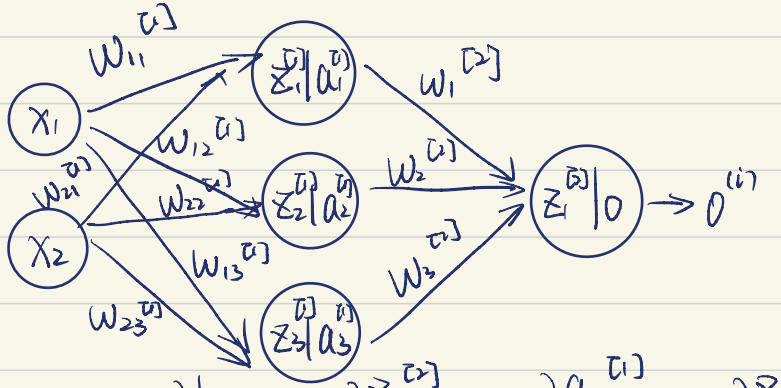
Question 1:

a.



Forward Propagation:

$$z_2^{[1]} = w_{12}^{[1]} x_1 + w_{22}^{[1]} x_2$$

$$a_2^{[1]} = \sigma(z_2^{[1]})$$

$$z_1^{[2]} = w_1^{[2]} a_1^{[1]} + w_2^{[2]} a_2^{[1]} + w_3^{[2]} a_3^{[1]}$$

$$o = \sigma(z_1^{[2]})$$

$$l = \frac{1}{m}\sum_{i=1}^{m}(o^{(i)} - y^{(i)})^2$$

$$\frac{\partial l}{\partial w_{12}^{[1]}} = \frac{\partial l}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial a_2^{[1]}} \cdot \frac{\partial a_2^{[1]}}{\partial z_2^{[1]}} \cdot \frac{\partial z_2^{[1]}}{\partial w_{12}^{[1]}}$$

$$\therefore \frac{\partial l}{\partial w^{[2]}} = \frac{\partial}{\partial w^{[2]}}\left(\frac{1}{m}\sum_{i=1}^{m}(o^{(i)} - y^{(i)})^2\right)$$

$$= \frac{\partial}{\partial w^{[2]}}\frac{1}{m}\sum_{i=1}^{m}\left(\sigma(w^{[2]}a^{[1]}) - y^{(i)}\right)^2 = \frac{1}{m}\sum_{i=1}^{m} 2o^{(i)} \cdot o^{(i)} \cdot (1-o^{(i)}) \cdot a^{[1]} - 2y^{(i)} o^{(i)}(1-o^{(i)}) \cdot a^{[1]}$$

$$= \frac{2}{m}\sum_{i=1}^{m} o^{(i)}(1-o^{(i)}) \cdot a^{[1]}(o^{(i)} - y^{(i)})$$

$$a^{[1]} = \frac{\partial z^{[2]}}{\partial w^{[2]}}$$

$$\therefore \frac{\partial L}{\partial z_1^{[2]}} = \frac{2}{m}\sum_{i=1}^{m} o^{(i)}(1-o^{(i)})(o^{(i)}-y^{(i)})$$

$$\therefore \frac{\partial L}{\partial w_{12}^{[1]}} = \frac{2}{m}\sum_{i=1}^{m}\left(o^{(i)}(1-o^{(i)})(o^{(i)}-y^{(i)})\right)\cdot w_2^{[2]}\cdot a_2^{[1]}(1-a_2^{[1]})\cdot x_1^{(i)}$$

(b) The decision boundary of "0": $x_1 = 0.25$, $x_2 = 0.75$, $x_1 + x_2 = 4.5$

$z_1^{[1]} = w_{01}^{[1]} + w_{11}^{[1]} x_1 + w_{21}^{[1]} x_2$

want: when $x_1 > 0.75$, $z_1^{[1]} < 0$

$0.75\, w_{11}^{[1]} + w_{01}^{[1]} = 0 \Rightarrow w_{11}^{[1]} = -1,\ w_{01}^{[1]} = 0.75$

same for $z_2^{[1]}$: $0.25 w_{22}^{[1]} + w_{02}^{[1]} = 0 \Rightarrow w_{22}^{[1]} = -1,\ w_{02}^{[1]} = 0.25$

$z_3^{[1]} = w_{03}^{[1]} + w_{13}^{[1]} x_1 + w_{23}^{[1]} x_2$

$\Rightarrow w_{03}^{[1]} = 4.5,\ w_{13}^{[1]} = 1,\ w_{23}^{[1]} = 1$

$$\Rightarrow W^{[1]} = \begin{pmatrix} 0.25 & -1 & 0 \\ 0.25 & 0 & -1 \\ -0.5 & 1 & 1 \end{pmatrix}$$

for output layer, if 0, then $W_{01}^{[2]} + 0 < 0 \Rightarrow W_{01}^{[2]} = -1$

$$\Rightarrow W^{[2]} = (-1 \quad 1 \quad 1 \quad 1)$$

(c) The problem is not linearly separable. Therefore, it can not make the loss 0.

Question 2:

$$MAP = \left(\prod_{i=1}^{m} \sum_{z^{(i)}} p(x^{(i)}, z^{(i)} | \theta) \right) p(\theta).$$

$$\Rightarrow l(\theta) = \log\left(\prod_{i=1}^{m} \sum_{z^{(i)}} P(x^{(i)}, z^{(i)} \mid \theta)\right) \cdot P(\theta)$$

$$= \log P(\theta) + \sum_{i=1}^{m} \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)} \mid \theta)$$

$$= \log P(\theta) + \sum_{i=1}^{m} \sum_{z^{(i)}} \log P(x^{(i)}, z^{(i)} \mid \theta).$$

E-step: $$l(\theta) = \log P(\theta) + \sum_{i=1}^{m} \sum_{z^{(i)}} \log Q_i(z^{(i)}) \cdot \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}$$

$$= \log P(\theta) + \sum_{i=1}^{m} \log E_{z^{(i)} \sim Q_i} \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})} \geq \log P(\theta) + \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i} \log \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}$$

$$= \log P(\theta) + \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \cdot \log \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}$$

want: $$\log E_{z^{(i)} \sim Q_i} \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})} = E_{z^{(i)} \sim Q_i} \log \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}$$

$$\Rightarrow \frac{P(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})} \text{ is constant.} \Rightarrow Q_i(z^{(i)}) = \frac{1}{\lambda} P(x^{(i)}, z^{(i)} \mid \theta).$$

$$\because \sum_{z^{(i)}} Q_i(z^{(i)}) = 1 \Rightarrow \lambda = \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta).$$

$$\Rightarrow Q_i(z^{(i)}) = \frac{P(x^{(i)}, z^{(i)}; \theta)}{\sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta)} = \frac{P(x^{(i)}, z^{(i)}; \theta)}{P(x^{(i)}; \theta)} = P(z^{(i)} | x^{(i)}; \theta)$$

M-step: $\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} + \log P(\theta)$.

Take the gradient of θ and set to 0.

Proof of convergence: see lecture notes

Question 3:

(a) i. $x^{pr} = y^{pr} + z^{pr} + \epsilon^{pr}$

$\because y^{pr}, z^{pr}, \epsilon^{pr}$ are all Gaussian

$\Rightarrow x^{pr}$ is Gaussian

$\Rightarrow$ the mean for (x^{pr}, y^{pr}, z^{pr}) is $(\mu_p + v_r, \mu_p, v_r)^T$

and the covariance is $\begin{pmatrix} \sigma_p^2+\sigma_r^2+\sigma^2 & \sigma_p^2 & \sigma_r^2 \\ \sigma_p^2 & \sigma_p^2 & 0 \\ \sigma_r^2 & 0 & \sigma_r^2 \end{pmatrix}$

ii. E-step: $\frac{p(x^{pr}, y^{pr}, z^{pr})}{Q_{pr}(y^{pr}, z^{pr})}$ is constant.

$\because \sum_{r=1}^{R}\sum_{p=1}^{P} Q_{pr}(y^{pr}, z^{pr}) = 1.$

$\therefore Q_{pr} = \frac{p(x^{pr}, y^{pr}, z^{pr})}{\sum_{r=1}^{R}\sum_{p=1}^{P} p(x^{pr}, y^{pr}, z^{pr})} = \frac{p(x^{pr}, y^{pr}, z^{pr})}{p(x^{pr})} = p(y^{pr}, z^{pr} \mid x^{pr})$

$\Rightarrow Q_{pr}$ is also a Gaussian.

$\mu = [\mu_p \cdot V_r] + \frac{x^{pr} - \mu_p - V_r}{\sigma_p^2+\sigma_r^2+\sigma^2}[\sigma_p^2 \sigma_r^2]$

$\Sigma = [\sigma_p^2 \; 0 \; 0 \; \sigma_r^2] - \frac{1}{\sigma_p^2+\sigma_r^2+\sigma^2}[\sigma_p^4 \; \sigma_p^2\sigma_r^2 \; \sigma_r^2\sigma_p^2 \; \sigma_r^4]$

(b) M-step: $\underset{\mu_p, \sigma_p, v_r, J_r}{\operatorname{argmax}} Q_{pr} \cdot \log \frac{P(x^{pr}, y^{pr}, z^{pr}; \mu_p, \sigma_p, v_r, J_r)}{Q_{pr}}$

use gradient descent, set all the gradient to 0.

Question 4:

(a) $KL(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = -\sum_x P(x) \log \frac{Q(x)}{P(x)} \geq -\log \sum_x P(x) \frac{Q(x)}{P(x)}$ (Jensen inequality)

$\geq -\log \sum_x Q(x) \geq -\log 1 \geq 0.$

If $P = Q$, then $KL(P\|Q) = \sum_x P(x) \log 1 = 0.$

If $KL(P\|Q) = 0$, then $-\sum_x P(x) \log \frac{Q(x)}{P(x)} = -\log \sum_x P(x) \cdot \frac{Q(x)}{P(x)}$

$\Rightarrow \frac{Q(x)}{P(x)}$ is constant $\Rightarrow P = Q.$

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

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

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

$\because \sum_y P(y|x) = 1$

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

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

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

(c) $KL(\hat{P} \| P_\theta) = \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)} = -\sum_x \hat{P}(x) \log \frac{P_\theta(x)}{\hat{P}(x)}$

$$= -\frac{1}{m} \sum_x \sum_{i=1}^{m} I\{x^{(i)} = x\} \cdot \log \frac{P(x;\theta)}{\frac{1}{m}\sum_{i=1}^{m} I\{x^{(i)} = x\}}$$

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

Thus, minimizing $KL(\hat{P} \| P_\theta)$ is equivalent to maximizing $\sum_{i=1}^{m} \log P_\theta(x^{(i)}) = l(\theta)$.