PS3

Problem I: (a).

(1) $z^{[1]} = W^{[1]} x^{(i)} + b^{[1]}$: $\cancel{z_{1,1}^{[1]} = W_{1,1}^{[1]} x_1^{(i)} + b_{0,1}^{[1]} \cdot z_{2,1}^{[1]} \cdot W_{2,1}^{[1]} x_2^{(i)} +}$

$1\times3 \quad 3\times2 \quad 2\times1 \quad 3\times1$

$z_1^{[1]} = W_{1,1}^{[1]} x_1^{(i)} + W_{2,1}^{[1]} x_2^{(i)} + b_{0,1}^{[1]}$

(2). $a^{[1]} = g(z^{[1]}) = \frac{1}{1+e^{-z^{[1]}}}$ $\quad a_1^{[1]} = \frac{1}{1+e^{-z_1^{[1]}}}$

(3). $z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$

$1\times1 \quad 1\times3 \quad 3\times1 \quad 1\times1$

(4). $a^{[2]} = g(z^{[2]}) = \frac{1}{1+e^{-z^{[2]}}}$

$\frac{\partial}{\partial z^{[2]}}$: $g(z^{[2]})(1-g(z^{[2]}))$

$\frac{\partial z^{[2]}}{\partial a_2^{[1]}}$: $W_2^{[2]}$

$\frac{\partial a_{(2)}^{[1]}}{\partial z_{(2)}^{[1]}}$: $g(z_2^{[1]})(1-g(z_2^{[1]}))$

$\frac{\partial z_{(2)}^{[1]}}{\partial W_{1,2}^{[1]}}$: $x_1^{(i)}$

$\frac{\partial o}{\partial W_{1,2}^{[1]}}$: $g(z^{[2]})(1-g(z^{[2]}))\, W_2^{[2]} \cdot g(z_2^{[1]})(1-g(z_2^{[1]})) \cdot x_1$

$W_{1,2}^{[1]} := W_{1,2}^{[1]} - \partial \frac{\partial o}{\partial W_{1,2}^{[1]}}$

(b). It is possible.

The dataset can be perfectly separated by a triangle. For ~~each~~ 1st hidden layer, it determins if the point is on the left side of $x_1 = 0.5$.

The 2nd hidden layer determines $x_2 \geq 0.5$, The third $x+y-4 \geq 0$.

If any of the above is true, then $y^{(i)} = 1$.

(c). It is not possible.

When combine all the neurons with linear activation function. The final result is linear

There is no way to separate the dataset perfectly with a single line in 2-dimension.

Problem 2

(a). $D_{KL} = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$.

$$-D_{KL} = -\sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} = \sum_{x \in X} P(x) \log \frac{Q(x)}{P(x)} \leq \log \sum_{x \in X} P(x) \cdot \frac{Q(x)}{P(x)}$$

$$\leq \log 1$$

$-D_{KL} \leq 0$

$D_{KL} \geq 0$.

$P = Q \Rightarrow D_{KL}(P \| Q) = 0$.

$D_{KL}(P \| Q) = \sum P(x) \log \frac{P(x)}{Q(x)} = \sum P(x) \cdot 0 = 0$.

$D_{KL}(P \| Q) = 0 \Rightarrow P == Q$.

if $D_{KL}(P \| Q) = 0$. then $D_{KL}(P \| Q) = \sum P(x) \log \frac{P(x)}{Q(x)} = 0 = -D_{KL} = \sum P(x) \log \frac{Q(x)}{P(x)}$

$\sum P(x) \log \frac{Q(x)}{P(x)} \leq \log \sum Q(x) = 0$

We know that $\log x$ is ~~concave~~ strictly concave, and $\sum P(x) \log \frac{Q(x)}{P(x)} = \log \sum P(x) \cdot \frac{Q(x)}{P(x)} = 0$.

Then $\frac{Q(x)}{P(x)}$ is a constant. Suppose $\frac{Q(x)}{P(x)} = c$.

Then $\sum P(x) \log \frac{P(x)}{Q(x)} = \sum P(x) \cdot \log c = \log(c) \cdot \sum P(x) = \log(c) = 0$.

Therefore $c = 1$, thus ~~P(x) and Q(x)~~ $P = Q$.

(b).

$$D_{KL}(P(X,Y)\|Q(X,Y)) = \sum_{\substack{x\in X\\ y\in Y}} P(x,y)\log\frac{P(x,y)}{Q(x,y)}$$

$$= \sum_{x\in X}\sum_{y\in Y} P(x)P(y|x)\log\frac{P(x)P(y|x)}{Q(x)Q(y|x)}$$

$$= \sum_{x\in X}\sum_{y\in Y} P(x)P(y|x)\left(\log\frac{P(x)}{Q(x)} + \log\frac{P(y|x)}{Q(y|x)}\right).$$

$$= \sum_{x\in X}\sum_{y\in Y} P(y|x)P(x)\log\frac{P(x)}{Q(x)} + \sum_{x\in X}\sum_{y\in Y} P(x)P(y|x)\log\frac{P(y|x)}{Q(y|x)}$$

$$= \sum_{y\in Y} P(y|x)\sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)} + \sum_{x\in X} P(x)\sum_{y\in Y} P(y|x)\log\frac{P(y|x)}{Q(y|x)}$$

$$= \sum_{y\in Y} P(y|x)\, D_{KL}(P(X)\|Q(X)) + D_{KL}(P(Y|X)\|Q(Y|X))$$

$$= D_{KL}(P(X)\|Q(X)) + D_{KL}(P(Y|X)\|Q(Y|X)).$$

Q.E.D.

c). $$\arg\min_\theta D_{KL}(\hat{P}\|P_\theta) = \arg\min_\theta \sum_i^m \hat{P}(x^{(i)})\log\frac{\hat{P}(x^{(i)})}{P_\theta(x^{(i)})}$$

since $\hat{P}$ is uniformly distributed, $\hat{P}(x^{(i)}) = \frac{1}{m}$, then the equation is.

$$\arg\min_\theta \sum_i^m \frac{1}{m}\log\frac{1}{mP_\theta(x^{(i)})}$$

$$= \arg\min_\theta \frac{1}{m}\sum_i^m -\log(mP_\theta(x^{(i)}))$$

$$= \arg\max_\theta \frac{1}{m}\sum_i^m \log m + \log P_\theta(x^{(i)})$$

$$= \arg\max_\theta \frac{1}{m}\sum_i^m \log m + \frac{1}{m}\sum_i^m \log P_\theta(x^{(i)})$$

$$= \arg\max_\theta \sum_i^m \log P_\theta(x^{(i)})$$ which is the maximum likelihood.

Problem 3

(a). $E_{y \sim P(y;\theta)} [\nabla_{\theta'} \log P(y;\theta') |_{\theta'=\theta}] \overset{df}{=} \int_{-\infty}^{\infty} P(y;\theta) \nabla_{\theta'} \log P(y;\theta') |_{\theta'=\theta} \, dy$

$$= \int_{-\infty}^{\infty} P(y;\theta) \cdot \frac{1}{P(y;\theta')|_{\theta'=\theta}} \cdot \nabla_{\theta'} P(y;\theta') |_{\theta'=\theta} \, dy$$

$$= \int_{-\infty}^{\infty} \nabla_{\theta'} P(y;\theta') |_{\theta'=\theta} \, dy$$

$$= \nabla_{\theta'} \left(\int_{-\infty}^{\infty} P(y;\theta') \, dy \right)$$

$$= \nabla_{\theta'} 1$$

$$= 0.$$

(b). $I(\theta) = \mathrm{Cov}_{y \sim p(y;\theta)} [\nabla_{\theta'} \log p(y;\theta') |_{\theta'=\theta}]$

$$= E_{y \sim p(y;\theta)} [\nabla_{\theta'} \log P(y;\theta') \nabla_{\theta'} \log P(y;\theta')^T |_{\theta'=\theta}] - (E_{y \sim P(y;\theta)} [\nabla_{\theta'} \log P(y;\theta') |_{\theta'=\theta}])^2$$

$$= E_{y \sim p(y;\theta)} [\nabla_{\theta'} \log P(y;\theta') \nabla_{\theta'} \log P(y;\theta')^T |_{\theta'=\theta}]$$

(c). $I(\theta) = E [\nabla_{\theta'} \log P(y) \nabla_{\theta'} \log P(y)^T |_{\theta'=\theta}]$.

For i, j entry of the corvariance matrix, $\left(\frac{1}{P(y;\theta')} \cdot \nabla_{\theta'_i} P(y;\theta')\right) \cdot \left(\frac{1}{P(y;\theta')} \cdot \nabla_{\theta'_j} P(y;\theta')\right) |_{\theta'=\theta}$

$$= \frac{1}{(P(y;\theta'))^2} \nabla_{\theta'_i} P(y;\theta') \nabla_{\theta'_j} P(y;\theta') |_{\theta'=\theta}.$$

For $E_{y \sim P(y;\theta)} [-\nabla^2_{\theta'} \log P(y;\theta') |_{\theta'=\theta}]$ the i, j entry of the ~~resu~~ result matrix is.

$$-\nabla_{\theta'_j} (\nabla_{\theta'_i} \log P(y;\theta') |_{\theta'=\theta}) = -\nabla_{\theta'_j} \left(\frac{1}{P(y;\theta')} \nabla_{\theta'_i} P(y;\theta') |_{\theta'=\theta} \right)$$

$$= -\left((-1) \frac{1}{P(y;\theta')^2} \cdot \nabla_{\theta'_j} P(y;\theta') \cdot \nabla_{\theta'_i} P(y;\theta') + \frac{1}{P(y;\theta')} \nabla_{\theta'_i} \nabla_{\theta'_j} P(y;\theta') \right) |_{\theta'=\theta}$$

$$= \underbrace{\frac{1}{P(y;\theta')^2} \cdot \nabla_{\theta'_i} P(y;\theta') \nabla_{\theta'_j} P(y;\theta')}_{A} - \underbrace{\frac{1}{P(y;\theta')} \nabla_{\theta'_i} \nabla_{\theta'_j} P(y;\theta') |_{\theta'=\theta}}_{B}$$

Prob 3. (c) continue.

We know $E_{y\sim p(y;\theta)}[A - B] = E_{y\sim p(y;\theta)}[A] - E_{y\sim p(y;\theta)}[B]$.

Also, $E_{y\sim p(y;\theta)}[B] = \int_{-\infty}^{\infty} p(y;\theta) \cdot \frac{1}{p(y;\theta)} \nabla_{\theta_i} \nabla_{\theta_j} p(y;\theta')\, dy$.

$$= \int_{-\infty}^{\infty} \nabla_{\theta_i} \nabla_{\theta_j} p(y;\theta')\, dy$$

$$= \nabla_{\theta_i} \nabla_{\theta_j} \int_{-\infty}^{\infty} p(y;\theta')\, dy$$

$$= \nabla_{\theta_i} \nabla_{\theta_j} \ 1$$

$$= 0.$$

Thus, the original i,j entry $= E_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta)^2} \nabla_{\theta_i} p(y;\theta) \nabla_{\theta_j} p(y;\theta)\right]$

$$= E_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta') \nabla_{\theta'} \log p(y;\theta')^T \Big|_{\theta'=\theta}\right]$$

$$= I(\theta').$$

(d). $D_{KL}(p_\theta \| p_{\theta+d}) = f(\hat{\theta}) \approx D_{KL}(p_\theta \| p_\theta) + (\hat{\theta}-\theta)^T \nabla_{\theta'} D_{KL}(p_\theta \| p_{\theta'})|_{\theta'=\theta} + \frac{1}{2}(\hat{\theta}-\theta)^T (\nabla_{\theta'}^2 D_{KL}(p_\theta \| p_{\theta'})|_{\theta'=\theta})(\hat{\theta}-\theta)$.

$$= 0 + d^T \nabla_{\theta'} p(y;\theta) \log \frac{p(y;\theta)}{p(y;\theta')}\Big|_{\theta'=\theta} + \frac{1}{2} d^T \nabla_{\theta'}^2 p(y;\theta) \log \frac{p(y;\theta)}{p(y;\theta')}\Big|_{\theta'=\theta}\, d$$

$$= d^T \int_{-\infty}^{\infty} \cancel{p(y;\theta)} \cdot \frac{\cancel{p(y;\theta')}}{\cancel{p(y;\theta)}} \cdot (-1) \frac{1}{\cancel{p(y;\theta')}^{\,p(y;\theta)}} \cdot \nabla_{\theta'} p(y;\theta')\Big|_{\theta'=\theta} + \frac{1}{2} d^T \nabla_{\theta'} \left(\cancel{p(y;\theta)} \cdot \frac{\cancel{p(y;\theta')}}{\cancel{p(y;\theta)}} (-1) \frac{1}{\cancel{p(y;\theta')}^{\,p(y;\theta)}} \nabla_{\theta'} p(y;\theta')\Big|_{\theta'=\theta}\right. d.$$

$$= \underbrace{d^T (-1) \int_{-\infty}^{\infty} \frac{\cancel{1}}{\cancel{p(y;\theta)}} \cdot \nabla_{\theta'} p(y;\theta')\Big|_{\theta'=\theta} dy}_{A} + \underbrace{\cdots\ dy.}_{B}$$

$$A = d^T (-1) \nabla_{\theta'} \int_{-\infty}^{\infty} p(y;\theta')\, dy \Big|_{\theta'=\theta}$$

$$= d^T (-1) \nabla_{\theta'} (1) \Big|_{\theta'=\theta}$$

$$= 0.$$

$$B = \frac{1}{2} d^T \int_{-\infty}^{\infty} \nabla_{\theta'} \left(\cancel{P(y;\theta)} \cdot \cancel{\frac{P(y;\theta')}{P(y;\theta)}} \cdot P(y;\theta) \cdot (-1) \cdot \frac{1}{P(y;\theta')} \cdot \nabla_{\theta'} P(y;\theta') \right)_{|\theta'=\theta} dy$$

$$= \frac{1}{2} d^T (-1) \int_{-\infty}^{\infty} \nabla_{\theta'} P(y;\theta) \cdot \frac{1}{P(y;\theta')} \nabla_{\theta'} P(y;\theta')_{|\theta'=\theta} \, dy.$$

$$= \frac{1}{2} d^T (-1) \int_{-\infty}^{\infty} P(y;\theta) \left((-1) \frac{1}{P(y;\theta')^2} \nabla_{\theta'} P(y;\theta') \cdot \nabla_{\theta'} P(y;\theta')^T + \frac{1}{P(y;\theta)} \nabla^2_{\theta'} P(y;\theta') \right)_{|\theta'=\theta} d \, dy \; d$$

$$= \frac{1}{2} d^T \Bigg[\underbrace{\int_{-\infty}^{\infty} P(y;\theta) \cdot \frac{1}{P(y;\theta)^2} \nabla_{\theta'} P(y;\theta') \nabla_{\theta'} P(y;\theta)^T_{|\theta'=\theta} \, dy}_{C} - \underbrace{\int_{-\infty}^{\infty} \nabla^2_{\theta'} P(y;\theta)_{|\theta'=\theta} \, dy}_{D} \Bigg] d$$

$$D = \nabla^2_{\theta'} \int_{-\infty}^{\infty} P(y;\theta') \, dy_{\,|\theta'=\theta} = \nabla^2_{\theta'} (1)_{|\theta'=\theta} = 0$$

$$C = \int_{-\infty}^{\infty} P(y;\theta) \cdot \frac{\nabla_{\theta'} P(y;\theta')}{P(y;\theta)} \cdot \frac{\nabla_{\theta'} P(y;\theta')^T}{P(y;\theta)} {}_{|\theta'=\theta} \, dy$$

$$= E_{y \sim P(y;\theta)} \left[\nabla_{\theta'} \log P(y;\theta') \nabla_{\theta'} \log P(y;\theta')^T_{|\theta'=\theta} \right] d$$

$$= I(\theta)$$

Thus. $D_{KL}(P_\theta \| P_{\theta+d}) = \frac{1}{2} d^T I(\theta) d$

Q.E.D.

(e). $\mathcal{L}(d,\lambda) = \ell(\theta+d) - \lambda\left[D_{KL}(p_\theta \| p_{\theta+d}) - c\right]$

$$= \ell(\theta) + d^T \nabla_{\theta'} \ell(\theta')|_{\theta'=\theta} - \lambda\left(\tfrac{1}{2} d^T I(\theta) d - c\right)$$

$$= \log P(y;\theta) + d^T \frac{\nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}}{P(y;\theta)} - \lambda\left(\tfrac{1}{2} d^T I(\theta) d - c\right)$$

$$\nabla_d \mathcal{L}(d,\lambda) = \frac{\nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}}{P(y;\theta)} - \lambda(I(\theta) d = 0$$

$$\lambda I(\theta) d = \frac{\nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}}{P(y;\theta)}$$

$$d = \frac{1}{\lambda} I^{-1}(\theta) \cdot \frac{\nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}}{P(y;\theta)}$$

$$\nabla_\lambda \mathcal{L}(d,\lambda) = c - \tfrac{1}{2} d^T I(\theta) d = 0$$

$$d^T I(\theta) d = 2c$$

$$\left(\frac{1}{\lambda} I^{-1}(\theta) \cdot \frac{\nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}}{P(y;\theta)}\right)^T I(\theta) \left(\frac{1}{\lambda} I^{-1}(\theta) \cdot \frac{\nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}}{P(y;\theta)}\right) = 2C$$

$$\frac{1}{\lambda^2 P(y;\theta)^2} \cdot \nabla_{\theta'} P(y;\theta')^T|_{\theta'=\theta}\, I^{-1}(\theta)\, I(\theta)\, I^{-1}(\theta)\, \nabla_{\theta'} P(y;\theta')|_{\theta'=\theta} = 2C$$

$$\lambda^2 = \frac{1}{2c} \cdot \frac{1}{P(y;\theta)^2} \nabla_{\theta'} P(y;\theta')^T|_{\theta'=\theta}\, I^{-1}(\theta)\, \nabla_{\theta'} P(y;\theta')|_{\theta'=\theta}$$

$$\lambda = \frac{1}{P(y;\theta)\sqrt{2c}} \sqrt{\nabla_{\theta'} P(y;\theta')^T I^{-1}(\theta) \nabla_{\theta'} P(y;\theta)}$$

$$\lambda = \frac{1}{\sqrt{2c}} \sqrt{\nabla_\theta \ell(\theta) I^{-1}(\theta) \nabla_\theta \ell(\theta)}$$

$$d^* = \sqrt{\frac{2C}{\nabla_\theta \ell(\theta) I^{-1}(\theta) \nabla_\theta \ell(\theta)}}\, I^{-1}(\theta) \nabla_\theta \ell(\theta) = \sqrt{\frac{2c}{\nabla_\theta P(y;\theta) I^{-1}(\theta) \nabla_\theta P(y;\theta)}}\, I^{-1}(\theta) \nabla_\theta P(y;\theta)$$

(f). ~~In Newton's method, the update is~~ ~~$-H^{-1}\nabla_\theta \ell(\theta)$.~~ ~~and we know from log likelyhood is concave. Then eventually~~ ~~$\ell(\theta)$~~ ~~will converge to the maximum. which is H is P.S.D. Since natural parameter also has a maximum that maximize the log~~ ~~$P(y;\theta)$~~ ~~and minimize~~ ~~D_{KL}.~~ ~~the update~~ ~~$\frac{1}{\lambda} I^{-1}(\theta)\nabla_\theta \ell(\theta) = \frac{1}{\lambda}(-I^{-1}(\theta))\nabla_\theta \ell(\theta)$.~~ ~~we know~~ ~~$\lambda$~~ ~~is positive, which does not contribute the direction of~~

For Generalized linear model, $\ell(\theta) = \arg\max_\theta P(y;\theta) = b(y)\exp(\theta^T T(y) - a(\theta))$

From problem set 1 question 4, we know $H_{ij} = \frac{\partial^2}{\partial \eta^2} a(\eta)$ where η is the natural parameter

In natural gradient, the parameter is θ. And,

$$I(\theta)_{ij} = \int_{-\infty}^{\infty} P(y;\theta) - \nabla_\theta^2 \log P(y;\theta)\, dy$$

$$= \int_{-\infty}^{\infty} P(y;\theta) - \nabla_\theta^2 \log\left(b(y)\exp(\theta^T T(y) - a(\theta))\right) dy.$$

$$= \int_{-\infty}^{\infty} P(y;\theta) - \nabla_\theta^2 \left(\log b(y) + \theta^T T(y) - a(\theta)\right) dy.$$

$$= \int_{-\infty}^{\infty} P(y;\theta) - \nabla_\theta \left(0 + T(y) - \nabla_\theta a(\theta)\right) dy.$$

$$= \int_{-\infty}^{\infty} P(y;\theta) + \nabla_\theta^2 a(\theta)\, dy.$$

$$= \nabla_\theta^2 a(\theta) \int_{-\infty}^{\infty} P(y;\theta)\, dy$$

$$= \nabla_\theta^2 a(\theta)$$ ~~which is~~ Thus, $I(\theta) = H$

Thus, $I(\theta)^{-1}\nabla_\theta \log P(y;\theta) = H^{-1}\nabla_\theta \log P(y;\theta)$ which indicates the same direction. since $\frac{1}{\lambda}$ is a positive scalar that does not contribute to the direction.

(f) Answer:

Newton's Method: ~~θ~~ ~~θ~~ $\theta := \theta - H^{-1}\nabla_\theta \ell(\theta)$.

Natural Gradient: $I(\theta) = E_{y\sim p(y;\theta)}[-\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]$

$$= E_{y\sim p(y;\theta)}[-\nabla^2_\theta \ell(\theta)]$$

$$= -E_{y\sim p(y;\theta)}[\nabla^2_\theta \ell(\theta)]$$

$$\theta := \theta + \tilde{d}$$

$$= \theta + \frac{1}{\lambda} I(\theta)^{-1} \nabla_\theta \ell(\theta).$$

$$= \theta + \frac{1}{\lambda} E_{y\sim p(y;\theta)}[H]^{-1} \nabla_\theta \ell(\theta).$$

14)-(a).

$$l_{semi\text{-}sup}(\theta^{(t+1)}) = l_{unsup}(\theta^{(t+1)}) + \alpha\, l_{sup}(\theta^{(t+1)})$$

expanding $$= \sum_{i=1}^{m} \cancel{\log} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \frac{P(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

Jensen's inequality $$\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} + \alpha \sum_{i=1}^{M} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

θ^{t+1} is the arg max $$\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} + \alpha \sum_{i=1}^{m} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{t})$$

$$= l_{unsup}(\theta) + \alpha\, l_{sup}(\theta)$$

$$= l_{semi\text{-}sup}(\theta).$$

(b). latent variables. : $W_j^{(i)}$ for $\forall i \in \{1, \dots m\}$ and $\forall j$ $j \in \mathbb{R}^n$

$$W_j^{(i)} := Q_i(z^{(i)} = j) = P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma).$$

$$= \frac{P(z^{(i)}, x^{(i)}; \phi, \mu, \Sigma)}{P(x^{(i)})}$$

$$= \frac{P(x^{(i)} \mid z^{(i)}) \cdot P(z^{(i)})}{\sum_{z^{(i)}} P(x^{(i)} \mid z^{(i)}) \cdot P(z^{(i)})}$$ with μ, Σ, ϕ

~~(c). latent variables~~ ~~parameters : $\mu^{(t+1)}, \Sigma^{(t+1)}, \phi^{(t+1)}$~~

~~$l_{semi\text{-}sup}(\theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \mu, \Sigma, \phi)}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \mu, \Sigma, \phi)$~~

$$\tilde{W}_j^{(i)} := 1\{\tilde{z}^{(i)} = j\} \text{ for } \forall i \in \{1, \dots, \tilde{m}\}\ \forall j \in \mathbb{R}^n$$

(c).

$$l_{semi\text{-}sup}(\theta) = l_{unsup}(\theta) + \alpha\, l_{sup}(\theta)$$

$$\theta^{(t+1)} := \arg\max_{\theta} \sum_{i=1}^{m}\left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i^{(t)}(z^{(i)})}\right) + \alpha\left(\sum_{i=1}^{\tilde{m}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)};\theta)\right)$$

where $Q_i^{(t)}(z^{(i)}) = P(z^{(i)}=j \mid x^{(i)}; \phi, \mu, \Sigma)$ and

$$P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma) = P(x^{(i)} \mid z^{(i)}=j; \mu, \Sigma)\cdot P(z^{(i)}=j;\phi)$$

$$= \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot \phi_j$$

$$\nabla_{\mu_l}\, \theta^{(t+1)} = \nabla_{\mu_l} \sum_{i=1}^{m}\sum_{z^{(i)}}^{k} -w_j^{(i)} \cdot \frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)$$

$$= -\frac{1}{2}\sum_{i=1}^{m} \nabla_{\mu_l} w_l^{(i)} \left(x^{(i)T}\Sigma_l^{-1}x^{(i)} - x^{(i)T}\Sigma_j^{-1}\mu_l - \mu_l \Sigma_j^{-1} x^{(i)} + \mu_l^T \Sigma_l^{-1}\mu_l\right)$$

$$= -\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\left(-x^{(i)T}\Sigma_l^{-1} - \Sigma_l^{-1}x^{(i)} + 2\Sigma_l^{-1}\mu_l\right)$$

$$= -\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\left(-2\Sigma_l^{-1}x^{(i)} + 2\Sigma_l^{-1}\mu_l\right)$$

$$= \sum_{i=1}^{m} w_l^{(i)}\left(\Sigma_l^{-1}x^{(i)} - \Sigma_l^{-1}\mu_l\right)$$

$$\cancel{\sum_{i=1}^{m} w_l^{(i)}\Sigma_l^{-1}x^{(i)} - \sum_{i=1}^{m} w_l^{(i)}\Sigma_l^{-1}\mu_l = 0}$$

$$\cancel{\Sigma_l^{-1}\sum_{i=1}^{m} w_l^{(i)}x^{(i)} - \Sigma_l^{-1}\sum_{i=1}^{m} w_l^{(i)}\mu_l = 0}$$

$$= \Sigma_l^{-1}\sum_{i=1}^{m} w_l^{(i)}\left(x^{(i)}-\mu_l\right)$$

$$\cancel{\mu_l \sum_{i=1}^{m} w_l^{(i)} = \sum_{i=1}^{m} w_l^{(i)} x^{(i)}}$$

$$\cancel{\mu_l = \frac{\sum_{i=1}^{m} w_l^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_l^{(i)}}}$$

(c). continued.

$$\nabla_{\mu_l} \alpha \sum_{i=1}^{\tilde{m}} \log \left(P(\tilde{x}^{(i)} | \tilde{z}^{(i)}; \mu_l, \Sigma_l) P(\tilde{z}^{(i)}; \phi \right)$$

$$= \alpha \sum_{i=1}^{\tilde{m}} \nabla_{\mu_l} \log \left(\frac{1}{(2\pi)^{n/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(\tilde{x}^{(i)} - \mu_l)^T \Sigma_l^{-1} (\tilde{x}^{(i)} - \mu_l)\right)\right) \cdot \prod_{j=1}^{K} \phi_j^{1\{\tilde{z}^{(i)}=1\}}$$

$$= \alpha \sum_{i=1}^{\tilde{m}} \nabla_{\mu_l} \left(-\frac{1}{2}(\tilde{x}^{(i)} - \mu_l)^T \Sigma_l^{-1} (\tilde{x}^{(i)} - \mu_l)\right)$$

$$= \alpha \sum_{i=1}^{\tilde{m}} (-1) \Sigma_l^{-1} (\tilde{x}^{(i)} - \mu_l) \cdot (-1)$$

$$= \alpha \sum_{i=1}^{\tilde{m}} \Sigma_l^{-1} (\tilde{x}^{(i)} - \mu_l) = \underset{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}{\alpha \sum_{i=1}^{\tilde{m}} \Sigma_l^{-1} \left(\tilde{x}^{(i)} 1\{\tilde{z}^{(i)}=l\} - \mu_l 1\{\tilde{z}^{(i)}=l\}\right)}$$

Combine the two terms. $\nabla_{\mu_l} l_{semi\text{-}sup}(\theta) = \sum_{i=1}^{m} \Sigma_l^{-1} w_l^{(i)} (x^{(i)} - \mu_l) + \alpha \sum_{i=1}^{\tilde{m}} \Sigma_l^{-1} (\tilde{x}^{(i)} - \mu_l) = 0$

$\tilde{x}^{(i)}$ belongs to μ_l

$$\Sigma_l^{-1} \left[\sum_{i=1}^{m} w_l^{(i)} (x^{(i)} - \mu_l) + \alpha \sum_{i=1}^{\tilde{m}} (\tilde{x}^{(i)} - \mu_l) \right] = 0.$$

$$\sum_{i=1}^{m} w_l^{(i)} x^{(i)} - \mu_l \sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \tilde{x}^{(i)} - \alpha \mu_l \cdot \text{count}(\tilde{x}^{(i)}) =$$

$$\mu_l \sum_{i=1}^{m} w_l^{(i)} + \alpha \mu_l \cdot \text{count}(\tilde{x}) = \sum_{i=1}^{m} w_l^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \tilde{x}^{(i)}$$

$$\mu_l = \frac{\sum_{i=1}^{m} w_l^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \tilde{x}^{(i)} 1\{\tilde{z}^{(i)}=l\}}{\sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=l\}}$$

$$\nabla_\phi \, l_{unsup}(\theta) = \nabla_\phi \sum_{i=1}^{m}\sum_{j=1}^{K} w_j^{(i)} \log \phi_j$$

$$\nabla_\phi \, \alpha \, l_{sup}(\theta) = \nabla_\phi \, \alpha \sum_{i=1}^{\tilde{m}} \log \prod_{j=1}^{K} \phi^{\tilde{z}_j^{(i)}} = \nabla_\phi \, \alpha \sum_{i=1}^{\tilde{m}}\sum_{j=1}^{K} \tilde{z}_j^{(i)} \log \phi_j$$

$$\nabla_\phi \left(\sum_{i=1}^{m}\sum_{j=1}^{K} w_j^{(i)} \log \phi_j + \alpha \sum_{i=1}^{\tilde{m}}\sum_{j=1}^{K} \tilde{z}_j^{(i)} \log \phi_j \right)$$

We know $\sum_{i=1}^{K} \phi_j = 1$ thus $\sum_{i=1}^{K} \phi_j - 1 = 0$.

$$\mathcal{L}(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{K} w_j^{(i)} \log \phi_j + \alpha \sum_{i=1}^{\tilde{m}}\sum_{j=1}^{K} \tilde{z}_j^{(i)} \log \phi_j + \beta\left(\sum_{i=1}^{K} \phi_j - 1\right).$$

$$\frac{\partial}{\partial \phi_j} \mathcal{L}(\phi) = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \alpha \frac{\sum_{i=1}^{\tilde{m}} \tilde{z}_j^{(i)}}{\phi_j} + \beta = 0.$$

$$= \frac{1}{\phi_j} \sum_{i=1}^{m} w_j^{(i)} + \frac{1}{\phi_j} \alpha \sum_{i=1}^{\tilde{m}} \tilde{z}_j^{(i)} + \beta = 0.$$

$$\phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \tilde{z}_j^{(i)}}{-\beta}.$$

We know $\sum_j \phi_j = \sum_j \frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \tilde{z}_j^{(i)}}{-\beta} = 1$

$$\sum_j \sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_j \sum_{i=1}^{\tilde{m}} \tilde{z}_j^{(i)} = -\beta$$

$$m + \alpha \tilde{m} = -\beta$$

Thus $$\phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \tilde{z}_j^{(i)}}{m + \alpha \tilde{m}}$$

$$\nabla_{\Sigma_\ell^{(t+1)}} \ell_{unsup}(\theta) = \nabla_{\Sigma_\ell} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

$$= \nabla_{\Sigma_\ell} \sum_{i=1}^{m} \sum_{j=1}^{k} W_j^{(i)} \log \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} + w_j^{(i)} \left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right))$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} \nabla_{\Sigma_\ell} w_j^{(i)} \log |\Sigma_j|^{-1/2} - \frac{1}{2} w_j^{(i)} \left((x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j) \right)$$

$$= \sum_{i=1}^{m} -\frac{1}{2} W_\ell^{(i)} \Sigma_\ell^{-1} + \sum_{i=1}^{m} -\frac{1}{2} (w_\ell^{(i)} (-\Sigma^{-1}) (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T) \Sigma^{-1}$$

$$= \sum_{i=1}^{m} -\frac{1}{2} W_\ell^{(i)} \Sigma_\ell^{-1} + \sum_{i=1}^{m} \frac{1}{2} W_\ell^{(i)} (\Sigma^{-1} (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T) \Sigma^{-1}$$

$$\nabla_{\Sigma_\ell} \ell_{sup}(\theta) = -\frac{\alpha \tilde{m}}{2} \Sigma^{-1} + \alpha \sum_{i=1}^{\tilde{m}} \frac{1}{2} \Sigma^{-1} (\tilde{x}^{(i)}-\mu_\ell)(\tilde{x}^{(i)}-\mu_\ell)^T \Sigma^{-1}$$

$$\sum_{i=1}^{m} -\frac{1}{2} W_\ell^{(i)} + \sum_{i=1}^{m} \frac{1}{2} W_\ell^{(i)} \Sigma^{-1} (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T - \frac{\alpha \tilde{m}}{2} + \alpha \sum_{i=1}^{\tilde{m}} \frac{1}{2} \Sigma^{-1} (\tilde{x}^{(i)}-\mu_\ell)(\tilde{x}^{(i)}-\mu_\ell)^T = 0$$

$$\Sigma^{-1} \left(\sum_{i=1}^{m} \frac{1}{2} W_\ell^{(i)} (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{m}} \frac{1}{2} (\tilde{x}^{(i)}-\mu_\ell)(\tilde{x}^{(i)}-\mu_\ell)^T \right) =$$

$$\sum_{i=1}^{m} \frac{1}{2} W_\ell^{(i)} + \frac{1}{2} \alpha \tilde{m}$$

$$\sum_{i=1}^{m} W_\ell^{(i)} (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{m}} (\tilde{x}^{(i)}-\mu_\ell)(\tilde{x}^{(i)}-\mu_\ell)^T = (\Sigma) \left(\sum_{i=1}^{m} W_\ell^{(i)} + \alpha \tilde{m} \right) I$$

$$\Sigma = \frac{\sum_{i=1}^{m} W_\ell^{(i)} (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{m}} (\tilde{x}^{(i)}-\mu_\ell)(\tilde{x}^{(i)}-\mu_\ell)^T}{\left(\sum_{i=1}^{m} W_\ell^{(i)} + (\alpha \tilde{m}) \right) I}$$

$$\Sigma_\ell = \left(\sum_{i=1}^{m} W_\ell^{(i)} (x^{(i)}-\mu_\ell)(x^{(i)}-\mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{m}} (\tilde{x}^{(i)}-\mu_\ell)(\tilde{x}^{(i)}-\mu_\ell)^T \right) \left[\left(\sum_{i=1}^{m} W_\ell^{(i)} + \alpha \tilde{m} \right) I \right]^{-1}$$

4f).

	unsupervised	semi-supervised
(i)	117, 119, 102	35, 34, 34.
(ii)	not stable (different clusters)	very stable.
(iii).	not good quality. (not Gaussian shape)	good quality (Gussian shape).

Problem 5

(b). $\frac{256 \times 256 \times 256}{16} \approx ?\ 1M$

$3 \times 8 = 24$ bit. to store colors.

4 bits for 16 colors.

$\frac{24}{4} = 6$

4f). unsupervised | semi-supervised

(i) 117, 119, 102 | 35, 34, 34.

(ii) not stuble (different clusters) | very stable.

(iii). not good quality. (not Gaussian shape) | good quality (Gussian shape).

$\boxed{\text{Problem 5}}$

(b). $\frac{256 \times 256 \times 256}{16} \approx ? \ 1M$

$3 \times 8 = 24$ bit. to store colors.

4 bits for 16 colors.

$\frac{24}{4} = 6$

$\boxed{\text{Problem I Rewrite}}$

$$h_1 = \sigma(x_1 w_{1,1}^{[1]} + x_2 W_{2,1}^{[1]} + W_{0,1}^{[1]})$$

$$h_2 = \sigma(x_1 W_{1,2}^{[1]} + x_2 W_{2,2}^{[1]} + W_{0,2}^{[1]})$$

$$h_3 = \sigma(x_1 W_{1,3}^{[1]} + x_2 W_{2,3}^{[1]} + W_{0,3}^{[1]})$$

$$O = \sigma(h_1 W_1^{[2]} + h_2 W_2^{[2]} + h_3 W_3^{[2]} + W_0^{[2]})$$

$$\frac{\partial l}{\partial W_{1,2}^{[1]}} = \frac{\partial l}{\partial O}\frac{\partial O}{\partial h_2}\frac{\partial h_2}{\partial W_{1,2}^{[1]}} = \frac{1}{m}\sum_{i=1}^{m} 2(O^{(i)} - y) \cdot O^{(i)} \cdot (1-O^{(i)}) \cdot W_2^{[2]} \cdot h_2 \cdot (1-h_2) \cdot x_1$$