**The Australian National University**
**School of Computing**
Group member name(s): Campbell Rankine
Group member UID(s): u6997593

Semester 1, 2022
Assignment 1

# COMP4670: Statistical Machine Learning

## 1 Section 1: Bayesian Estimate vs Probabalistic hunches

**Answer to Question 1.1: Your hunch**

The class belongs to Program B

**Answer to Question 1.2: Computing the probabilities**

**Find:**

$$P(ProgramA \mid Observed) \tag{1}$$
$$P(ProgramB \mid Observed) \tag{2}$$

**Equation 1:**

$$P(ProgramA \mid Observed) = \frac{P(Observed \mid ProgramA)P(ProgramA)}{P(Observed)}$$

Since both have an equal number of classes P(Program A) must be 0.5.

$$P(ProgramA \mid Observed) = \frac{P(Observed \mid ProgramA)0.5}{P(Observed)}$$

We already know the fraction of observed boys is 0.55.

$$P(ProgramA \mid Observed) = \frac{P(Observed \mid ProgramA)0.5}{0.55}$$

The following $P(Observed \mid ProgramA)$ can be represented as a binary distribution by the function:

$$P(Observed \mid ProgramA) = \binom{n}{x}(0.65)^x(0.35)^{n-x}$$

Since X is the number of successes we set x = 0.55n and yield:

$$P(Observed \mid ProgramA) = \binom{n}{0.55n}(0.65)^{0.55n}(0.35)^{0.45n}$$

**Equation 2: By similar calculation and reasoning.**

Since P(Program B) Will be the same, as they have the same class numbers, we yield:

$$P(ProgramB \mid Observed) = \frac{P(Observed \mid ProgramB)0.5}{0.55}$$

By setting the number of successes x=0.55 as above. We yield the equation for $P(Observed \mid ProgramB)$ as:

$$P(Observed \mid ProgramB) = \binom{n}{0.55n}(0.45)^{0.55n}(0.55)^{0.45n}$$

**Since we have now found both equations Compare ratio between the two:**

The 0.5/0.55 in each cancels out, as does the $\binom{n}{0.55n}$. Therefore we are left with:

$$\frac{P(ProgramA \mid Observed)}{P(ProgramB \mid Observed)} = \frac{(0.65)^{0.55n}(0.35)^{0.45n}}{(0.45)^{0.55n}(0.55)^{0.45n}}$$

by using the power of quotient property, obtain:

$$\frac{P(ProgramA \mid Observed)}{P(ProgramB \mid Observed)} = (\frac{0.65}{0.45})^{0.55n}(\frac{0.35}{0.55})^{0.45n}$$

$$\frac{P(ProgramA \mid Observed)}{P(ProgramB \mid Observed)} = (1.44)^{0.55n}(0.63)^{0.45n}$$

$$\frac{P(ProgramA \mid Observed)}{P(ProgramB \mid Observed)} = (1.224)^{n}(0.81)^{n}$$

$$\frac{P(ProgramA \mid Observed)}{P(ProgramB \mid Observed)} = (0.9914)^{n}$$

Since we assume we will not have a class number $n < 1$ the ratio between the probabilities is $< 1$ for all values n. Meaning that program B is more likely regardless of the value of n.

**Answer to Question 1.3: Well-reasoned hunch**

Since for a binary distribution as p-¿0.5 and n tends to infinity the distribution will become closer to symmetrical and closer to normal. Since 0.45 is closer to 0.5 than 0.65 the variance in B is larger meaning that it has a greater chance of being the observed class.

# 2 Section 2: Exponential Families

### Answer to Question 2.1: Verifying valid distribution

Valid probability distribution function if non negative and integrates to one.

**Non-Negative:**

As we are dealing with exponential there are no values for each formulae such that this function will return a negative value

**Integrates to 1:**

$$\int q(x \mid \boldsymbol{\eta})dx = \int Exp(\boldsymbol{\eta}^T u(x) - \psi(\boldsymbol{\eta}))dx$$

$$\int q(x \mid \boldsymbol{\eta})dx = \int Exp(\boldsymbol{\eta}^T u(x))Exp(-\psi(\boldsymbol{\eta}))dx$$

By definition $Exp(-\psi(\boldsymbol{\eta}))$ is equal to $g(\boldsymbol{\eta})$ and since we are integrating with respect to x we can treat $g(\boldsymbol{\eta})$ as a constant. This allows us to pull it out of the integral:

$$\int q(x \mid \boldsymbol{\eta})dx = g(\boldsymbol{\eta}) \int Exp(\boldsymbol{\eta}^T u(x))$$

Sub back in for $g(\boldsymbol{\eta})$, and use definition to substitute for $\psi(\boldsymbol{\eta})$:

$$\int q(x \mid \boldsymbol{\eta})dx = Exp(-\ln \int Exp(\boldsymbol{\eta}^T u(x))) \int Exp(\boldsymbol{\eta}^T u(x))$$

We know from previous math courses that $Exp(-\ln something) = \frac{1}{something}$ therefore:

$$\int q(x \mid \boldsymbol{\eta})dx = \frac{1}{\int Exp(\boldsymbol{\eta}^T u(x))} \int Exp(\boldsymbol{\eta}^T u(x))$$

These cancel out and leave us with a final result of 1, which means this is a valid probability distribution

### Answer to Question 2.2: A Bayesian example (Part 1)

Given $q(x \mid \mu) = \mathcal{N}(\mu, \sigma^2)$ and we know that we are following a gaussian distribution. We can obtain the following:

$$q(x \mid \mu) = h(x)g(\eta)Exp(\boldsymbol{\eta}^T u(x))$$

$$\boldsymbol{\eta} = (\mu/\sigma^2, \frac{-1}{2\sigma^2})^T$$

To express $q(\mu \mid x)$ we employ bayes theorum which leads us to:

$$q(\mu \mid x) = \frac{h(x)g(\eta)Exp(\boldsymbol{\eta}^T u(x))p(\mu)}{p(x)}$$

We know p(x) becomes the regularization term as an integral of the top, and $p(\mu)$ is given. This leads us to:

$$q(\mu \mid x) = \frac{h(x)g(\eta)Exp(\boldsymbol{\eta}^T u(x))Exp(\boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))}{\int h(x)g(\eta)Exp(\boldsymbol{\eta}^T u(x))Exp(\boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))dx}$$

we know h(x) and $g(\boldsymbol{\eta})$ are constants as they are given in the definition of gaussian distribution. So these can be pulled out of the integral on the bottom, leading us to:

$$q(\mu \mid x) = \frac{h(x)g(\eta)Exp(\boldsymbol{\eta}^T u(x))Exp(\boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))}{h(x)g(\eta)\int Exp(\boldsymbol{\eta}^T u(x))Exp(\boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))dx}$$

We can easily see h(x) and $g(\boldsymbol{\eta})$ cancel:

$$q(\mu \mid x) = \frac{Exp(\boldsymbol{\eta}^T u(x))Exp(\boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))}{\int Exp(\boldsymbol{\eta}^T u(x))Exp(\boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))dx}$$

We can simplify the top and the bottom by using the rules of exponents to yield us:

$$q(\mu \mid x) = \frac{Exp(\boldsymbol{\eta}^T u(x) + \boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))}{\int Exp(\boldsymbol{\eta}^T u(x) + \boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))dx}$$

Expand gaussian definition:

$$q(\mu \mid x) = \frac{Exp(\frac{-1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) + \boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))}{\int Exp(\frac{-1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) + \boldsymbol{\eta}^t u(\mu) - \psi(\boldsymbol{\eta}))dx}$$

We can see this is easily a dot product such that:

$$\hat{u}(x) = (1, -1, -1, \mu, \mu^2)$$

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{\eta} u(x), \psi(\boldsymbol{\eta}), \frac{x}{\sigma^2}, \frac{x^2}{\sigma^2}, \frac{-1}{2\sigma^2})$$

yielding us:

$$q(\mu \mid x) = \frac{Exp(\hat{\boldsymbol{\eta}}^T \hat{u}(x))}{\int Exp(\hat{\boldsymbol{\eta}}^T \hat{u}(x))dx}$$

where our bottom part matches the definition for the log partition function so

$$q(\mu \mid x) = Exp(\hat{\boldsymbol{\eta}}^T \hat{u}(x) - \psi(\hat{\boldsymbol{\eta}}))$$

**Answer to Question 2.3: A Bayesian example (Part 2)**

We now know some value $\mu_0$ and $\sigma_0^2$ and we are given the similar definition as 2.2:

$$q(x \mid \mu) = \mathcal{N}(\mu, \sigma^2)$$

$$q(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$$

To find $Exp(u, \boldsymbol{\eta})$ start by expanding definition for $\mathcal{N}(\mu_0, \sigma_0^2)$:

$$Exp(u, \boldsymbol{\eta}) = (\frac{1}{2\pi\sigma_0^2})^{1/2} Exp(\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2)$$

$$Exp(u, \boldsymbol{\eta}) = (\frac{1}{2\pi\sigma_0^2})^{1/2} Exp(\frac{-1}{2\sigma_0^2}\mu^2 + \frac{\mu_0}{\sigma_0^2}\mu - \frac{-1}{2\sigma_0^2}\mu_0^2)$$

$$Exp(u,\boldsymbol{\eta}) = (\frac{1}{2\pi})^{1/2}(\frac{1}{\sigma_0^2})^{1/2}Exp(\frac{-1}{2\sigma_0^2}\mu^2 + \frac{\mu_0}{\sigma_0^2}\mu - \frac{-1}{2\sigma_0^2}\mu_0^2)$$

Set h(x) = $(\frac{1}{2\pi})^{1/2}$:

$$Exp(u,\boldsymbol{\eta}) = h(x)(\frac{1}{\sigma_0^2})^{1/2}Exp(\frac{-1}{2\sigma_0^2}\mu^2 + \frac{\mu_0}{\sigma_0^2}\mu - \frac{-1}{2\sigma_0^2}\mu_0^2)$$

We can factor mu and the mu$_0 or sigma_0 terms inside the exponents the dot product is visible$ :

$$Exp(u,\boldsymbol{\eta}) = h(x)(\frac{1}{\sigma_0^2})^{1/2}Exp(\begin{pmatrix} \frac{\mu_0}{\sigma_0^2} \\ \frac{-1}{2\sigma_0^2} \end{pmatrix}^T \begin{pmatrix} \mu \\ \mu^2 \end{pmatrix} - \frac{-1}{2\sigma_0^2}\mu_0^2)$$

We now remove the last part of the exponent to put the equation in exponential family form ($\boldsymbol{\eta}^t u(x)$):

$$Exp(u,\boldsymbol{\eta}) = h(x)(\frac{1}{\sigma_0^2})^{1/2}Exp(-\frac{\mu_0^2}{2\sigma_0^2})Exp(\begin{pmatrix} \frac{\mu_0}{\sigma_0^2} \\ \frac{-1}{2\sigma_0^2} \end{pmatrix}^T \begin{pmatrix} x \\ x^2 \end{pmatrix})$$

We now can clearly see our $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \begin{pmatrix} \frac{\mu_0}{\sigma_0^2} \\ \frac{-1}{2\sigma_0^2} \end{pmatrix}$$

We can also see that this matches our definition of u(mu) that was given to us, confirming this is a gaussian distribution:

$$u(x) = \begin{pmatrix} \mu \\ \mu^2 \end{pmatrix}$$

$$Exp(u,\boldsymbol{\eta}) = h(x)(\frac{1}{\sigma_0^2})^{1/2}Exp(-\frac{\mu_0^2}{2\sigma_0^2})Exp(\boldsymbol{\eta}^T u(x))$$

We should now put everything into terms of the $\boldsymbol{\eta}$ to make sure that this is valid that means every term containing $\mu_0$ or $\sigma_0^2$ must be a linear combination of elements in $\boldsymbol{\eta}$:

$$(\frac{1}{\sigma_0^2})^{1/2} = (-2\boldsymbol{\eta}_2)^{1/2}$$

$$Exp(u,\boldsymbol{\eta}) = h(x)(-2\boldsymbol{\eta}_2)^{1/2}Exp(-\frac{\mu_0^2}{2\sigma_0^2})Exp(\boldsymbol{\eta}^T u(x))$$

$$-\frac{\mu_0^2}{2\sigma_0^2} = (\boldsymbol{\eta}_1^2/b\boldsymbol{\eta}_2)$$

Where B is some real value

$$-\frac{\mu_0^2}{2\sigma_0^2} = (\frac{\mu_0^2}{\sigma_0^4}/\frac{-b}{2\sigma_0^2})$$

b=4 so:

$$-\frac{\mu_0^2}{2\sigma_0^2} = (\boldsymbol{\eta}_1^2/4\boldsymbol{\eta}_2)$$

$$Exp(u, \boldsymbol{\eta}) = h(x)(-2\boldsymbol{\eta}_2)^{1/2} Exp(-\frac{\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2}) Exp(\boldsymbol{\eta}^T u(x))$$

So therefor the parameters of $Exp(u, \boldsymbol{\eta})$ are:

$$\boldsymbol{\eta} = \begin{pmatrix} \frac{\mu_0}{\sigma_0^2} \\ \frac{-1}{2\sigma_0^2} \end{pmatrix}$$

Which matches our Gaussian distribution equations. This allows us to simply plug in $\boldsymbol{\eta}$ into the equation from question 2.2 yielding the parameters for $\mu \mid x$ as:

$$\hat{\boldsymbol{\eta}} = (\begin{pmatrix} \frac{\mu_0}{\sigma_0^2} \\ \frac{-1}{2\sigma_0^2} \end{pmatrix} \begin{pmatrix} \mu \\ \mu^2 \end{pmatrix}, (-2\boldsymbol{\eta}_2)^{1/2} Exp(-\frac{\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2}), \frac{x}{2\sigma_0^2}, \frac{x^2}{\sigma_0^2}, \frac{-1}{2\sigma_0^2})$$

$$\hat{u}(x) = (1, -1, -1, \mu_0, \mu_0^2)$$

**Answer to Question 2.5: KL-Divergence for exponential families**

$$\mathbb{E}_{x \sim \mathrm{EXP}(\boldsymbol{u}, \boldsymbol{\eta})} [\boldsymbol{u}(x)]$$

From our definition of KL-Divervence:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \int q(x \mid \boldsymbol{\eta}_1) ln(\frac{q(x \mid \boldsymbol{\eta}_1)}{q(x \mid \boldsymbol{\eta}_2)}) dx$$

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \int q(x \mid \boldsymbol{\eta}_1) ln(\frac{Exp(\boldsymbol{\eta}_1^T u(x) - \psi(\boldsymbol{\eta}_1))}{Exp(\boldsymbol{\eta}_2^T u(x) - \psi(\boldsymbol{\eta}_2))}) dx$$

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \int Exp(\boldsymbol{\eta}_1^T u(x) - \psi(\boldsymbol{\eta}_1)) ln(\frac{Exp(-\psi(\boldsymbol{\eta}_1)) Exp(\boldsymbol{\eta}_1^T u(x))}{Exp(-\psi(\boldsymbol{\eta}_2)) Exp(\boldsymbol{\eta}_2^T u(x))}) dx$$

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \int q(x \mid \boldsymbol{\eta}_1) ln(Exp(\psi(\boldsymbol{\eta}_2)) Exp(-\psi(\boldsymbol{\eta}_1)) Exp(-\boldsymbol{\eta}_2^T u(x)) Exp(\boldsymbol{\eta}_1^T u(x))) dx$$

Use property of logs to simplify:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \int q(x \mid \boldsymbol{\eta}_1)(\psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \boldsymbol{\eta}_2^T u(x) + \boldsymbol{\eta}_1^T u(x)) dx$$

Distribute the first term

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \boldsymbol{\eta}_1 \int q(x \mid \boldsymbol{\eta}_1) u(x) dx - \boldsymbol{\eta}_2 \int q(x \mid \boldsymbol{\eta}_1) u(x) dx - \int q(x \mid \boldsymbol{\eta}_1) \psi(\boldsymbol{\eta}_1) dx + \int q(x \mid \boldsymbol{\eta}_1) \psi(\boldsymbol{\eta}_2) dx$$

Since $\psi(\boldsymbol{\eta}_i)$ can also be treated as a constant:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \boldsymbol{\eta}_1 \int q(x \mid \boldsymbol{\eta}_1) u(x) dx - \boldsymbol{\eta}_2 \int q(x \mid \boldsymbol{\eta}_1) u(x) dx - \psi(\boldsymbol{\eta}_1) \int q(x \mid \boldsymbol{\eta}_1) dx + \psi(\boldsymbol{\eta}_2) \int q(x \mid \boldsymbol{\eta}_1) dx$$

We proved earlier in 2.1 that $\int Exp(\boldsymbol{\eta}_1^T u(x) - \psi(\boldsymbol{\eta}_1)) dx = 1$, therefore:
Since $\psi(\boldsymbol{\eta}_i)$ can also be treated as a constant:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \boldsymbol{\eta}_1 \int q(x \mid \boldsymbol{\eta}_1) u(x) dx - \boldsymbol{\eta}_2 \int q(x \mid \boldsymbol{\eta}_1) u(x) dx - \psi(\boldsymbol{\eta}_1) + \psi(\boldsymbol{\eta}_2)$$

We know that the expected value of U here follows the form:

$$\mathbb{E}_{x\sim \mathrm{EXP}(\boldsymbol{u},\boldsymbol{\eta})}\left[\boldsymbol{u}(x)\right] = \int xu(x)dx = \int exp(u,\boldsymbol{\eta_i})u(x)dx$$

Since we clearly have that form we can introduce the expected u value as:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \boldsymbol{\eta}_1\mathbb{E}_{x\sim \mathrm{EXP}(\boldsymbol{u},\boldsymbol{\eta_1})}\left[\boldsymbol{u}(x)\right] - \boldsymbol{\eta}_2\mathbb{E}_{x\sim \mathrm{EXP}(\boldsymbol{u},\boldsymbol{\eta_1})}\left[\boldsymbol{u}(x)\right] - \psi(\boldsymbol{\eta}_1) + \psi(\boldsymbol{\eta}_2)$$

By equation 2.7:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \boldsymbol{\eta}_1\lambda_1 - \boldsymbol{\eta}_2\lambda_1 - \psi(\boldsymbol{\eta}_1) + \psi(\boldsymbol{\eta}_2)$$

Rearrange:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \boldsymbol{\eta}_2\lambda_1 + \boldsymbol{\eta}_1\lambda_1$$

Pull out -$\lambda_1$:

$$D_{kl}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)$$

**Answer to Question 2.6: Pythagorean Theorem for exponential families**

To prove that this property holds under a certain condition we set assume the formula is true:

$$\psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) + \psi(\boldsymbol{\eta}_3) - \psi(\boldsymbol{\eta}21) - \lambda_2^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_2) = \psi(\boldsymbol{\eta}_3) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_1)$$

All lambdas cancel:

$$-\lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) - \lambda_2^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_2) = -\lambda_1^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_1)$$

Distribute:

$$-\lambda_1^T\boldsymbol{\eta}_2 + \lambda_1^T\boldsymbol{\eta}_1 - \lambda_2^T\boldsymbol{\eta}_3 + \lambda_2^T\boldsymbol{\eta}_2 = -\lambda_1^T\boldsymbol{\eta}_3 + \lambda_1^T\boldsymbol{\eta}_1$$

$$-\lambda_1^T\boldsymbol{\eta}_2 - \lambda_2^T\boldsymbol{\eta}_3 + \lambda_2^T\boldsymbol{\eta}_2 = -\lambda_1^T\boldsymbol{\eta}_3$$

$$-\lambda_1^T\boldsymbol{\eta}_2 + \lambda_1^T\boldsymbol{\eta}_3 - \lambda_2^T\boldsymbol{\eta}_3 + \lambda_2^T\boldsymbol{\eta}_2 = 0$$

$$\lambda_1^T(-\boldsymbol{\eta}_2 + \boldsymbol{\eta}_3) + \lambda_2^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3) = 0$$

Rearrange by multiplying each side by -1

$$\lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3) - \lambda_2^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3) = 0$$

$$(\lambda_1^T - \lambda_2^T)(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3) = 0$$

This equation says that these two must be perpendicular and therefore the property holds if they are perpendicular

# 3 Section 3: Utilising Expectation Maximization

**Answer to Question 3.1: Deriving the EMM Expectation Step**

Start by expanding bayes theorum:

$$p(x, z \mid v) = p(x_n, v)p(z) = \sum^{K} \pi_k q(x \mid \boldsymbol{\eta}_k) \prod^{k} \pi_k^{z_{nk}}$$

Substitute back into our initial equation:

$$\sum^{z} p(z \mid x, v^{old}) ln(\sum^{K} \pi_k q(x \mid \boldsymbol{\eta}_k) \prod^{k} \pi_k^{z_{nk}})$$

distribute log:

$$\sum^{z} p(z \mid x, v^{old}) \sum^{K} ln(\pi_k) + ln(q(x \mid \boldsymbol{\eta}_k)) z^{nk} \sum^{K} ln(\pi_k))$$

$$\sum^{z} p(z \mid x, v^{old}) z^{nk} \sum^{N} \sum^{K} ln(\pi_k) + ln(q(x \mid \boldsymbol{\eta}_k)) \sum^{K} ln(\pi_k))$$

$$\sum^{z} p(z \mid x, v^{old}) z^{nk} \sum^{N} \sum^{K} ln(\pi_k) + ln(q(x \mid \boldsymbol{\eta}_k)$$

From equation 3.7:

$$p(z_n = 1 \mid x_n, v^{old}) \sum^{N} \sum^{K} ln(\pi_k) + ln(q(x \mid \boldsymbol{\eta}_k)$$

Expand on $p(z_n = 1 \mid x_n, v^{old})$ with bayes theorum:

$$p(z_n = 1 \mid x_n, v^{old}) = \frac{p(x_n \mid z_n = 1, v^{old}) p(z_n = 1)}{p(x_n, v^{old})}$$

$$p(z_n = 1 \mid x_n, v^{old}) = \frac{\prod^{K} q(x_n \mid \boldsymbol{\eta}_k)^{z_{nk}} \prod^{K} (\pi_k^{old})^{z_{nk}}}{p(x_n, v^{old})}$$

$$p(z_n = 1 \mid x_n, v^{old}) = \frac{\prod^{K} (\pi_k^{old})^{z_{nk}} q(x_n \mid \boldsymbol{\eta}_k)^{z_{nk}}}{p(x_n, v^{old})}$$

we know that the bottom term should be the intgral of the top, since the integral of the product is the sum over the product we introduce:

$$p(z_n = 1 \mid x_n, v^{old}) = \frac{\prod^{K} (\pi_k^{old})^{z_{nk}} q(x_n \mid \boldsymbol{\eta}_k)^{z_{nk}}}{\sum^{J} \prod^{K} (\pi_k^{old})^{z_{nk}} q(x_n \mid \boldsymbol{\eta}_k)^{z_{nk}}}$$

raising to the $z_{nk}$ cancels:

$$p(z_n = 1 \mid x_n, v^{old}) = \frac{\prod^{K} (\pi_k^{old}) q(x_n \mid \boldsymbol{\eta}_k)}{\sum^{J} \prod^{K} (\pi_k^{old}) q(x_n \mid \boldsymbol{\eta}_k)}$$

$$p(z_n = 1 \mid x_n, v^{old}) = \frac{(\pi_k^{old})q(x_n \mid \boldsymbol{\eta}_k)}{\sum^J (\pi_k^{old})q(x_n \mid \boldsymbol{\eta}_k)}$$

We now insert this back into our original equation giving us:

$$\frac{(\pi_k^{old})q(x_n \mid \boldsymbol{\eta}_k)}{\sum^J (\pi_k^{old})q(x_n \mid \boldsymbol{\eta}_k)} \sum^N \sum^K ln(\pi_k) + ln(q(x \mid \boldsymbol{\eta}_k)$$

$$\sum^N \sum^K \frac{(\pi_k^{old})q(x_n \mid \boldsymbol{\eta}_k)}{\sum^J (\pi_k^{old})q(x_n \mid \boldsymbol{\eta}_k)} (ln(\pi_k) + ln(q(x \mid \boldsymbol{\eta}_k))$$

Arriving us at our derivation.

**Answer to Question 3.2: Deriving the EMM Maximisation Step**

Since for a binary distribution as p-¿0.5 and n tends to infinity the distribution will become closer to symmetrical and closer to normal. Since 0.45 is closer to 0.5 than 0.65 the variance in B is larger meaning that it has a greater chance of being the observed class.