Random Projections as Regularizers – Supplementary Material

Robert J. Durrant BOBD@WAIKATO.AC.NZ
Department of Statistics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.

Ata Kabán A.KABAN@CS.BHAM.AC.UK
School of Computer Science, University of Birmingham, Edgbaston, B15 2TT, UK.

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

There are three cases to consider:

1.1. Case $k < \rho - 1$

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

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

where $\lambda_{\min\neq 0}$ denotes the smallest nonzero eigenvalue of $\hat{\Lambda}$ (and of $\hat{\Sigma}$), and $\lambda_{\max}$ its largest eigenvalue.

Recall that as a result of lemmas 3 and 4 we only need consider the diagonal entries of this expectation as the off-diagonal terms are known to be zero.

Now, we evaluate the remaining expectation. There are two cases: If $i > \rho$ then r_i is independent from the denominator and we have $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] = \mathrm{E}\left[r_i^2\right]\mathrm{E}\left[1/\sum_{j=1}^{\rho} r_j^2\right] = \frac{1}{\rho-2}$, where we used the expectation of the inverse-χ^2 with ρ degrees of freedom, and the fact that $\mathrm{E}\left[r_i^2\right] = 1$. When $i \leqslant \rho$, then in turn we have $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] = \mathrm{E}\left[\frac{r_i^2}{\|r\|^2}\right] = \frac{1}{\rho}$. That is,

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

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

bounded by:

$$\kappa \leqslant \frac{\rho}{\rho - 2} \cdot \frac{\lambda_{\max}}{\lambda_{\min \neq 0}} \tag{1}$$

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

$$\left[\begin{array}{c|c} \mathrm{E}[P^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} P] & 0 \\ \hline 0 & \mathrm{E}[S^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} S] \end{array}\right] \tag{2}$$

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

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

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

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

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

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

$$\mathrm{E}[P^T(PP^T)^{-1}P]/\lambda_{\min \neq 0} \succcurlyeq \mathrm{E}\left[P^T(P\underline{\hat{\Lambda}}P^T)^{-1}P\right] \tag{3}$$

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

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

$$\mathrm{E}[P^T(PP^T)^{-1}P] = \mathrm{E}[P^T P]/\rho = \frac{k}{\rho} \cdot I_\rho \tag{4}$$

Similarly to eq. (3) we can also show that:

$$\mathrm{E}\left[P^T(P\underline{\hat{\Lambda}}P^T)^{-1}P\right] \succcurlyeq \mathrm{E}[P^T \left(PP^T\right)^{-1} P]/\lambda_{\max} \tag{5}$$

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

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

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

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

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

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

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

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

where we used the expectation of the inverse Wishart. Likewise,

$$\mathrm{Tr}\left(\mathrm{E}\left[\left(P\underline{\hat{\Lambda}}P^T\right)^{-1}\right]\right) \geqslant \frac{k}{\rho-k-1}\cdot\frac{1}{\lambda_{\max}} \tag{6}$$

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

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

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

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

$$\kappa \leqslant \frac{\rho}{\rho-k-1}\cdot\frac{\lambda_{\max}}{\lambda_{\min\neq 0}} \tag{7}$$

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

1.2. Case $k > \rho + 1$

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

We use the form in eq. (2) again, with P a $k \times \rho$ matrix and S a $k \times (d-\rho)$ matrix. Since here we have $k > \rho + 1$ we replace $\left(P\hat{\underline{\Lambda}}P^T\right)^{-1}$ by its pseudo-inverse. Then we can rewrite this (Penrose, 1955) as:

$$\left(P\hat{\underline{\Lambda}}P^T\right)^{+} = P\hat{\underline{\Lambda}}^{\frac{1}{2}}(\hat{\underline{\Lambda}}^{\frac{1}{2}}P^TP\hat{\underline{\Lambda}}^{\frac{1}{2}})^{-2}\hat{\underline{\Lambda}}^{\frac{1}{2}}P^T \tag{8}$$

$$= P(P^TP)^{-1}\hat{\underline{\Lambda}}^{-1}(P^TP)^{-1}P^T \tag{9}$$

Using this, the first diagonal block becomes:

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

$$= \hat{\underline{\Lambda}}^{-1} \tag{10}$$

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

$$= \mathrm{Tr}\left(\mathrm{E}\left[\left(P\hat{\underline{\Lambda}}P^T\right)^{+}\right]\right)\cdot I_{d-\rho}$$

$$= \mathrm{E}\left[\mathrm{Tr}\left(\left(P\hat{\underline{\Lambda}}P^T\right)^{+}\right)\right]\cdot I_{d-\rho}$$

$$= \mathrm{E}\left[\mathrm{Tr}\left(P^TP\left(P^TP\right)^{-1}\hat{\underline{\Lambda}}^{-1}\left(P^TP\right)^{-1}\right)\right]\cdot I_{d-\rho}$$

$$= \frac{\mathrm{Tr}(\hat{\underline{\Lambda}}^{-1})}{k-\rho-1}\cdot I_{d-\rho} \tag{11}$$

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

Hence, in this case we obtained the exact form:

$$\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{+}R\right] = \left[\begin{array}{c|c} \hat{\underline{\Lambda}}^{-1} & 0 \\ \hline 0 & \frac{\mathrm{Tr}(\hat{\underline{\Lambda}}^{-1})}{k-\rho-1}\cdot I_{d-\rho} \end{array}\right] \tag{12}$$

It follows that implicitly when $k > \rho+1$ the data space covariance estimate gets regularized only in its null space by the ensemble, with zero eigenvalues replaced by $\frac{k-\rho-1}{\mathrm{Tr}(\hat{\underline{\Lambda}}^{-1})}$. We see that, unlike the previous case, here the amount of regularization increases as we *increase* k.

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

For a single RP-FLD, the choice $k = \rho$ or $k = \rho \pm 1$ is bad because the ML covariance estimate in the projection space remains poorly conditioned. Detailed analysis in Hoyle

(2011) has shown that having the number of points or the rank of covariance equal to the dimensionality (which is k for RP-FLD) performs even worse than pseudo-inverse FLD would when having less points than dimensions (which is also bad, cf. the analysis in Bickel and Levina (2004)).

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

To make this connection, we again use the block-diagonal form of equation (2). Now, because $k = \rho$, P is a square matrix with independent $\mathcal{N}(0, 1)$ entries and therefore it is invertible with probability 1. Hence, the upper block is

$$\mathrm{E}[P^T \left(P\hat{\underline{\Lambda}}P^T\right)^{-1} P] = \mathrm{E}[P^T \left(P^T\right)^{-1} \hat{\underline{\Lambda}}^{-1} P^{-1} P] = \hat{\underline{\Lambda}}^{-1} \tag{13}$$

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

$$\mathrm{E}\left[R^T \left(R\hat{\Lambda}R^T\right)^{-1} R\right] = \left[\begin{array}{c|c} \hat{\underline{\Lambda}}^{-1} & 0 \\ \hline 0 & \mathrm{diag}(+\infty) \end{array}\right] \tag{14}$$

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

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

2. Generalization error of the ensemble for a fixed training set

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

Theorem 1 (Exact generalization error with Gaussian classes) *Let* $x_q \sim \sum_{y=0}^{1} \pi_y \mathcal{N}(\mu_y, \Sigma)$, *where* $\Sigma \in \mathcal{M}_{d\times d}$ *is a full rank covariance matrix. Let* $R \in \mathcal{M}_{k\times d}$ *be a random projection matrix with i.i.d. Gaussian entries and denote* $\hat{S}^{-1} := E_R\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$. *Then the exact generalization error of the converged randomly projected ensemble classifier is given by:*

$$\Pr_{x_q}\{\hat{h}_{ens}(x_q) \neq y\} = \sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{S}^{-1}(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}}\right) \tag{15}$$

The proof of this theorem is similar in spirit to the one for a single FLD in Pattison and Gossink (1999). For completeness we give it below.

Proof of Theorem 1

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

$$\Pr_{x_q}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1}\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\} \tag{16}$$

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

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

and so:

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

which is a univariate Gaussian. Therefore:

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

Hence, for the query point x_q we have the probability (16) is given by:

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

where Φ is the c.d.f of the standard Gaussian.
A similar argument deals with the case when x_q belongs to class 1, and applying the law of total probability completes the proof.
Indeed equation (15) has the same form as the error of the data space FLD (See Bickel and Levina, 2004; Pattison and Gossink, 1999, for example.) and the converged ensemble,

inspected in the original data space, produces exactly the same mean estimates and covariance matrix eigenvector estimates as FLD working on the original data set. However it has different eigenvalue estimates that result from the sophisticated regularization scheme that we analyzed in section 1.

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

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

We will use the following concentration lemma:

Lemma 2 (Concentration inequalities for exponential random variables) *Let $X = (X_1, X_2, X_3, \ldots, X_d)$ be a sequence of Gaussian random variables in $\mathbb{R}^d$ with mean vector $E[X] = \mu$ and covariance matrix Σ. Let $\epsilon > 0$. Then:*

$$Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(Tr\left(\Sigma\right) + \|\mu\|^2\right)\right\} \leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon} - 1\right)^2\right) \tag{17}$$

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

$$Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(Tr\left(\Sigma\right) + \|\mu\|^2\right)\right\} \leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon} - 1\right)^2\right) \tag{18}$$

Proof: We prove the statement of eq. (17) fully, and outline the proof of (18) which is very similar. Let $t > 0$ be a positive real constant (to be optimized later), then:

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right)\right\} \\
&= \Pr\left\{\exp\left(t\|X\|^2\right) \geqslant \exp\left(t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right)\right)\right\} \\
&\leqslant \exp\left(-t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right)\right) \mathrm{E}\left[\exp\left(t\|X\|^2\right)\right]
\end{aligned} \tag{19}$$

Where (19) follows by Markov's inequality. Now, $X \sim \mathcal{N}(\mu, \Sigma)$ and so $\|X\|^2 = \sum_{i=1}^{\infty} X_i^2$ has a non-central χ^2 distribution, and therefore $\mathrm{E}\left[\exp\left(t\|X\|^2\right)\right]$ is the moment generating function of a non-central χ^2 distribution. Hence (e.g. Maniglia and Rhandi, 2004, proposition 1.2.8) for all $t \in (0, 1/2\lambda_{\max}(\Sigma))$ we have (19) is equal to:

$$\begin{aligned}
&= \exp\left(-t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right)\right) \prod_{i=1}^{\infty} \left(1 - 2t\lambda_i\right)^{-\frac{1}{2}} \exp\left(\frac{t\mu_i^2}{1-2t\lambda_i}\right) \\
&= \exp\left(-t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right)\right) \prod_{i=1}^{\infty} \left(1 + \frac{2t\lambda_i}{1-2t\lambda_i}\right)^{\frac{1}{2}} \exp\left(\frac{t\mu_i^2}{1-2t\lambda_i}\right) \\
&\leqslant \exp\left(-t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right)\right) \prod_{i=1}^{\infty} \exp\left(\frac{1}{2}\frac{2t\lambda_i}{1-2t\lambda_{\max}(\Sigma)}\right) \exp\left(\frac{t\mu_i^2}{1-2t\lambda_{\max}(\Sigma)}\right) \\
&= \exp\left(-t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right) + \frac{t\left(\sum_{i=1}^{\infty}\lambda_i + \mu_i^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right) \\
&= \exp\left(-t\,(1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right) + \|\mu\|^2\right) + \frac{t\left(\mathrm{Tr}(\Sigma) + \|\mu\|^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right)
\end{aligned} \tag{20}$$

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 (20) yields, after some algebra, (17):

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

The second inequality (18) is proved similarly. We begin by noting:

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

and then complete the proof as before, substituting in the optimal $t = \frac{1+(1-\epsilon)^{-\frac{1}{2}}}{2\lambda_{\max}(\Sigma)}$ to give the bound.

Now we can bound the generalization error of the RP-FLD ensemble. We begin by decomposing the numerator of the generalization error term (for a single class) obtained in Theorem 1 as follows:

$$\begin{aligned}&\left(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0\right)^T \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right)\\ =\ &\left(\hat{\mu}_1 - \hat{\mu}_0\right)^T \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right) + 2\left(\hat{\mu}_0 - \mu_0\right)^T \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right)\end{aligned} \tag{21}$$

Using this decomposition we can rewrite the argument of the first term in Theorem 1 in the following form:

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

Where:

$$A = \frac{\left(\hat{\mu}_1 - \hat{\mu}_0\right)^T \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}} \tag{22}$$

and:

$$B = \frac{2\left(\mu_0 - \hat{\mu}_0\right)^T \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}} \tag{23}$$

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

3.1. Lower-bounding the term A

Applying the Kantorovich inequality (e.g. Horn and Johnson, 1985, Thm 7.4.41), A is lower bounded by:

$$\|\Sigma^{-\frac{1}{2}}\left(\hat{\mu}_1 - \hat{\mu}_0\right)\| \cdot \frac{2\sqrt{\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})}}{1+\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})} \tag{24}$$

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

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

Applying the concentration bound Lemma 2, (18), we have:

$$\|\Sigma^{-\frac{1}{2}}(\hat{\mu}_1 - \hat{\mu}_0)\| \geqslant \sqrt{(1-\epsilon)\left(\frac{d \cdot N}{N_0 N_1} + \|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2\right)} \tag{25}$$

with probability at least:

$$1 - \exp\left(-\frac{d + \|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 N_0 N_1/N}{2}\left(\sqrt{1-\epsilon} - 1\right)^2\right) \tag{26}$$

To complete the bounding of the term A, we denote $g(a) := \frac{\sqrt{a}}{1+a}$, and observe that this is a monotonic decreasing function on $[1, \infty)$. So, replacing a with the condition number $\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}}) \in [1, \infty)$ we need to upper bound this condition number in order to lower bound g. Denoting this upper bound by $\bar{\kappa}$, which will be quantified in Section 3.3, then the term A is lower bounded with high probability by:

$$A \geqslant 2g(\bar{\kappa})\sqrt{(1-\epsilon)\left(\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 + \frac{d \cdot N}{N_0 N_1}\right)} \tag{27}$$

3.2. Upper-bounding the term B

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

$$B \leqslant \frac{2\|\Sigma^{-\frac{1}{2}}(\mu_0 - \hat{\mu}_0)\| \cdot \|\Sigma^{\frac{1}{2}}\hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\|}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}} \tag{28}$$

After cancellation, this simplifies to:

$$= 2\|\Sigma^{-\frac{1}{2}}(\mu_0 - \hat{\mu}_0)\| \tag{29}$$

and so by Lemma 2, (17), we have:

$$B \quad \leqslant \quad 2\sqrt{(1+\epsilon)d/N_0} \tag{30}$$

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

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

Upper-bound on largest eigenvalue

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

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

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

with probability at least $1-\exp(-\epsilon^2/2), \forall \epsilon > 0$, where throughout we use the fact that the non-zero eigenvalues of AB are the same as non-zero eigenvalues of BA, in the second to last step we used the fact that for invertible matrices A we have $\lambda_{\max}(A) = 1/\lambda_{\min}(A^{-1})$, and in the last step we used that for any full row-rank matrix R, $(R\Sigma R^T)^{-\frac{1}{2}}R\hat{\Sigma}R^T(R\Sigma R^T)^{-\frac{1}{2}}$ is distributed as a k-dimensional Wishart with $N-2$ degrees of freedom and scale matrix I_k (e.g. Mardia et al., 1979, Corr. 3.4.1.2), and used the high probability lower-bound for the smallest eigenvalue of such a matrix, Eq. (2.3) in Vershynin (2011).

Lower-bound on smallest eigenvalue

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

$$
\begin{aligned}
&\lambda_{\min}(\Sigma^{\frac{1}{2}}\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)]\Sigma^{\frac{1}{2}}) \\
&= \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}(\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)])^{-1}\Sigma^{-\frac{1}{2}})} \\
&= \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}\{\hat{\Sigma}+(\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)])^{-1}-\hat{\Sigma}\}\Sigma^{-\frac{1}{2}})}
\end{aligned} \tag{31}
$$

Now, using Weyl's inequality, and the SVD decomposition $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$ combined with Lemma 3, the denominator in (31) is upper-bounded by:

$$
\begin{aligned}
&\lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{U}\{(\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1}-\hat{\Lambda}\}\hat{U}^T\Sigma^{-\frac{1}{2}}) \\
&\leqslant \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \lambda_{\max}((\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1}-\hat{\Lambda})/\lambda_{\min}(\Sigma)
\end{aligned} \tag{32}
$$

Now observe from Lemma 4 that the matrix $\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1}-\hat{\Lambda}$ is diagonal and, from our analysis in Section 1, it has the upper ρ diagonal entries in the interval:

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

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

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

. Hence,

$$
\lambda_{\max}((\mathrm{E}_R[R^T(R\hat{\Lambda}R)^{-1}R])^{-1}-\hat{\Lambda}) \leqslant \frac{\rho}{k}\lambda_{\max}(\hat{\Lambda})
$$

and so the lower-bounding of (32) continues as:

$$\geqslant \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \frac{\rho}{k}\frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\Sigma)}} \tag{33}$$

Now observe that $\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}$ is a d-dimensional standard Wishart with $N-2$ degrees of freedom and scale matrix I_d (e.g. Mardia et al., 1979, Corr. 3.4.1.2), and using the bound in Vershynin (2011) for largest eigenvalues of standard Wishart matrices we get (33) lower-bounded as

$$\geqslant \frac{1}{(\sqrt{N-2}+\sqrt{d}+\epsilon)^2/N + \frac{\rho}{k}\frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\Sigma)}} \tag{34}$$

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

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

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

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

$$\kappa = \frac{\lambda_{\max}(\Sigma^{\frac{1}{2}}\cdot \mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R]\cdot\Sigma^{\frac{1}{2}})}{\lambda_{\min}(\Sigma^{\frac{1}{2}}\cdot \mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R]\cdot\Sigma^{\frac{1}{2}})} \tag{35}$$

$$\leqslant \frac{(\sqrt{N-2}+\sqrt{d}+\epsilon)^2(1+\rho/k\cdot\kappa(\Sigma))}{(\sqrt{N-2}-\sqrt{k}-\epsilon)^2} =: \bar{\kappa}(\epsilon) \tag{36}$$

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

3.4. Putting everything together

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

Theorem 3 *Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$ be a set of training data of size $N = N_0 + N_1$, subject to $N < d$ and $N_y > 1\ \forall y$ with Gaussian class-conditionals $x|y \sim \mathcal{N}(\mu_y, \Sigma)$. Let ρ be the rank of the maximum likelihood estimate of the covariance matrix and let $k < \rho - 1$ be an integer. Then for any $\delta \in (0,1)$ and any training set of size N, the generalization error of the converged ensemble of randomly projected FLD classifiers is upper-bounded w.p. $1-\delta$ by the following:*

$$\Pr_{x_q}(\hat{h}_{ens}(x_q) \neq y) \leqslant \sum_{y=0}^{1}\pi_y\Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right)\left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2 + \frac{dN}{N_0N_1}}\ldots\right.\right.\right.$$
$$\left.\left.\left.\ldots - \sqrt{\frac{2N}{N_0N_1}\log\frac{5}{\delta}}\right]_+ - \sqrt{\frac{d}{N_y}}\left(1+\sqrt{\frac{2}{d}\log\frac{5}{\delta}}\right)\right]\right) \tag{37}$$

where $\bar{\kappa}$ is given by eq. (36) and $g(\cdot)$ is the function $g(a) := \frac{\sqrt{a}}{1+a}$.

Having completed the groundwork, the proof is now simple algebraic manipulation – the details are as follows:

There are five terms to simultaneously bound with high probability, namely the two B_y, A, and the two extreme eigenvalues involved in the condition number bound. We use the standard approach of setting each of the confidence probabilities no greater than $\delta/5$ and solving for ϵ (or a function of ϵ appearing in the bound) then back-substituting and applying the union bound to derive a guarantee which holds with probability $1-\delta$.
Firstly, for the extreme eigenvalues we have (twice):

$$\begin{aligned} \exp\left(-\epsilon_3^2/2\right) &\leqslant \delta/5 \\ \Longrightarrow \sqrt{2\log(5/\delta)} &\leqslant \epsilon_3 \end{aligned} \tag{38}$$

For the upper bounds on the B_y we have:

$$\exp\left(-\frac{d}{2}\left(\sqrt{1+\epsilon_y}-1\right)^2\right) \leqslant \delta/5$$

and solving for $\sqrt{1+\epsilon_y}$ we obtain:

$$\begin{aligned} \sqrt{\frac{2\log(5/\delta)}{d}} &\leqslant \pm\left(\sqrt{1+\epsilon_y}-1\right) \\ \Longrightarrow 1+\sqrt{\frac{2\log(5/\delta)}{d}} &\geqslant \sqrt{1+\epsilon_y} \end{aligned} \tag{39}$$

Finally, for the lower bound on A (which holds for both classes simultaneously) we solve for $\sqrt{1-\epsilon_2}$ to obtain:

$$\exp\left(-\left(\frac{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}{2N/N_0N_1}\right)\left(\sqrt{1-\epsilon_2}-1\right)^2\right) \leqslant \delta/5$$

$$\begin{aligned} \Longleftrightarrow \frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2} &\leqslant \left(\sqrt{1-\epsilon_2}-1\right)^2 \\ \Longleftrightarrow \sqrt{\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}} &\leqslant \pm\left(\sqrt{1-\epsilon_2}-1\right) \\ \Longrightarrow 1-\sqrt{\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}} &\geqslant \sqrt{1-\epsilon_2} \end{aligned} \tag{40}$$

Plugging the left hand sides of the inequalities (38), (39) and (40) into the bounds on κ, B_0, B_1 and A for ϵ_3, $\sqrt{1+\epsilon_0}$, $\sqrt{1+\epsilon_1}$ and $\sqrt{1-\epsilon_2}$ respectively gives, after some algebra, the theorem.

References

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.

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.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning; data mining, inference, and prediction.* Springer, 2001.

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

D. Hoyle. Accuracy of Pseudo-Inverse Covariance Learning – A Random Matrix Theory Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(7):1470 – 81, 2011.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

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.

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

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

R. Penrose. A generalized inverse for matrices. 51(3):406–413, 1955.

S. Raudys and R.P.W. Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19(5):385–392, 1998.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, University of Michigan, March 2011. URL http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf.