0\right)_N^T\hat{\Sigma}_N^{-1}\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N > 0\middle| y=0\right\} \quad (3.3)$$

Define $a_N^T := \left(\hat{\mu}_1-\hat{\mu}_0\right)_N^T\hat{\Sigma}_N^{-1}$ 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:

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

which is a univariate Gaussian. Therefore:

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

Hence, for the query point x_q we have the probability (3.3) is given by:

$$\begin{aligned}&\Pr_{x_q}\left\{\left(\hat{\mu}_1-\hat{\mu}_0\right)_N^T\hat{\Sigma}_N^{-1}\left(x_q-\tfrac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N > 0\middle| y=0\right\}\\ &= \Phi\left(\frac{a_N^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)_N}{\sqrt{a_N\Sigma_N a_N^T}}\right)\\ &= \Phi\left(-\tfrac{1}{2}\frac{\left(\hat{\mu}_1-\hat{\mu}_0\right)_N^T\hat{\Sigma}_N^{-1}\left(\hat{\mu}_0+\hat{\mu}_1-2\mu_0\right)_N}{\sqrt{\left(\hat{\mu}_1-\hat{\mu}_0\right)_N^T\hat{\Sigma}_N^{-1}\Sigma_N\hat{\Sigma}_N^{-1}\left(\hat{\mu}_1-\hat{\mu}_0\right)_N}}\right)\end{aligned}$$

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

[1] 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.

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

[3] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and KR Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48. IEEE, 2002.

[4] T. Diethe, Z. Hussain, D. Hardoon, and J. Shawe-Taylor. Matching pursuit kernel fisher discriminant analysis. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 121–128, 2009.

[5] S.Y. Huang, C.R. Hwang, and M.H. Lin. Kernel Fisher's discriminant analysis in Gaussian reproducing kernel Hilbert space. Technical report, Institute of Statistical Science, Academia Sinica, 2005.

[6] P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984.

[7] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In M. Hutter, R. Servedio, and E. Takimoto, editors, *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer Berlin / Heidelberg, 2007.

[8] B. Schölkopf and A.J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. The MIT Press, 2002.

[9] 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.

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

[11] J.O. Ramsay and B.W. Silverman. *Functional data analysis*. Springer, 2nd edition, 2005.

[12] R. Herbrich. *Learning kernel classifiers: theory and algorithms*. The MIT Press, 2002.

[13] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[14] S. Dasgupta. *Learning Probability Distributions*. PhD thesis, Berkeley, 2000.

[15] J. Shawe-Taylor and N. Cristianini. Estimating the moments of a random vector with applications. In *Proceedings of GRETSI 2003 conference*, volume 1, pages 47–52, 2003.

[16] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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

[18] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2nd edition, 2000.

[19] T.P. Centeno and N.D. Lawrence. Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. *The Journal of Machine Learning Research*, 7:455–491, 2006.

[20] M.S. Srivastava. Singular Wishart and multivariate beta distributions. *Annals of Statistics*, pages 1537–1560, 2003.

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

[22] S. Maniglia and A. Rhandi. Gaussian measures on separable Hilbert spaces and applications. *Quaderni del Dipartimento di Matematica dell'Università del Salento*, 2004(1), 2004.