Question 1.2: Computing the probabilities (4 Points)

Can you compute probabilities of the class being in program A and program B given the observation, and from there reason about which program is more likely?

Hint: You may need to write out expressions involving the number of students in this class, say n, which is fixed, but not provided in this question. Do the probability of the observed class being in program A depend on class size n?
You may want to consider the ratio of likelihoods.

Name : Yixi Rao
UID : U6826541

let X be a random variable indicating the program I came in. So. $X = \{A, B\}$

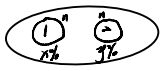
let O be a random variable indicating the percentage of boys who are in the program I came in
so $O = \{0\%, 1\% \cdots 100\%\}$

$$P(X=A \mid O=55\%) = \frac{P(O=55\% \mid X=A)\cdot P(X=A)}{P(O=55\%)} = \frac{P(O=55\% \mid X=A)\cdot P(X=A)}{\sum_x P(X=x, O=55\%)} = \frac{P(O=55\% \mid X=A)\cdot P(X=A)}{\sum_x P(X=x \mid O=55\%)\, P(X=x)}$$

$$= \frac{P(O=55\% \mid X=A)\cdot P(X=A)}{P(O=55\% \mid X=A)\cdot P(X=A) + P(O=55\% \mid X=B)\cdot P(X=B)}$$

Suppose that there are two classes in Program A and B

define $x\%$ be the percentage of boys in class 1 in Program A, and $y\%$ for class 2 in Program A

e.g. 

So we have $\begin{cases} nx + ny = 2n \cdot 0.65 \Rightarrow x+y = 1.3 \\ 0 \le x \le 1 \\ 0 \le y \le 1 \end{cases}$, we can conclude that $\begin{cases} 0.3 \le x \le 1 \\ 0.3 \le y \le 1 \end{cases}$

So the x or y has 70 possible percentage's values, so $P(O=55\% \mid X=A) = \frac{1}{70}$

do the same trick to calculate $P(O=55\% \mid X=B)$, We can get $P(O=55\% \mid X=B) = \frac{1}{90}$

And $P(X=A) = P(X=B) = \frac{1}{2}$ Since I enter in a classroom in random

So $P(X=A \mid O=55\%) = \frac{9}{16}$ $P(X=B \mid O=55\%) = \frac{7}{16}$

$P(X=A \mid O=55\%) > P(X=B \mid O=55\%)$

Question 1.3: Well-reasoned hunch (1 Points)

Without computing the posterior probabilities for program A or B, could you make a reasoned argument, based on properties of the binomial distribution, that program B is more likely?

define X_A is a random variable indicating the number of boys sampled from program A in n trials

so $P(X_A = k) = C_n^k 0.65^k 0.35^{n-k}$

define X_B is a random variable indicating the number of boys sampled from program B in n trials

so $P(X_B = k) = C_n^k 0.45^k 0.55^{n-k}$

$\begin{cases} D(X_A) = n \cdot 0.65 \cdot 0.35 = 0.225n \\ D(X_B) = n \cdot 0.45 \cdot 0.55 = 0.2475 \end{cases} \Rightarrow D(X_B) > D(X_A)$

The variance of binomal for program B is larger than the variance of binomal for program A, so Program B has more variability in the number of boys. so it is likely that B has as many as 55% boys.

Question 2.1: Verifying valid distribution (5 Points)

Show that $q(x \mid \eta)$ given by Eqs. (2.1) and (2.2) is a valid probability density function.

$q(x|\eta) = \exp(\eta^T \cdot u(x) - \psi(\eta))$

$\psi(x) = \log \int \exp(\eta^T \cdot u(x))\, dx$

$q(x|\eta) = \frac{1}{\exp(\psi(x))} \cdot \exp(\eta^T \cdot u(x)) = \frac{1}{\exp(\log \int \exp(\eta^T \cdot u(x)) dx)} \cdot \exp(\eta^T \cdot u(x))$

$= \frac{\exp(\eta^T \cdot u(x))}{\int \exp(\eta^T \cdot u(x))\, dx}$

To prove $q(x|\eta)$ is a valid probability density function, we need to prove

① $q(x|\eta)$ is nonnegative for each value of the random variable

② $\int_x q(x|\eta)\, dx = 1$

①: $q(x|\eta) = \frac{\exp(\eta^T \cdot u(x))}{\int \exp(\eta^T \cdot u(x))\, dx}$, and we know $\forall x.\ \exp(x) > 0$

so $q(x|\eta)$ must > 0

②: $\int_x q(x|\eta)\, dx = \int_x \frac{\exp(\eta^T \cdot u(x))}{\int \exp(\eta^T \cdot u(x))\, dx} = \frac{\int_x \exp(\eta^T \cdot u(x))\, dx}{\int_x \exp(\eta^T \cdot u(x))\, dx} = 1$

Question 2.2: A Bayesian example (Part 1) (13 Points)

Suppose that we have a likelihood and prior distribution given by,

$$x \mid \mu \sim \mathcal{N}(\mu, \sigma^2) \qquad q(x \mid \eta) = \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta})),$$
$$\mu \sim \textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}), \qquad \psi(\boldsymbol{\eta}) = \log \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x))\, \mathrm{d}x$$

where $\mathcal{N}$ is a 1-dimensional Gaussian distribution with mean μ and standard deviation σ. Show that $\mu \mid x \sim \textsc{Exp}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\eta}})$, for some $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\eta}}$.

Make sure to clearly specify the function $\hat{\boldsymbol{u}}$ and parameters $\hat{\boldsymbol{\eta}}$.

Hint: The integral for $q(\mu)$ does not need to be simplified / solved.

want to find $\hat{u}, \hat{\eta}$ to satisfy $p(\mu|x, \hat{\eta}) = \textsc{Exp}(\hat{u}, \hat{\eta}) = \frac{\exp(\hat{\eta}^T \cdot \hat{u}(\mu))}{\int \exp(\hat{\eta}^T \cdot \hat{u}(\mu))\, d\mu}$

↓

$$p(u|x) = \frac{p(x|u)p(u)}{p(x)} = \frac{p(x|u)p(u)}{\int p(x|u)p(u)\,du} = \frac{\mathcal{N}(u,\sigma^2)\cdot\frac{\exp(\eta^T\cdot u(u))}{\int\exp(\eta^T\cdot u(u))\,du}}{\int \mathcal{N}(u,\sigma^2)\cdot\frac{\exp(\eta^T\cdot u(u))}{\int\exp(\eta^T\cdot u(u))\,du}\,du} \quad ① \quad ②$$

find $\hat{\eta}, \hat{u}$ to match $\mathcal{N}(u,\sigma^2)\cdot\frac{\exp(\eta^T\cdot u(u))}{\int\exp(\eta^T\cdot u(u))\,du} = \exp(\hat{\eta}^T\cdot\hat{u}(u))$

$$① = \frac{1}{\sqrt{2\pi\sigma^2}}\cdot\exp\left(-\frac{(x-u)^2}{2\sigma^2}\right)\cdot\frac{\exp(\eta^T\cdot u(u))}{\int\exp(\eta^T\cdot u(u))\,du}$$

$$① = \exp\left(\log(2\pi\sigma^2)^{-\frac{1}{2}}\right)\cdot\exp\left(-\frac{1}{2\sigma^2}(x^2-2ux+u^2)\right)\cdot\frac{\exp(\eta^T\cdot u(u))}{\int\exp(\eta^T\cdot u(u))\,du}$$

$$① = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2)\right)\cdot\exp\left(-\frac{x^2}{2\sigma^2}+\frac{ux}{\sigma^2}-\frac{u^2}{2\sigma^2}\right)\cdot\frac{\exp(\eta^T\cdot u(u))}{\int\exp(\eta^T\cdot u(u))\,du}$$

$$① = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2)-\frac{x^2}{2\sigma^2}+\frac{ux}{\sigma^2}-\frac{u^2}{2\sigma^2}+\eta^T\cdot u(u)\right)\cdot\frac{1}{\int\exp(\eta^T\cdot u(u))\,du}$$

$$① = \exp\left(\frac{ux}{\sigma^2}-\frac{u^2}{2\sigma^2}+\eta^T\cdot u(u)-\frac{1}{2}\log(2\pi\sigma^2)-\frac{x^2}{2\sigma^2}\right)\cdot\exp\left(-\log\left(\int\exp(\eta^T\cdot u(u))\,du\right)\right)$$

$$① = \exp\left(\frac{ux}{\sigma^2}-\frac{u^2}{2\sigma^2}+\eta^T\cdot u(u)-\frac{1}{2}\log(2\pi\sigma^2)-\frac{x^2}{2\sigma^2}-\log\left(\int\exp(\eta^T\cdot u(u))\,du\right)\right)$$

$$① = \exp\left(\begin{bmatrix}u & u^2 & 1\end{bmatrix}\begin{bmatrix}\frac{x}{\sigma^2}\\ -\frac{1}{2\sigma^2}\\ \eta^T\cdot u(u)-\frac{1}{2}\log(2\pi\sigma^2)-\frac{x^2}{2\sigma^2}-\log\left(\int\exp(\eta^T\cdot u(u))\,du\right)\end{bmatrix}\right)$$

so $\hat{u}(u) = \begin{bmatrix}u\\ u^2\\ 1\end{bmatrix}$

and $\hat{\eta} = \begin{bmatrix}\frac{x}{\sigma^2}\\ -\frac{1}{2\sigma^2}\\ \eta^T u(u)-\frac{1}{2}\log(2\pi\sigma^2)-\frac{x^2}{2\sigma^2}-\log\left(\int\exp(\eta^T\cdot u(u))\,du\right)\end{bmatrix}$

Question 2.3: A Bayesian example (Part 2) (7 Points)

Suppose that we have the same likelihood as per Question 2.2 but with a Gaussian prior distribution

$$x \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2).$$

First, find the parameters of an exponential family such that $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}) = \mathcal{N}(\mu_0, \sigma_0^2)$ (in distribution) with $\boldsymbol{u} = (x, x^2)$.

Then use the result of Question 2.2 to find the parameters of $\mu \mid x$.

$$p(u \mid u_0, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(u-u_0)^2}{2\sigma_0^2}\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(u^2 - 2u_0 u + u_0^2)\right)$$

$$= \exp\left(\log(2\pi\sigma_0^2)^{-\frac{1}{2}}\right) \cdot \exp\left(-\frac{1}{2\sigma_0^2}\begin{bmatrix}-2u_0 & 1\end{bmatrix}\begin{bmatrix}u \\ u^2\end{bmatrix} - \frac{u_0^2}{2\sigma_0^2}\right)$$

$$= \exp\left(-\frac{1}{2}\log(2\pi\sigma_0^2)\right) \cdot \exp\left(-\frac{1}{2\sigma_0^2}\begin{bmatrix}-2u_0 & 1\end{bmatrix}\begin{bmatrix}u \\ u^2\end{bmatrix} - \frac{u_0^2}{2\sigma_0^2}\right)$$

$$= \exp\left(\begin{bmatrix}\frac{u_0}{\sigma_0^2} & -\frac{1}{2\sigma_0^2}\end{bmatrix}\begin{bmatrix}u \\ u^2\end{bmatrix} - \frac{u_0^2}{2\sigma_0^2} - \frac{1}{2}\log(2\pi\sigma_0^2)\right)$$

$$= \exp\left(\begin{bmatrix}\frac{u_0}{\sigma_0^2} & -\frac{1}{2\sigma_0^2}\end{bmatrix}\begin{bmatrix}u \\ u^2\end{bmatrix} - \left(\frac{u_0^2}{2\sigma_0^2} + \frac{1}{2}\log(2\pi\sigma_0^2)\right)\right)$$

$$\text{so } \eta = \begin{bmatrix}\frac{u_0}{\sigma_0^2} \\ -\frac{1}{2\sigma_0^2}\end{bmatrix} \qquad u(u) = \begin{bmatrix}u \\ u^2\end{bmatrix}$$

$$\hat{\eta} = \begin{bmatrix}\frac{x}{\sigma^2} \\ -\frac{1}{2\sigma^2} \\ \eta^T u(u) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} - \log\left(\int \exp(\eta^T \cdot u(u))\, du\right)\end{bmatrix}$$

$$= \begin{bmatrix}\frac{x}{\sigma^2} \\ -\frac{1}{2\sigma^2} \\ \frac{u_0 u}{\sigma_0^2} - \frac{u^2}{2\sigma_0^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} - \log\left(\int \exp\left(\frac{u_0 u}{\sigma_0^2} - \frac{u^2}{2\sigma_0^2}\right) du\right)\end{bmatrix}$$

We notice that $\hat{\eta}$ still has some u, so we re-factorise it

$$\hat{\eta}^T \cdot u(u) = \begin{bmatrix} \frac{x}{\sigma^2} \\ -\frac{1}{2\sigma^2} \\ \frac{\mu_0 u}{\sigma_0^2} - \frac{u^2}{2\sigma_0^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} - \log\left(\int \exp(\mu_0 u - \frac{u^2}{2\sigma_0^2})\,du\right) \end{bmatrix}^T \begin{bmatrix} u \\ u^2 \\ 1 \end{bmatrix}$$

$$= \frac{u \cdot x}{\sigma^2} - \frac{u^2}{2\sigma^2} + \frac{\mu_0 u}{\sigma_0^2} - \frac{u^2}{2\sigma_0^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} - \log\left(\int \exp(\eta^T \cdot u(u))\,du\right)$$

from the expression:

$$p(u|\mu_0, \eta_0) = \exp\left(\begin{bmatrix} \frac{\mu_0}{\sigma_0^2} & -\frac{1}{2\sigma_0^2} \end{bmatrix} \begin{bmatrix} u \\ u^2 \end{bmatrix} - \left(\frac{\mu_0^2}{2\sigma_0^2} + \frac{1}{2}\log(2\pi\sigma_0^2)\right)\right)$$

We know $\frac{\mu_0^2}{2\sigma_0^2} + \frac{1}{2}\log(2\pi\sigma_0^2)$ is $\psi(\eta)$

So $\log\int \exp(\eta^T \cdot u(u))\,du = \frac{\mu_0^2}{2\sigma_0^2} + \frac{1}{2}\log(2\pi\sigma_0^2)$

$$\hat{\eta}^T u(u) = -\frac{u^2}{2\sigma_0^2} + \frac{\mu_0 u}{\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2} - \frac{x^2}{2\sigma^2} + \frac{ux}{\sigma^2} - \frac{u^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log(2\pi\sigma_0^2)$$

$$\hat{\eta}^T \cdot u(u) = u^2\left(-\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma^2}\right) + u\left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right) - \frac{\mu_0^2}{2\sigma_0^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log(2\pi\sigma_0^2)$$

So $\hat{\eta} = \begin{bmatrix} -\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma^2} \\ \frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2} \\ -\frac{\mu_0^2}{2\sigma_0^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log(2\pi\sigma_0^2) \end{bmatrix}$ $\quad u(u) = \begin{bmatrix} u^2 \\ u \\ 1 \end{bmatrix}$

Question 2.5: KL-Divergence for exponential families (15 Points)

Show that the KL-divergence of distributions $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_1)$ and $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_2)$ within the same exponential family is given by:

$$D_{\text{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \boldsymbol{\lambda}_1^\top (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1).$$

Hint: Use Eq. (2.7).

derive Eq. (2.7).

$$\psi(\eta) = \log \int \exp(\eta^T \cdot u(x))\, dx$$

$$\exp(\psi(\eta)) = \int \exp(\eta^T \cdot u(x))\, dx$$

$$\exp(\psi(\eta)) \cdot \psi(\eta)' = \frac{\partial}{\partial \eta} \left(\int \exp(\eta^T \cdot u(x))\, dx \right)$$

$$= \int \exp(\eta^T \cdot u(x)) \cdot u(x)\, dx$$

$$\psi(\eta)' = \frac{1}{\exp(\psi(\eta))} \int \exp(\eta^T \cdot u(x)) \cdot u(x)\, dx$$

$$\psi(\eta)' = \int \exp(\eta^T \cdot u(x) - \psi(\eta)) \cdot u(x)\, dx$$

$$\Rightarrow \Delta\psi(\eta) = \psi(\eta)' = \int \exp(\eta^T \cdot u(x) - \psi(\eta)) \cdot u(x)\, dx = \int q(x|\eta) \cdot u(x)\, dx = \underset{x \sim \text{Exp}(u,\eta)}{E}[u(x)] = \lambda \quad (2.7.1)$$

now:

$$D_{KL}[q(x|\eta_1) : q(x|\eta_2)] = \int q(x|\eta_1) \log \frac{q(x|\eta_1)}{q(x|\eta_2)}\, dx$$

$$= \int \exp(\eta_1^T \cdot u(x) - \psi(\eta_1)) \cdot \log \frac{\exp(\eta_1 \cdot u(x))}{\exp(\psi(\eta_1))} \cdot \frac{\exp(\psi(\eta_2))}{\exp(\eta_2 u(x))}\, dx$$

$$= \int \exp(\eta_1^T \cdot u(x) - \psi(\eta_1)) \log \exp(\eta_1 u(x) + \psi(\eta_2) - \psi(\eta_1) - \eta_2 u(x))\, dx$$

$$= \int \exp(\eta_1^T \cdot u(x) - \psi(\eta_1)) \cdot (\eta_1 u(x) + \psi(\eta_2) - \psi(\eta_1) - \eta_2 u(x))\, dx$$

$$= \int \exp(\eta_1^T \cdot u(x) - \psi(\eta_1)) \cdot (\eta_1 - \eta_2) u(x)\, dx + \int \exp(\eta_1^T \cdot u(x) - \psi(\eta_1))\, \psi(\eta_2)\, dx - \cdots$$
$$- \int \psi(\eta_1) \cdot \exp(\eta_1^T \cdot u(x) - \psi(\eta_1))\, dx$$

according to Eq. (2.7.1), and $\int q(x|\eta)\, dx = 1$, we get

$$= (\eta_1 - \eta_2) \lambda_1^T + \psi(\eta_2) - \psi(\eta_1)$$

$$= \psi(\eta_2) - \psi(\eta_1) - \lambda_1^T \cdot (\eta_2 - \eta_1). \text{ proved.}$$

Question 2.6: Pythagorean Theorem for exponential families (10 Points)

Let $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_1)$, $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_2)$, and $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_3)$ be distributions within the same exponential family. Furthermore, let

$$a^2 = D_{\text{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2]; \quad b^2 = D_{\text{KL}}[\boldsymbol{\eta}_2 : \boldsymbol{\eta}_3]; \quad c^2 = D_{\text{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_3].$$

Prove that Fig. 1 holds. That is, show that $a^2 + b^2 = c^2$ iff the difference in natural parameters $(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3)$ is perpendicular to the difference in expectation parameters $(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)$.

We need to prove $a^2+b^2=c^2 \Leftrightarrow (\lambda_1-\lambda_2)^T\cdot(\eta_2-\eta_3)=0$ in two directions

① $a^2+b^2=c^2 \Rightarrow (\lambda_1-\lambda_2)^T\cdot(\eta_2-\eta_3)=0$

$a^2+b^2=c^2$

$D_{KL}[\eta_1:\eta_2]+D_{KL}[\eta_2:\eta_3]=D_{KL}[\eta_1:\eta_3]$ (use Eq in Q2.5)

$\psi(\eta_2)-\psi(\eta_1)-\lambda_1^T\cdot(\eta_2-\eta_1)+\psi(\eta_3)-\psi(\eta_2)-\lambda_2^T\cdot(\eta_3-\eta_2)=\psi(\eta_3)-\psi(\eta_1)-\lambda_1^T(\eta_3-\eta_1)$

$\psi(\eta_2)-\lambda_1^T\cdot(\eta_2-\eta_1)-\psi(\eta_2)-\lambda_2^T\cdot(\eta_3-\eta_2)=-\lambda_1^T(\eta_3-\eta_1)$

$\psi(\eta_2)-\lambda_1^T\eta_2+\lambda_1^T\eta_1-\psi(\eta_2)-\lambda_2^T\eta_3+\lambda_2^T\eta_2=-\lambda_1^T\eta_3+\lambda_1^T\eta_1$

$\psi(\eta_2)-\lambda_1^T\eta_2-\psi(\eta_2)-\lambda_2^T\eta_3+\lambda_2^T\eta_2=-\lambda_1^T\eta_3$

$-\lambda_1^T\eta_2-\lambda_2^T\eta_3+\lambda_2^T\eta_2=-\lambda_1^T\eta_3$

$-\lambda_1^T\eta_3+\lambda_1^T\eta_2+\lambda_2^T\eta_3-\lambda_2^T\eta_2=0$

$(\lambda_1^T-\lambda_2^T)\cdot(\eta_2-\eta_3)=0 \Rightarrow (\lambda_1-\lambda_2)^T\cdot(\eta_2-\eta_3)=0 \Rightarrow$ proved

② $(\lambda_1-\lambda_2)^T\cdot(\eta_2-\eta_3)=0 \Rightarrow a^2+b^2=c^2$

$(\lambda_1^T-\lambda_2^T)\cdot(\eta_2-\eta_3)=\lambda_1^T\eta_2-\lambda_1^T\eta_3-\lambda_2^T\eta_2+\lambda_2^T\eta_3=0$

So $\lambda_2^T\eta_2-\lambda_1^T\eta_2-\lambda_2^T\eta_3=-\lambda_1^T\eta_3$

$a^2+b^2=\psi(\eta_2)-\psi(\eta_1)-\lambda_1^T\cdot(\eta_2-\eta_1)+\psi(\eta_3)-\psi(\eta_2)-\lambda_2^T\cdot(\eta_3-\eta_2)$

$=\psi(\eta_2)-\psi(\eta_1)-\lambda_1^T\eta_2+\lambda_1^T\eta_1+\psi(\eta_3)-\psi(\eta_2)-\lambda_2^T\eta_3+\lambda_2^T\eta_2$

Replace $\rightarrow$ $=\lambda_2^T\eta_2-\lambda_1^T\eta_2-\lambda_2^T\eta_3+\psi(\eta_2)-\psi(\eta_1)+\lambda_1^T\eta_1+\psi(\eta_3)-\psi(\eta_2)$

$=-\lambda_1^T\eta_3+\psi(\eta_2)-\psi(\eta_1)+\lambda_1^T\eta_1+\psi(\eta_3)-\psi(\eta_2)$

$=-\lambda_1^T\eta_3-\psi(\eta_1)+\lambda_1^T\eta_1+\psi(\eta_3)=\psi(\eta_3)-\psi(\eta_1)-\lambda_1^T\eta_3+\lambda_1^T\eta_1=D_{KL}[\eta_1:\eta_3]=c^2$ proved.

Question 3.1: Deriving the EMM expectation step (8 Points)

Derive the expectation we will maximise in EMM EM. In particular, show that

$$\sum_{Z} p(Z \mid X, \vartheta^{\text{old}}) \log p(X, Z \mid \vartheta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma^{\text{old}}(z_{nk}) \left(\log \pi_k + \log q(x_n \mid \boldsymbol{\eta}_k)\right) \tag{3.5}$$

where

$$\gamma^{\text{old}}(z_{nk}) = \frac{\pi_k^{\text{old}} q(x_n \mid \boldsymbol{\eta}_k^{\text{old}})}{\sum_j \pi_j^{\text{old}} q(x_n \mid \boldsymbol{\eta}_j^{\text{old}})}. \tag{3.6}$$

let $\theta^{old} = \theta^t$

$$\sum_{Z} p(Z|X,\theta^t) \cdot \log p(X,Z|\theta)$$

$$= \sum_{Z} \log \prod_{n=1}^{N} p(x_n, z_n|\theta) \prod_{n=1}^{N} p(z_n|x_n,\theta^t)$$

$$= \sum_{Z} \sum_{n=1}^{N} \log p(x_n, z_n|\theta) \cdot \prod_{n=1}^{N} p(z_n|x_n,\theta^t) = \sum_{Z} \sum_{n=1}^{N} \log p(x_n|z_n,\theta) \cdot p(z_n|\theta) \cdot \prod_{n=1}^{N} p(z_n|x_n,\theta^t)$$

$$= \sum_{Z} \sum_{n=1}^{N} \log \prod_{k=1}^{K} \left(q(x_n|\eta_k)\right)^{z_{nk}} \cdot \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot p(Z|X,\theta^t)$$

$$= \sum_{Z} \sum_{n=1}^{N} \log \prod_{k=1}^{K} \left(q(x_n|\eta_k) \cdot \pi_k\right)^{z_{nk}} \cdot p(Z|X,\theta^t)$$

$$= \sum_{Z} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \cdot \log\left(q(x_n|\eta_k) \cdot \pi_k\right) \cdot p(Z|X,\theta^t)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{Z} z_{nk} \cdot p(Z|X,\theta^t) \cdot \log\left(q(x_n|\eta_k) \cdot \pi_k\right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk}=1 \mid x_n, \theta^t) \log\left(q(x_n|\eta_k) \cdot \pi_k\right) \quad \text{(using Eq. (3.7))}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{p(z_{nk}=1, x_n|\theta^t)}{p(x_n|\theta^t)} \log\left(q(x_n|\eta_k) \cdot \pi_k\right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{p(z_{nk}=1, x_n|\theta^t)}{\sum_{k=1}^{K} p(z_{nk}=1, x_n|\theta^t)} \log\left(q(x_n|\eta_k) \cdot \pi_k\right)$$

Now derive $p(z_{nk}=1, x_n|\theta^t)$

$$p(x_n, z_{nk}=1|\theta^t) = p(z_{nk}=1|\theta^t)\, p(x_n|z_{nk}=1,\theta^t)$$

$$= p(x_n|z_{nk}=1,\theta^t) \cdot p(z_{nk}=1|\theta^t) \quad \text{(then using Eq. (3.3))}$$

$$= q(x_n|\eta_k^t) \cdot \pi_k^t \quad \text{(using Eq. (3.4))}$$

So $\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{p(z_{nk}=1, x_n|\theta^t)}{\sum_{k=1}^{K} p(z_{nk}=1, x_n|\theta^t)} \log\left(q(x_n|\eta_k) \cdot \pi_k\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{q(x_n|\eta_k^t) \cdot \pi_k^t}{\sum_j \pi_j^t q(x_n|\eta_j^t)} \left(\log \pi_k + \log q(x_n|\eta_k)\right)$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} r^t(z_{nk}) \left(\log \pi_k + \log q(x_n|\eta_k)\right)$$

Proved.

Question 3.2: Deriving the EMM maximisation step (7 Points)

Derive the maximisation updates for EMM EM for mixing parameters π and expectation parameters $\mathbf{\Lambda}$. That is, assuming that Eq. (3.5) is concave in π and $\boldsymbol{H}$, show that

$$\pi_k = \frac{N_k}{N} \qquad (3.9) \qquad \text{and} \qquad \boldsymbol{\lambda}_k = \frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{u}(x_n)\cdot\gamma^{\text{old}}(z_{nk})), \qquad (3.10)$$

where $N_k = \sum_{n=1}^{N}\gamma^{\text{old}}(z_{nk})$.

Thus show that we have the update equation for the natural parameters $\boldsymbol{H}$ given by

$$\boldsymbol{\eta}_k = \nabla\varphi\left(\frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{u}(x_n)\cdot\gamma^{\text{old}}(z_{nk}))\right). \qquad (3.11)$$

let all $\gamma^{old} = \gamma^t$

(3.9)

$\pi_k^{t+1} = \arg\max_{\pi_k} \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\log\pi_k + \log q(x_n|\eta_k)\right)\cdot\gamma^t(z_{nk})\Big)$, s.t. $\sum_{k=1}^{K}\pi_k = 1$

We observe that this formular: $\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\log\pi_k + \log q(x_n|\eta_k)\right)\cdot\frac{q(x_n|\eta_k^t)\cdot\pi_k^t}{\sum_j \pi_j^t q(x_n|\eta_j^t)}$, only

$\sum_{n=1}^{N}\sum_{k=1}^{K}(\log\pi_k\cdot\gamma^t(z_{nk}))$ contains π_k, so

define $\mathcal{L}(\pi_k,\lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K}(\log\pi_k\cdot\gamma^t(z_{nk})) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$

$\frac{\partial\mathcal{L}}{\partial\pi_k} = \sum_{i=1}^{N}\frac{1}{\pi_k}\cdot\gamma^t(z_{nk}) + \lambda$

let $\frac{\partial\mathcal{L}}{\partial\pi_k} = 0$ so

$\sum_{n=1}^{N}\frac{1}{\pi_k}\cdot\gamma^t(z_{nk}) + \lambda = 0$ (times π_k on both sides)

$\sum_{n=1}^{N}\gamma^t(z_{nk}) + \pi_k\lambda = 0$ This is true for all $k\in\{1,2,3\cdots K\}$, so we can summarise it as

$\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma^t(z_{nk}) + \pi_k\lambda = 0$

$\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma^t(z_{nk}) + \sum_{k=1}^{K}\pi_k\lambda = 0$ (We know $\sum_{k=1}^{K}\pi_k = 1$)

$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma^t(z_{nk}) + \lambda = 0$ (We know $\sum_{k=1}^{K}\gamma^t(z_{nk}) = 1$)

$\Rightarrow N + \lambda = 0 \Rightarrow \lambda = -N$

This equation can be $\sum_{n=1}^{N}\gamma^t(z_{nk}) + \pi_k\lambda = 0 \equiv \sum_{n=1}^{N}\gamma^t(z_{nk}) - N\pi_k = 0 \Rightarrow \pi_k^{t+1} = \frac{1}{N}\sum_{n=1}^{N}\gamma^t(z_{nk}) = \frac{N_k}{N}$. proved.

(3.10)

$\eta_k^{t+1} = \arg\max_{\eta_k} \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\log\pi_k + \log q(x_n|\eta_k)\right)\cdot\gamma^t(z_{nk})$ only $\sum_{n=1}^{N}\sum_{k=1}^{K}\log(q(x_n|\eta_k))\cdot\gamma^t(z_{nk})$ contains η_k

So define $f(\eta_k) = \sum_{n=1}^{N}\sum_{k=1}^{K}\log(q(x_n|\eta_k))\cdot\gamma^t(z_{nk}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\eta_k^T u(x_n) - \psi(\eta_k)\right)\cdot\gamma^t(z_{nk})$

$\frac{\partial f(\eta_k)}{\partial\eta_k} = \sum_{n=1}^{N}(u(x_n) - \nabla\psi(\eta_k))\gamma^t(z_{nk}) = \sum_{n=1}^{N}(u(x_n) - \lambda_k)\gamma^t(z_{nk})$

let $\frac{\partial f(\eta_k)}{\partial \eta_k} = 0 = \sum_{n=1}^{N} u(x_n) \cdot r^t(z_{nk}) - \sum_{n=1}^{N} r^t(z_{nk}) \cdot \lambda_k$

$\sum_{n=1}^{N} u(x_n) r^t(z_{nk}) = \lambda_k \cdot \sum_{n=1}^{N} r^t(z_{nk}) = \lambda_k N_k$

$\therefore \lambda_k^{t+1} = \frac{1}{N_k} \sum_{n=1}^{N} u(x_n) r^t(z_{nk})$. proved.

(3.11)

We can use the E.q. (3.8) without proving, so we can use the proved E.q. (3.10) to show that

$\eta_k = \nabla\varphi(\lambda_k) = \nabla\varphi\left(\frac{1}{N_k} \sum_{n=1}^{N} (u(x_n) \cdot r^{old}(z_{nk}))\right)$. proved.....(3.11)