1. 1-D Fisher's Linear Discriminant (FLD)

Consider an in-distribution task and an out-of-distribution task specified by the distributions F_{in} and F_{out}, respectively. F_{in} is characterized by the class conditional densities,

$$f_{0,in} = \mathcal{N}(-\mu, \sigma^2) \tag{1}$$

$$f_{1,in} = \mathcal{N}(\mu, \sigma^2) \tag{2}$$

and F_{out} is characterized by the class conditional densities,

$$f_{0,out} = \mathcal{N}(-\mu + \Delta, \sigma^2) \tag{3}$$

$$f_{1,out} = \mathcal{N}(\mu + \Delta, \sigma^2) \tag{4}$$

Suppose that we have n samples $S_{in} = \{X_i, Y_i\}_{i=1}^n$ drawn from F_{in} and m samples $S_{out} = \{X_j, Y_j\}_{j=1}^m$ drawn from F_{out}. The samples are class-balanced. We are interested in generalizing on the in-distribution task using both S_{in} and S_{out}.

1.1 Single-Head FLD

Let M_0 and M_1 be the estimated means of classes 0 and 1 respectively. Note that each class comprises of samples from both in- and out-of-distribution tasks. Consider M_0, which is given by,

$$M_0 = \frac{\sum_{i=1}^{n/2} X_i + \sum_{j=1}^{m/2} X_j}{n/2 + m/2} \tag{5}$$

The mean and variance of M_0 are given by,

$$\mathbb{E}[M_0] = -\mu + \frac{m}{n+m}\Delta \tag{6}$$

$$\text{Var}[M_0] = \frac{2\sigma^2}{n+m} \tag{7}$$

By the central limit theorem, it can be shown that

$$M_0 \sim \mathcal{N}\Big(-\mu + \frac{m}{n+m}\Delta, \frac{2\sigma^2}{n+m}\Big) \tag{8}$$

Similarly,

$$M_1 \sim \mathcal{N}\Big(\mu + \frac{m}{n+m}\Delta, \frac{2\sigma^2}{n+m}\Big) \tag{9}$$

It can be noted that a sample X in the combined class 0 is drawn from a Gaussian mixture distribution given by,

$$X \sim f_0 = \frac{n}{n+m} f_{0,in} + \frac{m}{n+m} f_{0,out} \tag{10}$$

Therefore,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \frac{n}{n+m} f_{0,in}(x)dx + \int_{\mathbb{R}} x \frac{m}{n+m} f_{0,out}(x)dx = -\mu + \frac{m}{n+m}\Delta \tag{11}$$

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 \frac{n}{n+m} f_{0,in}(x)dx + \int_{\mathbb{R}} x^2 \frac{m}{n+m} f_{0,out}(x)dx = \frac{n}{n+m}(\sigma^2+\mu^2) + \frac{m}{n+m}(\sigma^2+(-\mu+\Delta)^2) \tag{12}$$

Hence, the variance of class 0 samples is given by,

$$\text{Var}[X] = \sigma^2 + \frac{mn}{(n+m)^2}\Delta^2 \tag{13}$$

It can be shown that class 1 samples have the same variance. Therefore, the variances Σ_0 and Σ_1 of class 0 and 1 are given by,

$$\Sigma_0 = \Sigma_1 = \Sigma = \sigma^2 + \frac{mn}{(n+m)^2}\Delta^2 \tag{14}$$

Lemma 1.1.1 The generalization risk of the in-distribution task is non-monotonic w.r.t to OOD sample size m, under certain shifts Δ.

Proof. The decision rule of the single-head FLD is given by,

$$g(x) = \begin{cases} 1, & \omega^\top x > c \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where, $\omega = (\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0)$ and $c = \omega^\top \frac{1}{2}(M_0 + M_1)$. In the single-head FLD, both in-distribution and OOD samples are used to estimate the projection vector ω and threshold c.

Consider the expression $\omega^\top x > c$.

$$\omega^\top x > c \tag{16}$$

$$(\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0) > (\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0)\frac{1}{2}(M_0 + M_1) \tag{17}$$

$$\frac{M_1 - M_0}{2(\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2)}x > \frac{M_1 - M_0}{2(\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2)}\frac{1}{2}(M_0 + M_1) \tag{18}$$

$$x > \frac{1}{2}(M_0 + M_1) \tag{19}$$

Therefore, Eq. (9) reduces to the following decision rule.

$$g(x) = \begin{cases} 1, & x > h \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where, $h = \frac{1}{2}(M_0 + M_1)$.

Now, consider a test input X from the in-distribution task, i.e. $X \sim F_{in}$. The generalization risk $L(h)$ is then given by,

$$\begin{aligned}
L(h) &= P[Y \neq g(X)|X = x] && (21)\\
&= P[Y = 1, g(X) = 0|X = x] + P[Y = 0, g(X) = 1|X = x] && (22)\\
&= P[Y = 1, X < h|X = x] + P[Y = 0, X > h|X = x] && (23)\\
&= P[X < h|Y = 1, X = x]P[Y = 1|X = x] + P[X > h|Y = 0, X = x]P[Y = 0|X = x] && (24)\\
&= \frac{1}{2}P[X < h|Y = 1, X = x] + \frac{1}{2}P[X > h|Y = 0, X = x] && (25)\\
&= \frac{1}{2}(P_{X\sim f_1}[X < h] + P_{X\sim f_0}[X > h]) && (26)\\
&= \frac{1}{2}(P_{X\sim f_{1,in}}[X < h] + 1 - P_{X\sim f_{0,in}}[X < h]) && (27)\\
&= \frac{1}{2}\left[1 - \Phi\left(\frac{h+\mu}{\sigma}\right) + \Phi\left(\frac{h-\mu}{\sigma}\right)\right] && (28)
\end{aligned}$$

Therefore,

$$L(h) = \frac{1}{2}\left[1 - \Phi\left(\frac{h+\mu}{\sigma}\right) + \Phi\left(\frac{h-\mu}{\sigma}\right)\right] \tag{29}$$

where, $h = \frac{1}{2}(M_0 + M_1) \sim \phi = \mathcal{N}(\frac{m}{n+m}\Delta, \frac{\sigma^2}{n+m})$.

The expected generalization risk is given by,

$$L_{n,m,\Delta} = \mathbb{E}[L(h)] \tag{30}$$

$$L_{n,m,\Delta} = \int_{-\infty}^{\infty} L(h)\phi(h)dh \tag{31}$$

Fig. 1.1 illustrates the non-monotonic nature of the expected generalization error w.r.t OOD sample size.

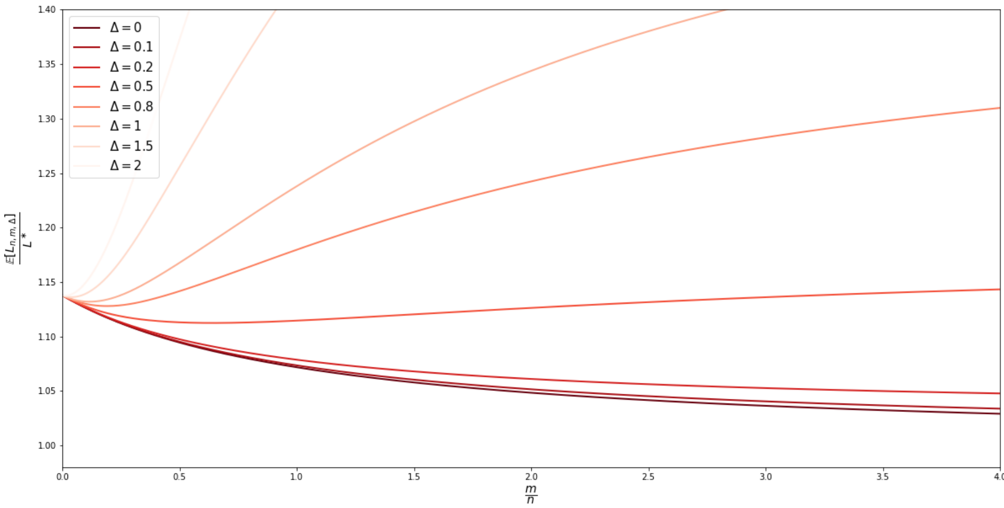

Figure 1.1.1 OOD sample size vs. expected generalization risk for single-head FLD under various shifts Δ. ($n = 5, \mu = 1, \sigma = 1$)

1.2 Multi-Head FLD

In the multi-head setting, the projection vector ω is estimated using both the in-distribution and OOD samples. However, now there would be 2 thresholds c_{in} and c_{out} reflecting the two task-specific heads of the FLD. The projection vector estimated using both the in-distribution and OOD samples is given by,

$$\omega = (\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0) = \frac{M_1 - M_0}{2(\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2)} \tag{32}$$

Next, consider c_{in}, the threshold specific to the in-distribution task.

$$c_{in} = \omega^\top \frac{1}{2}(M_{0,in} + M_{1,in}) \tag{33}$$

where, $M_{0,in} = \frac{1}{n/2}\sum_{i=1}^{n/2} X_i$ and $M_{1,in} = \frac{1}{n/2}\sum_{i=n/2+1}^{n} X_i$. By central limit theorem, $M_{0,in} \sim \mathcal{N}(-\mu, 2\sigma^2/n)$ and $M_{1,in} \sim \mathcal{N}(\mu, 2\sigma^2/n)$.

Lemma 1.2.1 The generalization risk of the in-distribution task is monotonic w.r.t to OOD sample size m.

Proof. The decision rule of the multi-head FLD specific to in-distribution task is given by,

$$g_{in}(x) = \begin{cases} 1, & \omega^\top x > c_{in} \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

where, $\omega = (\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0)$ and $c_{in} = \omega^\top \frac{1}{2}(M_{0,in} + M_{1,in})$. In the multi-head FLD, both in-distribution and OOD samples are used to estimate the projection vector ω and the threshold c_{in} is estimated only using the projected in-distribution data.

Consider the expression $\omega^\top x > c_{in}$.

$$\omega^\top x > c_{in} \tag{35}$$

$$(\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0) > (\Sigma_0 + \Sigma_1)^{-1}(M_1 - M_0)\frac{1}{2}(M_{0,in} + M_{1,in}) \tag{36}$$

$$\frac{M_1 - M_0}{2(\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2)} x > \frac{M_1 - M_0}{2(\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2)} \frac{1}{2}(M_{0,in} + M_{1,in}) \tag{37}$$

$$x > \frac{1}{2}(M_{0,in} + M_{1,in}) \tag{38}$$

Therefore, the decision rule reduces to,

$$g_{in}(x) = \begin{cases} 1, & x > h_{in} \\ 0, & \text{otherwise} \end{cases} \tag{39}$$

where, $h_{in} = \frac{1}{2}(M_{0,in} + M_{1,in})$.

As in the proof of Lemma 1.1.1, it can be shown that the generalization error $L(h)$ of the in-distribution task is given by,

$$L(h) = \frac{1}{2}\left[1 - \Phi\left(\frac{h_{in} + \mu}{\sigma}\right) + \Phi\left(\frac{h_{in} - \mu}{\sigma}\right)\right] \tag{40}$$

where, $h_{in} = \frac{1}{2}(M_{0,in} + M_{1,in}) \sim \phi = \mathcal{N}(0, \sigma^2/n)$. A special cases arises when $\Delta = 0$ where

$h_{in} = \frac{1}{2}(M_{0,in} + M_{1,in}) \sim \phi = \mathcal{N}(0, \sigma^2/(n+m))$

The expected generalization risk is given by,

$$L_{n,m,\Delta} = \mathbb{E}[L(h)] \tag{41}$$

$$L_{n,m,\Delta} = \int_{-\infty}^{\infty} L(h)\phi(h)dh \tag{42}$$

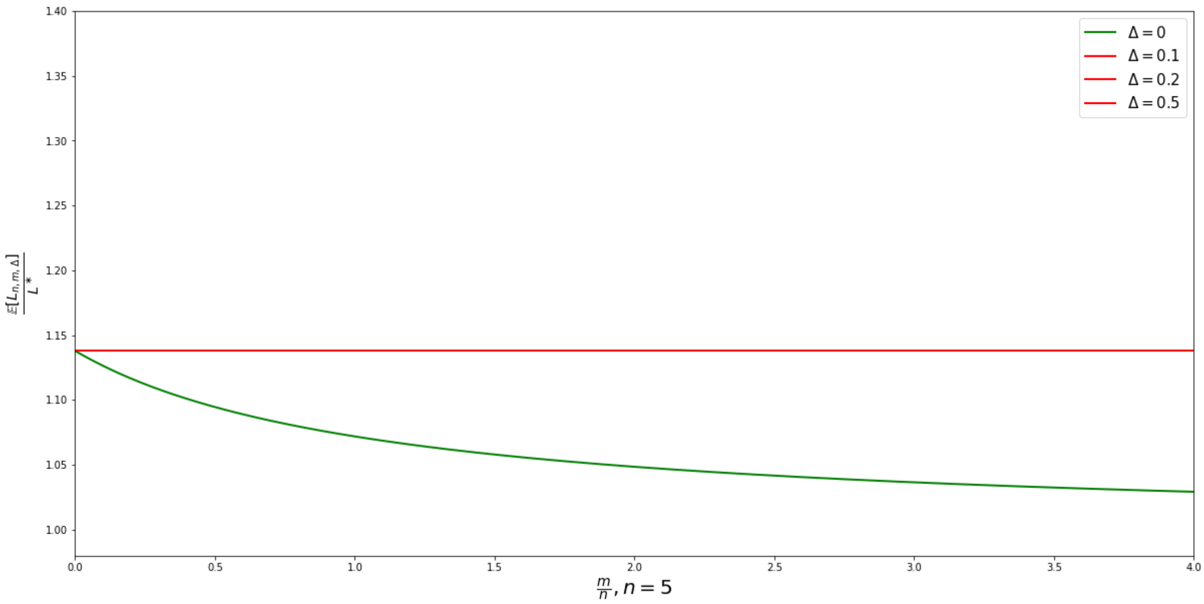

Figure 1.2.1 OOD sample size vs. expected generalization risk for multi-head FLD under various shifts Δ. ($n = 5, \mu = 1, \sigma = 1$)

1-D Multi-Head LDA with Unequal Class Priors

Let's begin with a general two-class problem X has a density $(1-p)f_0(x) + pf_1(x)$, where f_0 and f_1 are both mulitivariate normal distributions with parameters m_i, Σ_i for $i = 0, 1$. Then the Bayes rule is described by,

$$g_(x) = \begin{cases} 1, & pf_1(x) > (1-p)f_0(x) \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

Taking the logarithms, we observe that $g(x) = 1$ if and only if,

$$(x-m_1)^\top \Sigma_1^{-1}(x-m_1) - 2\log p + \log|\Sigma_1| < (x-m_0)^\top \Sigma_0^{-1}(x-m_0) - 2\log(1-p) + \log|\Sigma_1| \tag{44}$$

When $\Sigma_1 = \Sigma_0 = \Sigma$, this expression reduces to,

$$(x-m_1)^\top \Sigma^{-1}(x-m_1) - 2\log p < (x-m_0)^\top \Sigma^{-1}(x-m_0) - 2\log(1-p) \tag{45}$$

$$2(x^\top \Sigma^{-1} m_1 - x^\top \Sigma^{-1} m_0) > m_1^\top \Sigma^{-1} x - m_0^\top \Sigma^{-1} x + 2\log\frac{1-p}{p} \tag{46}$$

$$\left(2\Sigma^{-1}(m_1 - m_0)\right)^\top x > \left(2\Sigma^{-1}(m_1 - m_0)\right)^\top \left(\frac{m_0 + m_1}{2}\right) + 2\log\frac{1-p}{p} \tag{47}$$

$$\omega^\top x > c + 2\log\frac{1-p}{p} \tag{48}$$

where,

$$\omega = 2\Sigma^{-1}(m_1 - m_0) \tag{49}$$

$$c = \omega^\top \left(\frac{m_0 + m_1}{2}\right) + 2\log\frac{1-p}{p} \tag{50}$$

In 1-D setting, these expressions reduce to,

$$\omega = \frac{2}{\sigma^2}(m_1 - m_0) \tag{51}$$

$$c = \omega\left(\frac{m_0 + m_1}{2}\right) + 2\log\frac{1-p}{p} \tag{52}$$

In the multi-head setting,

$$\omega = \frac{2}{\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2}(\hat{m}_1 - \hat{m}_0) \tag{53}$$

$$c_{in} = \omega\left(\frac{\hat{m}_{0,in} + \hat{m}_{1,in}}{2}\right) + 2\log\frac{1-\pi_{in}}{\pi_{in}} \tag{54}$$

Therefore, the decision condition is given by,

$$x > \frac{\hat{m}_{0,in} + \hat{m}_{1,in}}{2} + \frac{\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2}{\hat{m}_1 - \hat{m}_0} \log \frac{1 - \pi_{in}}{\pi_{in}} \tag{55}$$

$$x > h + \frac{A}{g} \tag{56}$$

where,

$$h_{in} = \frac{\hat{m}_{0,in} + \hat{m}_{1,in}}{2} \tag{57}$$

$$g = \hat{m}_1 - \hat{m}_0 \tag{58}$$

$$A = \left(\sigma^2 + \frac{mn}{(n+m)^2}\Delta^2\right) \log \frac{1 - \pi_{in}}{\pi_{in}} \tag{59}$$

Hence the generalization error on the in-distribution task is given by,

$$L(h, g) = \frac{1}{2}\left[1 - \Phi\left(\frac{h_{in} + A/g + \mu}{\sigma}\right) + \Phi\left(\frac{h_{in} + A/g - \mu}{\sigma}\right)\right] \tag{60}$$

where, $h_{in} \sim \mathcal{N}(0, \sigma^2/n)$ and $g \sim \mathcal{N}(2\mu, 4\sigma^2/(n+m))$.

Proving Lemma 1.1?

Consider the risk of the 1-D single-head LDA.

$$\mathcal{E} = \frac{1}{2} - \frac{1}{2}\Phi\left[\frac{\bar{\mu} + \mu}{\sqrt{1 + \bar{\sigma}^2}}\right] + \frac{1}{2}\Phi\left[\frac{\bar{\mu} - \mu}{\sqrt{1 + \bar{\sigma}^2}}\right]$$

where $\bar{\mu} = \frac{(1-\alpha)m\Delta}{\alpha n + (1-\alpha)m}$ and $\bar{\sigma}^2 = \frac{((1-\alpha)^2 m + \alpha^2 n)\sigma^2}{(\alpha n + (1-\alpha)m)^2}$.

$$\frac{d\mathcal{E}}{dm} = -\frac{1}{2}\phi\left[\frac{\bar{\mu} + \mu}{\sqrt{1 + \bar{\sigma}^2}}\right]\frac{d}{dm}\left(\frac{\bar{\mu} + \mu}{\sqrt{1 + \bar{\sigma}^2}}\right) + \frac{1}{2}\phi\left[\frac{\bar{\mu} - \mu}{\sqrt{1 + \bar{\sigma}^2}}\right]\frac{d}{dm}\left(\frac{\bar{\mu} - \mu}{\sqrt{1 + \bar{\sigma}^2}}\right) \tag{61}$$

$$= \frac{1}{2}\phi\left[\frac{\bar{\mu} + \mu}{\sqrt{1 + \bar{\sigma}^2}}\right]\left[-\frac{d}{dm}\left(\frac{\bar{\mu} + \mu}{\sqrt{1 + \bar{\sigma}^2}}\right) + \exp\left(\frac{2\mu\bar{\mu}}{\sqrt{1 + \bar{\sigma}^2}}\right)\frac{d}{dm}\left(\frac{\bar{\mu} - \mu}{\sqrt{1 + \bar{\sigma}^2}}\right)\right] \tag{62}$$

$$= -\frac{1}{2}\phi\left[\frac{\bar{\mu} + \mu}{\sqrt{1 + \bar{\sigma}^2}}\right]\left[\frac{d}{dm}\left(\frac{\bar{\mu}}{\sqrt{1 + \bar{\sigma}^2}}\right)\left[1 - \exp\left(\frac{2\mu\bar{\mu}}{\sqrt{1 + \bar{\sigma}^2}}\right)\right] + \frac{d}{dm}\left(\frac{\mu}{\sqrt{1 + \bar{\sigma}^2}}\right)\right] \tag{63}$$

Consider,

$$A = \frac{d}{dm}\left(\frac{\bar{\mu}}{\sqrt{1+\bar{\sigma}^2}}\right)\left[1 - \exp\left(\frac{2\mu\bar{\mu}}{\sqrt{1+\bar{\sigma}^2}}\right)\right] + \frac{d}{dm}\left(\frac{\mu}{\sqrt{1+\bar{\sigma}^2}}\right)$$