[PS 3]

#1.

(a) $$\frac{\partial l}{\partial w_{12}^{[1]}} = \frac{\partial l}{\partial o}\frac{\partial o}{\partial z^o}\frac{\partial z^o}{\partial h_2}\frac{\partial h_2}{\partial z^{h_2}}\frac{\partial z^{h_2}}{\partial w_{12}^{[1]}}$$

$$= \frac{1}{m}\sum 2o^i \cdot o^i(1-o^i)\cdot w_2^{[2]}\cdot h_2^i(1-h_2^i)\cdot x_1^i$$

(b) from triangle hint,

$$\underbrace{\begin{bmatrix} -0.3 & 1 & 0 \\ -0.3 & 0 & 1 \\ 4.2 & -1 & -1 \end{bmatrix}}_{W^{[1]}} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$ : $1^{st}$ mat. mult.

$(W_{ij}^{[1]} = w_{j,i}\ ,\ j=0,1,2,\ i=1,2,3)$

If point is within triangle, $f(z^h) = [1\ 1\ 1]^T$.

$2^{nd}$ mat. mult (for points in triangle)

: $$\underbrace{\begin{bmatrix} -1 & -1 & -1 & 2.5 \end{bmatrix}}_{w^{[2]}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$(\{w_j^{[2]},\ j=0,1,2,3\})$

$\forall a$ s.t $2<a<3$ OK.

(c) No. Current classes can't be seperated linearly, and taking $1^{st}$ mat. mult is affine transform from 2D to 3D if activation is linear. So it's still not linearly-seperable, resulting error while solving by step function.

#2. $$\log \prod_i p(x^i,\theta) = \log p(\theta)\prod_i \sum_{z^i} p(x^i, z^i|\theta)$$

$$= \sum_i \log \sum_{z^i} p(x^i, z^i|\theta) + \log p(\theta)$$

$$\geq \sum_i \sum_{z^i} Q_i(z^i)\log\frac{p(x^i,z^i|\theta)}{Q_i(z^i)} + \log p(\theta)$$

to tighten lower bound

Since $Q_i$ is any dist. for $z^i$, we can set $Q_i$ s.t $\frac{p(x^i,z^i|\theta)}{Q_i(z^i)} = c$ in E step. $\sum_{z^i} Q_i(z^i) = 1 \Rightarrow c = \underbrace{\sum_{z^i} p(x^i, z^i|\theta)}_{=p(x^i|\theta)}$ & $Q_i(z^i) = \frac{p(x^i,z^i|\theta)}{p(x^i|\theta)}$

$= p(z^i|x^i,\theta)$

$\Rightarrow$ In E step, set : ( establish & tighten lower bound.)

Iterate

$$Q_i(z^i) = p(z^i \mid x^i, \theta)$$

In M step, maximise lower bound :

$$\theta = \underset{\theta}{\operatorname{argmax}} \left[ \sum_i \sum_{z^i} \left( Q_i(z^i) \log \frac{p(x^i, z^i \mid \theta)}{Q_i(z^i)} \right) + \log p(\theta) \right]$$

( tune $\theta$ to maximise lower bound)

// The only change to get $\theta_{MAP}$ compared to $\theta_{ML}$ is $\log p(\theta)$ term and it's tractable , so M-step is tractable. //

Now, let's see $\underbrace{\prod_i p(x^i, \theta^t)}_{A(\theta^t)}$ monotonically increases as $t$ grows up.

at each it. step $t$, $\begin{cases} Q_i^t = p(z^i \mid x^i, \theta^{t-1}) \\ \theta^t \leftarrow \theta^{t-1} + \underbrace{\text{GD term}} \end{cases}$

$$\log A(\theta^{t+1}) \geq \sum_i \sum_{z^i} Q_i^{t+1}(z^i) \log \frac{p(x^i, z^i \mid \theta^{t+1})}{Q_i^{t+1}(z^i)} + \log p(\theta^{t+1})$$

because $\theta$ update increase this whole term // $\rightarrow \geq$ $\displaystyle \sum_i \sum_{z^i} Q_i^{t+1}(z^i) \log \frac{p(x^i, z^i \mid \theta^t)}{Q_i^{t+1}(z^i)} + \log p(\theta^t)$

$$= \log A(\theta^t)$$

$\therefore$ EM monotonically increases $\log A$.

#3.

(a) $x^{pr} = y^{pr} + z^{pr} + \varepsilon^{pr}$, where $\varepsilon^{pr} \sim \mathcal{N}(0, \sigma^2)$

(i) since $y^{pr}, z^{pr}, \varepsilon^{pr}$ is indep. $\Rightarrow x^{pr} \sim \mathcal{N}(\mu_p + \nu_r, \sigma_p^2 + \tau_r^2 + \sigma^2)$

$\therefore$ let $\begin{cases} m^{pr} \\ \Sigma^{pr} \end{cases}$ be $\begin{cases} \text{mean vec.} \\ \text{covar. mat.} \end{cases}$ of $(y^{pr}, z^{pr}, x^{pr})$

$\therefore m^{pr} = [\mu_p, \nu_r, \mu_p + \nu_r]$, $\Sigma^{pr} = \begin{bmatrix} \sigma_p^2 & 0 & \sigma_p^2 \\ 0 & \tau_r^2 & \tau_r^2 \\ \sigma_p^2 & \tau_r^2 & \sigma_p^2 + \tau_r^2 + \sigma^2 \end{bmatrix}$

$\Big( \because \mathrm{Cov}(A, A+B)$ ($A, B$ : indep var.)

$= E[(A - E(A))(A + B - E(A+B))] = E[(A-E(A))(A-E(A))] + E[A - E(A)]E[B - E(B)]$

$= \mathrm{Cov}(A, A) \Big)$

(ii) $1: (y^{pr}, z^{pr})$, $2: x^{pr}$

$$\Rightarrow \begin{cases} \mu_1 = [\mu_p, \nu_r], \quad \Sigma_{11} = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix}, \quad \Sigma_{12} = [\sigma_p, \tau_r^2]^T = \Sigma_{21}^T, \quad \Sigma_{22} = \sigma_p^2 + \tau_r^2 + \sigma^2 \\ \mu_2 = \mu_p + \nu_r, \quad x_2 = x^{pr} \end{cases}$$

$$\Rightarrow \begin{cases} \mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = \sim \\ \Sigma_{1|2} = \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \sim \end{cases}$$

$$\Rightarrow Q_{pr}(y^{pr}, z^{pr}) = \frac{1}{\sqrt{2\pi}\,|\Sigma_{1|2}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} y^{pr} \\ z^{pr}\end{bmatrix} - \mu_{1|2}\right)^T \Sigma_{1|2}^{-1}\left(\begin{bmatrix} y^{pr} \\ z^{pr}\end{bmatrix} - \mu_{1|2}\right)\right)$$

(b) At E-step, we calculated $w_{(y^{pr}, z^{pr})} = Q_{pr}(y^{pr}, z^{pr})$

Then lower bound is

$$l(\mu_p, \nu_r, \sigma_p^2, \tau_r^2) = \sum_p \sum_r \sum_{y^{pr}, z^{pr}} \underbrace{w_{(y^{pr}, z^{pr})}}_{\text{from (a)-ii}} \log \frac{\overbrace{p(y^{pr}, z^{pr}, x^{pr})}^{\text{from (a)-i}}}{w_{(y^{pr}, z^{pr})}}$$

$$= \sum_p \sum_r \sum_{y^{pr}, z^{pr}} w_{(y^{pr}, z^{pr})} \left(\log \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_{pr}|^{\frac{1}{2}}} - \frac{1}{2}(a^{pr} - m^{pr})^T \Sigma_{pr}^{-1}(a^{pr} - m^{pr}) - \log w_{(y^{pr}, z^{pr})}\right)$$

* Update $\mu_i$ :

$$\frac{\partial l}{\partial \mu_i} = \sum_r \sum_{y,z} w_{(y^{ir}, z^{ir})} \underbrace{\left(\frac{\partial m^{ir}}{\partial \mu_i}\right)^T}_{[1,0,1]} \Sigma_{pr}^{-1}(a^{ir} - m^{pr})$$

$$= \cdots \text{ (should calculate } \Sigma_{pr}^{-1}) = 0$$

$$\Rightarrow \boxed{\mu_i = \frac{\sum_r \sum_{y,z} w_{y^{ir}, z^{ir}} \left(\frac{y^{ir}}{\sigma_i^2} - \frac{2(x^{ir} - \nu_r)}{\sigma^2}\right)}{\sum_r \sum_{y,z} w_{(y^{ir}, z^{ir})} \left(\frac{1}{\sigma_i^2} - \frac{2}{\sigma^2}\right)}}$$

* Update $\sigma_i^2$ :

$$\frac{\partial l}{\partial \sigma_i^2} = -\frac{1}{2}\sum_r \sum_{y,z} w_{y^{ir}, z^{ir}} \left(\frac{1}{|\Sigma_{ir}|} \cdot \frac{\partial |\Sigma_{ir}|}{\partial \sigma_i^2} + (a^{ir} - m^{ir})^T \frac{\partial \Sigma_{ir}^{-1}}{\partial \sigma_i^2}(a^{ir} - m^{ir})\right)$$

$$= \cdots \text{ (should calculate } |\Sigma_{pr}| \text{ and } \Sigma_{pr}^{-1})$$

$$\Rightarrow \boxed{\sigma_i^2 = \frac{\sum_r \sum_{y,z} w_{y^{ir}, z^{ir}} (y^{ir} - \mu_i)^2}{\sum_r \sum_{y,z} w_{y^{ir}, z^{ir}}}}$$

* Update $\nu_j$ :

$$\frac{\partial l}{\partial \nu_j} = \text{(similar to } \frac{\partial l}{\partial \mu_i}\text{, } [0,1,1] \text{ instead of } [1,0,1]) \cdots = 0$$

$$\Rightarrow \boxed{\nu_j = \frac{\sum_p \sum_{y,z} w_{y^{pj}, z^{pj}} \left(\frac{z^{pj}}{\tau_j^2} - \frac{2(x^{pj} - \mu_p)}{\sigma^2}\right)}{\sum_p \sum_{y,z} w_{y^{pj}, z^{pj}} \left(\frac{1}{\tau_j^2} - \frac{2}{\sigma^2}\right)}}$$

* Update $\tau_j^2$:

$$\frac{\partial l}{\partial \tau_j^2} = \left(\text{similar to } \frac{\partial l}{\partial \sigma_j^2}\right) \cdots = 0$$

$$\Rightarrow \boxed{\tau_j^2 = \frac{\sum_p \sum_{y,z} w_{y^{pi},z^{pi}} (z_{pj} - \nu_j)^2}{\sum_p \sum_{y,z} w_{y^{pi},z^{pi}}}}$$

#4.

(a) $KL(P\|Q) = -\sum_x P(x) \log \frac{Q(x)}{P(x)} = E[-\log \frac{Q}{P}] \boxed{\geq} -\log(E[\frac{Q}{P}])$

$= -\log \sum_x Q(x) = -\log 1 = \boxed{0}$

from hint, equality implies to $\frac{Q}{P} = E[\frac{Q}{P}] = 1$, $P = Q$

Thus $KL(P\|Q) = 0 \Leftrightarrow P = Q$.

(b) $KL(P(Y|X) \| Q(Y|X)) + KL(P(X) \| P(Y))$

$= \sum_x P(x) \left( \log \frac{P(x)}{Q(x)} + \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} \right)$

$= \sum_x P(x) \left( \sum_y P(y|x) \log \frac{P(x,y)}{Q(x,y)} \right)$

$= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)}$

$= KL(P(X,Y) \| Q(X,Y))$

(c) $KL(\hat{P}\|P_\theta) = -\sum_x \hat{P}(x) \log \frac{P_\theta(x)}{\hat{P}(x)} = -\sum_x \frac{1}{m} \sum_{i=1}^m 1\{x^i = x\} \log \frac{P_\theta(x)}{\frac{1}{m}\sum_{i=1}^m 1\{x^i = x\}}$

$= -\frac{1}{m} \sum_{i=1}^m \log \frac{P_\theta(x^i)}{\frac{1}{m}\sum_{i=1}^m 1} = -\frac{1}{m} \sum_i \log P_\theta(x^i)$

$\therefore \sim$ //

$\boxed{\text{\#5. see ps 3-5.ipynb}}$