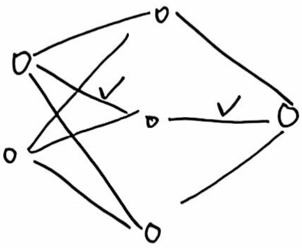
PS3.

1. a)



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

$$\cancel{\frac{\partial l}{\partial W_{12}^{[1]}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial o^{(i)}}{\partial W_{12}^{[1]}}}$$

$$= \sum_{i=1}^{m} \frac{\partial l}{\partial o^{(i)}} \cdot \cancel{\frac{\partial a_2^{[1]}}{\partial W_2^{[2]}}} \cdot \frac{\partial o^{(i)}}{\partial a_2^{[1]}} \cdot \frac{\partial a_2^{[1]}}{\partial z_2^{[1]}} \cdot \frac{\partial z_2^{[1]}}{\partial W_{12}^{[1]}}$$

$$\frac{\partial l}{\partial W_{1,2}^{[1]}} = \sum_{i=1}^{m} \frac{\partial l}{\partial o^{(i)}} \cdot \frac{\partial o^{(i)}}{\partial a_2^{[1]}} \cdot \frac{\partial a_2^{[1]}}{\partial z_2^{[1]}} \cdot \frac{\partial z_2^{[1]}}{\partial W_{1,2}^{[1]}}$$

$$= \sum_{i=1}^{m} \frac{1}{m} \cdot 2\left(o^{(i)} - y^{(i)}\right) \cdot W_2^{[2]} \cdot g'\left(z_2^{[1]}\right) \cdot x_1^{(i)}$$

$$= \sum_{i=1}^{m} \frac{1}{m} \cdot 2\left(o^{(i)} - y^{(i)}\right) W_2^{[2]} \cdot \frac{1}{1 + \cancel{e^{-(x_1^{(i)} W_{11}^{[1]} + W_{21}^{[1]} x_2^{(i)})}}\, e^{-(W_1^{[1]T} x^{(i)} + W_{01}^{[1]})}} \cdot \left(1 - \frac{1}{1 + e^{-(W_1^{[1]T} x^{(i)} + W_{01}^{[1]})}}\right) \cdot x_1^{(i)}$$

b). Yes.

Boundary: $x_1 \geq 0.5$
$x_2 \geq 0.5$
$x_1 + x_2 \leq 7$.

$$\Rightarrow \begin{aligned} x_1 + 0 \cdot x_2 - 0.5 &\geq 0 \\ 0 \cdot x_1 + x_2 - 0.5 &\geq 0 \\ -x_1 - x_2 + 7 &\geq 0 \end{aligned}$$

$W_{1,1}^{[1]} = 1 \quad W_{2,1}^{[1]} = 0 \quad W_{0,1}^{[1]} = -0.5$

$W_{1,2}^{[1]} = 0 \quad W_{2,2}^{[1]} = 1 \quad W_{0,2}^{[1]} = -0.5$

$W_{1,3}^{[1]} = -1 \quad W_{2,3}^{[1]} = -1 \quad W_{0,3}^{[1]} = 7$

o: h_1 & h_2 & h_3

$h_1 + h_2 + h_3 - 3 \geq 0$

$W_1^{[2]} = 1 \quad W_2^{[2]} = 1 \quad W_3^{[2]} = 1$

$W_0^{[2]} = -3$

(c) $h_1 = W_1^{[1]} x$

$h_2 = W_2^{[1]} x$

$h_3 = W_3^{[1]} x$

$$o = f\left(W^{[2]} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix}\right)$$

$$= f\left(W_0^{[2]} + W_1^{[2]} W_1^{[1]} x + W_2^{[2]} W_2^{[1]} x + W_3^{[2]} W_3^{[1]} x\right)$$

$$= f\left(\left(W_1^{[2]} W_1^{[1]} + W_2^{[2]} W_2^{[1]} + W_3^{[2]} W_3^{[1]}\right) x + W_0^{[2]}\right)$$

So there is ~~just one boundary.~~ one boundary.

2. a)

$\forall P, Q \quad D_{KL}(P \| Q) \geq 0$

Proof:

$$D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$$= \sum_x P(x)\left(-\log \frac{Q(x)}{P(x)}\right)$$

Since $-\log t$ is convex

$$\therefore D_{KL}(P\|Q) \geq -\log\left(\sum_x P(x) \frac{Q(x)}{P(x)}\right)$$

$$= -\log \sum_x Q(x) = 0$$

If we want $D_{KL}(P\|Q) = 0$, according to Jense's inequality,

$$\frac{Q(x)}{P(x)} = c.$$

$$Q(x) = c\,P(x)$$

$$1 = \sum Q(x) = \sum_x c\,P(x) = c$$

$$\therefore Q(x) = P(x)$$

b) $\cancel{D_{KL}(P(X|Y))}$ ||

left: $D_{KL}(P(X,Y)\|Q(X,Y))$

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

right: $D_{KL}(P(X)\|Q(X)) + D_{KL}(P(Y|X)\|Q(Y|X))$

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

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

left − right:

$$\sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)} - \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$$- \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)} \cdot \frac{Q(x)}{P(x)}$$

$$= \sum_{x,y} P(x,y) \log \left(\frac{\cancel{P(x,y)}}{\cancel{Q(x,y)}} \cdot \frac{\cancel{Q(x,y)}}{\cancel{P(x,y)}} \cdot \frac{P(x)}{Q(x)} \right)$$

$$- \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

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

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

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} - \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$= 0$

$\therefore$ left = right

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

(c) $\arg\min_\theta D_{KL}(\hat{P}\|P_\theta)$

$$= \arg\min_\theta \sum_{x\in Val(x)} \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)}$$

$$= \arg\min_\theta \sum_{x\in Val(x)} \hat{P}(x) \left(-\log \frac{P_\theta(x)}{\hat{P}(x)} \right)$$

$$= \arg\max_\theta \sum_{x\in Val(x)} \hat{P}(x) \log \frac{P_\theta(x)}{\hat{P}(x)}.$$

2

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

$$= \arg\max_\theta \left(\sum_{i=1}^m \log P_\theta(x^{(i)}) \right) - \underbrace{\sum_{i=1}^m \log \hat{P}(x^{(i)})}_{\text{constant}}$$

$$= \arg\max_\theta \sum_{i=1}^m \left(\log P_\theta(x^{(i)}) - \log \hat{P}(x^{(i)}) \right)$$

$$= \arg\max_\theta \sum_{x\in Val(x)} \sum_{x^{(i)}=x} \left(\log P_\theta(x^{(i)}) - \log \hat{P}(x^{(i)}) \right)$$

given $x^{(i)}=x$, $\log P_\theta(x^{(i)})$ and $\log \hat{P}(x^{(i)})$ are constant.

$$= \arg\max_\theta \sum_{x\in Val(x)} \left(\sum_{i=1}^m 1\{x^{(i)}=x\} \cdot \left(\log P_\theta(x) - \log \hat{P}(x) \right) \right)$$

$$= \arg\max_\theta \frac{1}{m} \sum_{x\in Val(x)} \left(\sum_{i=1}^m 1\{x^{(i)}=x\} \cdot \left(\log P_\theta(x) - \log \hat{P}(x) \right) \right)$$

$$= \arg\max_\theta \sum_{x\in Val(x)} \frac{1}{m} \sum_{i=1}^m 1\{x^{(i)}=x\} \cdot \left(\log P_\theta(x) - \log \hat{P}(x) \right)$$

$$= \arg\max_\theta \sum_{x\in Val(x)} \hat{P}(x) \left(\log \frac{P_\theta(x)}{\hat{P}(x)} \right)$$

$$= \arg\min D_{KL}(\hat{P}\|P_\theta).$$

3. a)

$$E_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')|\theta'=\theta]=\vec{0}$$

i.d. $\int_{-\infty}^{+\infty} p(y;\theta)\,\nabla_{\theta'}\log p(y;\theta')\,dy=\vec{0}$

i.d. $\forall\theta_i$ $\int_{-\infty}^{+\infty} p(y;\theta)\frac{\partial}{\partial\theta_i}\log p(y;\theta)\,dy=0.$

$$\int_{-\infty}^{+\infty} p(y;\theta)\frac{\partial}{\partial\theta_i}\log p(y;\theta)\,dy$$

$$=\int_{-\infty}^{+\infty} p(y;\theta)\frac{1}{p(y;\theta)}\frac{\partial p(y;\theta)}{\partial\theta_i}\,dy.$$

$$=\int_{-\infty}^{+\infty}\frac{\partial p(y;\theta)}{\partial\theta_i}\,dy.$$

$$=\frac{\partial\int_{-\infty}^{+\infty}p(y;\theta)\,dy}{\partial\theta_i}$$

$$=\frac{\partial 1}{\partial\theta_i}=0$$

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

$$=E\Big((\nabla_{\theta'}\log p(y;\theta')-E(\nabla_{\theta'}\log p(y;\theta')))$$

$$\cdot(\nabla_{\theta'}\log p(y;\theta')-E(\nabla_{\theta'}\log p(y;\theta')))^T\Big|\theta'=\theta\Big)$$

$$=E\left(\nabla_{\theta'}\log p(y;\theta')\nabla_{\theta'}\log p(y;\theta')^T\middle|\theta'=\theta\right)$$

c) $\cancel{\frac{\partial}{\partial\theta_i}\log p(y;\theta')}$

$$(\nabla_{\theta'}\log p(y;\theta'))_i=\frac{\partial}{\partial\theta_i}\log p(y;\theta')$$

$$=\frac{1}{p(y;\theta')}\cdot\frac{\partial p(y;\theta')}{\partial\theta_i}$$

$$H_{i,j}=\frac{\partial}{\partial\theta_j}\left(\frac{1}{p(y;\theta')}\cdot\frac{\partial p(y;\theta')}{\partial\theta_i}\right).$$

$$=(-1)p(y;\theta')^{-2}\cdot\frac{\partial p(y;\theta')}{\partial\theta_j}\cdot\frac{\partial p(y;\theta')}{\partial\theta_i}$$

$$+\,p(y;\theta')^{-1}\frac{\partial^2 p(y;\theta')}{\partial\theta_j\partial\theta_i}$$

$$E_{y\sim p(y;\theta)}[-H_{i,j}|\theta'=\theta]$$

$$=\int_{-\infty}^{+\infty}p(y;\theta')\Big[p(y;\theta')^{-2}\cdot\frac{\partial p(y;\theta')}{\partial\theta_j}\cdot\frac{\partial p(y;\theta')}{\partial\theta_i}-p(y;\theta')^{-1}\frac{\partial^2 p(y;\theta')}{\partial\theta_j\partial\theta_i}\Big].$$

$dy.$

Since.

$$\int_{-\infty}^{+\infty}\frac{\partial^2 p(y;\theta')}{\partial\theta_j\partial\theta_i}\,dy.$$

$$=\frac{\partial^2\int_{-\infty}^{+\infty}p(y;\theta')\,dy}{\partial\theta_j\,\partial\theta_i}$$

$$=0$$

$$\therefore E_{y\sim p(y;\theta)}[-H_{ij}|\theta'=\theta].$$

$$=\int_{-\infty}^{+\infty}p(y;\theta')\left(p(y;\theta')^{-2}\frac{\partial p(y;\theta')}{\partial\theta_j}\cdot\frac{\partial p(y;\theta')}{\partial\theta_i}\right)$$

$dy.$

$$=E\left((\nabla_{\theta'}\log p(y;\theta'))_i(\nabla_{\theta'}\log p(y;\theta'))_j\middle|\theta'=\theta'\right)=I_{ij}(\theta)$$

~~$\therefore E_{y\sim p(y;\theta)}[-H]$~~

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

(d) let $\tilde{\theta}=\theta+d$

Tylor expansion at θ_0:

$$D_{KL}(P_\theta\|P_{\tilde\theta})\approx D_{KL}(P_\theta\|P_{\tilde\theta})|_{\tilde\theta=\theta_0}$$

$$+(\tilde\theta-\theta_0)^T\nabla_{\tilde\theta}D_{KL}(P_\theta\|P_{\tilde\theta})|_{\tilde\theta=\theta_0}$$

$$+\frac{1}{2}(\tilde\theta-\theta_0)^T\nabla^2_{\tilde\theta}D_{KL}(P_\theta\|P_{\tilde\theta})|_{\tilde\theta=\theta_0}$$

$$\cdot(\tilde\theta-\theta_0)^T$$

(Holds ~~for~~ when $\tilde\theta$ close to θ_0)

$$\frac{\partial}{\partial \tilde\theta_i} D_{KL}(P_\theta \| P_{\tilde\theta}) = \frac{\partial}{\partial \tilde\theta_i} \sum_x P_\theta(x) \log \frac{P_\theta(x)}{P_{\tilde\theta}(x)}$$

$$= \frac{\partial}{\partial \tilde\theta_i} \sum_x P_\theta(x) \left(\log P_\theta(x) - \log P_{\tilde\theta}(x) \right)$$

$$= \sum_x P_\theta(x) \left(- \frac{1}{P_{\tilde\theta}(x)} \right) \cdot \frac{\partial P_{\tilde\theta}(x)}{\partial \tilde\theta_i}$$

$$\therefore \nabla_{\tilde\theta} D_{KL}(P_\theta \| P_{\tilde\theta}) = \sum_x P_\theta(x) \left(- \frac{1}{P_{\tilde\theta}(x)} \right) \nabla_{\tilde\theta} P_{\tilde\theta}(x)$$

$$\frac{\partial^2}{\partial \tilde\theta_i \partial \tilde\theta_j} D_{KL}(P_\theta \| P_{\tilde\theta}) = \frac{\partial}{\partial \tilde\theta_i} \sum_x P_\theta(x) \left(- \frac{1}{P_{\tilde\theta}(x)} \right) \cdot \frac{\partial P_{\tilde\theta}(x)}{\partial \tilde\theta_j}$$

$$= \sum_x P_\theta(x) \left[(-1)(P_{\tilde\theta}(x))^{-2} \frac{\partial P_{\tilde\theta}(x)}{\partial \tilde\theta_i} \frac{\partial P_{\tilde\theta}(x)}{\partial \tilde\theta_j} + \left(-\frac{1}{P_{\tilde\theta}(x)} \right) \frac{\partial^2 P_{\tilde\theta}(x)}{\partial \tilde\theta_i \partial \tilde\theta_j} \right].$$

$$= \sum_x P_\theta(x) \left(- (P_{\tilde\theta}(x))^{-2} \frac{\partial P_{\tilde\theta}(x)}{\partial \tilde\theta_i} \frac{\partial P_{\tilde\theta}(x)}{\partial \tilde\theta_j} \right)$$

$$- \sum_x P_\theta(x) \frac{1}{P_{\tilde\theta}(x)} \frac{\partial^2 P_{\tilde\theta}(x)}{\partial \tilde\theta_i \partial \tilde\theta_j}$$

let $\theta_0 = \theta$

$$\therefore \nabla_{\tilde\theta} D_{KL}(P_\theta \| P_{\tilde\theta}) \Big|_{\tilde\theta = \theta} = \sum_x P_\theta(x) \left(- \frac{1}{P_\theta(x)} \right) \nabla_\theta P_\theta(x)$$

$$\frac{\partial}{\partial \tilde\theta_i} D_{KL}(P_\theta \| P_{\tilde\theta}) = \frac{\partial}{\partial \tilde\theta_i} \int_{-\infty}^{+\infty} P_\theta(x) \log \frac{P_\theta(x)}{P_{\tilde\theta}(x)} dx$$

$$= \int_{-\infty}^{+\infty} P_\theta(x) \frac{\partial}{\partial \tilde\theta_i} \left(- \log P_{\tilde\theta}(x) \right) dx$$

$$= - \int_{-\infty}^{+\infty} P_\theta(x) \frac{1}{P_{\tilde\theta}(x)} \frac{\partial}{\partial \tilde\theta_i} P_{\tilde\theta}(x) \, dx$$

$$\nabla_{\tilde\theta} D_{KL}(P_\theta \| P_{\tilde\theta}) = - \int_{-\infty}^{+\infty} P_\theta(x) \frac{1}{P_{\tilde\theta}(x)} \nabla_{\tilde\theta} P_{\tilde\theta}(x) \, dx.$$

$$\frac{\partial^2}{\partial \tilde\theta_i \partial \tilde\theta_j} D_{KL}(P_\theta \| P_{\tilde\theta}) = \frac{\partial}{\partial \tilde\theta_i} \cdot \left(- \int_{-\infty}^{+\infty} P_\theta(x) \frac{1}{P_{\tilde\theta}(x)} \frac{\partial}{\partial \tilde\theta_i} P_{\tilde\theta}(x) \, dx \right)$$

$$\nabla^2_{\tilde\theta} D_{KL} = - \int_{-\infty}^{+\infty} P_\theta(x) \nabla^2_{\tilde\theta} \log P_{\tilde\theta}(x) \, dx. = I(\tilde\theta)$$

$$\therefore D_{KL}(P_\theta \| P_{\tilde\theta}) \approx D_{KL}(P_\theta \| P_{\tilde\theta}) \Big|_{\tilde\theta = \theta_0}$$

$$+ (\tilde\theta - \theta_0)^T \nabla_{\tilde\theta} D_{KL}(P_\theta \| P_{\tilde\theta}) \Big|_{\tilde\theta = \theta_0}$$

$$+ \frac{1}{2} (\tilde\theta - \theta_0)^T \nabla^2_{\tilde\theta} D_{KL}(P_\theta \| P_{\tilde\theta}) (\tilde\theta - \theta_0) \Big|_{\tilde\theta = \theta_0}$$

let $\theta_0 = \theta$,

$$\therefore D_{KL}(P_\theta \| P_{\theta + d})$$

$$\approx D_{KL}(P_\theta \| P_\theta) + d^T \left(- \int_{-\infty}^{+\infty} P_\theta(x) \nabla_\theta \log P_\theta(x) \right)$$

$$+ \frac{1}{2} d^T I(\theta) \, d$$

$$= \frac{1}{2} d^T I(\theta) \, d.$$

(e) ~~$d^* = \arg\max_d l(\theta + d)$ s.t. $D_{KL}(P_\theta \| P_{\theta+d}) = c$~~

~~$l(\theta + d) \approx l(\theta) + d^T \frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y)$~~

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

$$\mathcal{L}(d, \lambda) = - d^T \frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y) + \lambda \left(\frac{1}{2} d^T I(\theta) d - c \right)$$

~~Since $d^T \frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y)$ is convex,~~

~~$\frac{1}{2} d^T I(\theta) d - c$ is convex.~~

~~$\max_d \min_\lambda \mathcal{L}(d, \lambda) = \min_\lambda \max_d \mathcal{L}(d, \lambda)$~~

~~$\min_d \max_\lambda \mathcal{L}(d, \lambda) = \max_\lambda \min_d \mathcal{L}(d, \lambda)$~~

~~and d^* in dual problem is the same as that we required.~~

$$\nabla_d \mathcal{L}(d, \lambda) = - \frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y) + \lambda I(\theta) d = 0$$

According to KKT, there must be some d satisfy this formula.

$$d^* = \frac{1}{\lambda P_\theta(y)} I^{-1}(\theta) \nabla_\theta P_\theta(y)$$

Assume $I(\theta)$ is invertable Matrix. Hessian is invertible.

Plug back into $\mathcal{L}(d, \lambda)$:

~~$\mathcal{L}(d, \lambda) = - \frac{1}{\lambda P_\theta(y)} I^{-1}(\theta) \nabla_\theta P_\theta(y)$~~

$$\mathcal{L}(d, \lambda) = d^T \left(- \frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y) + \frac{1}{2} \lambda I(\theta) d \right) - \lambda c$$

$$= d^T \left(- \frac{1}{2} \frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y) \right) - \lambda c$$

$$= \left(\frac{1}{\lambda P_\theta(y)} \nabla_\theta P_\theta^T(y) I^{-1}(\theta)\right)\left(-\frac{1}{2}\frac{1}{P_\theta(y)} \nabla_\theta P_\theta(y)\right) - \lambda C$$

Since $I(\theta) = E_{y \sim P_\theta(y)}\left(-\nabla_{\theta'}^2 \log P_{\theta'}(y) \mid \theta' = \theta\right)$

$$I^T(\theta) = I(\theta)$$
$$I^{-T}(\theta) = I^{-1}(\theta)$$

$$\therefore \mathcal{L}(d,\lambda) = -\frac{1}{2\lambda}\left(\frac{1}{P_\theta^2(y)}\right) \nabla_\theta P_\theta^T(y) I^{-1}(\theta) \nabla_\theta P_\theta(y) - \lambda C$$

$$\frac{\partial \mathcal{L}(d,\lambda)}{\partial \lambda} = \frac{1}{2}\left(\frac{1}{P_\theta^2(y)}\right) \nabla_\theta P_\theta^T(y) I^{-1}(\theta) \nabla_\theta P_\theta(y) \cdot \lambda^{-2} - C = 0$$

$$\therefore \lambda = \sqrt{\frac{1}{2C P_\theta^2(y)} \nabla_\theta P_\theta^T(y) I^{-1}(\theta) \nabla_\theta P_\theta(y)}$$

$$\therefore d^* = \frac{1}{\lambda P_\theta(y)} I^{-1}(\theta) \nabla_\theta P_\theta(y)$$

$$= \sqrt{\frac{2C}{\nabla_\theta P_\theta^T(y) I^{-1}(\theta) \nabla_\theta P_\theta(y)}} \; I^{-1}(\theta) \nabla_\theta P_\theta(y)$$

(f) Since $\sqrt{\frac{2C}{\nabla_\theta P_\theta^T(y) I^{-1}(\theta) \nabla_\theta P_\theta(y)}}$ is a scalar.

$I^{-1}(\theta) \nabla_\theta P_\theta(y)$ determine the direction of d^*

$$I^{-1}(\theta) = E_{y \sim P_\theta(y)}(-H^{-1})$$

In Newton's method.

$$\theta' = \theta - \alpha H \nabla_\theta \underbrace{h_\theta(y)}_{P_\theta(y)}$$

4. (a)

$$l_{semi\text{-}sup}(\theta^{(t)}, Q^{(t)}) = \sum_{i=1}^{m} \sum_{z} Q^{(t)}(z) \log \frac{P(x^{(i)}, z; \theta^{(t)})}{Q^{(t)}(z)} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})$$

Since.

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

$$\therefore l_{semi\text{-}sup}(\theta^{(t+1)}, Q^{(t)}) \geq l_{semi\text{-}sup}(\theta^{(t)}, Q^{(t)})$$

$$l_{semi\text{-}sup}(\theta^{(t+1)}, Q^{(t+1)}) = \sum_{i=1}^{m} \sum_{z} Q^{(t+1)}(z) \log \frac{P(x^{(i)}, z; \theta^{(t+1)})}{Q^{(t+1)}(z)} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

~~Since~~

According to Jense's inequality.

$$\sum_z Q(z) \log \frac{P(x^{(i)}, z; \theta^{(t+1)})}{Q(z)} \leq \sum_z \log Q(z) \frac{P(x^{(i)}, z; \theta^{(t+1)})}{Q(z)}$$

if and only if $Q(z) = P(z \mid x^{(i)}; \theta^{(t+1)})$ we have the equality.

and $Q^{(t+1)}(z) = P(z \mid x^{(i)}; \theta^{(t+1)})$

$$\therefore \sum_z Q_{(i)}^{(t+1)}(z) \log \frac{P(x^{(i)}, z; \theta^{(t+1)})}{Q_{(i)}^{(t+1)}(z)} \geq \sum_z Q_{(i)}^{(t)}(z) \log \frac{P(x^{(i)}, z; \theta^{(t+1)})}{Q_{(i)}^{(t)}(z)}$$

$$\therefore l_{semi\text{-}sup}(\theta^{(t+1)}, Q^{(t+1)}) \geq l_{semi\text{-}sup}(\theta^{(t+1)}, Q^{(t)}) \geq l_{semi\text{-}sup}(\theta^{(t)}, Q^{(t)})$$

So the $l_{semi\text{-}sup}(\theta)$ monotonically increases with each iteration of E and M step.

b) ~~$l(\theta) = \sum_{i=1}^{m} \log Q_{(i)}(z) \frac{P}{Q_{(i)}(z)}$~~

$$l(\theta) = \sum_{i=1}^{m} \log \sum_{z} Q_{(i)}(z) \frac{P(x^{(i)}, z; \theta)}{Q_{(i)}(z)} + \alpha \sum_{i=1}^{\tilde{m}} \log P(x^{(i)}, \tilde{z}^{(i)}; \theta)$$

$$\geq \sum_{i=1}^{m} \sum_{z} Q_{(i)}(z) \log \frac{P(x^{(i)}, z; \theta)}{Q_{(i)}(z)} + \alpha \sum_{i=1}^{\tilde{m}} \log P(x^{(i)}, \tilde{z}^{(i)}; \theta)$$

Since we want to get equation by adjusting Q

So, we need $Q_{(i)}(z^{(j)}) = P(z^{(j)} | x^{(i)}; \theta)$

$$= \frac{P(x^{(i)} | z^{(j)}; \theta) P(z^{(j)}; \phi)}{\sum_{l=1}^{k} P(x^{(i)} | z^{(l)}; \theta) P(z^{(l)}; \phi)}.$$

$$= \frac{\frac{1}{\sqrt{(2\pi)^k} |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \phi_j}{\sum_{l=1}^{k} \frac{1}{\sqrt{(2\pi)^k} |\Sigma_l|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\right) \phi_l}$$

$$= \frac{|\Sigma_j|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \phi_j}{\sum_{l=1}^{k} |\Sigma_l|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\right) \phi_l}$$

(c) $\text{ELBO}(x; Q^{(t)}, \theta^{(t)})$

$$= \sum_{i=1}^{m} \sum_{z} Q^{(t)}_{(i)}(z) \log \frac{P(x^{(i)}, z; \theta)}{Q^{(t)}_{(i)}(z; \phi)} + \sum_{i=1}^{\tilde{m}} \log P(x^{(i)}, \tilde{z}^{(i)}; \theta)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} Q^{(t)}_{(i)}(z_j) \log \frac{P(x^{(i)} | z_j; \theta) P(z_j; \phi)}{Q^{(t)}_{(i)}(z_j)} + \sum_{i=1}^{\tilde{m}} \log P(x^{(i)} | \tilde{z}^{(i)}; \theta) P(\tilde{z}^{(i)}; \phi).$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} Q^{(t)}_{(i)}(z_j) \log \frac{\frac{1}{\sqrt{(2\pi)^k |\Sigma_j|}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)) \phi_j}{Q^{(t)}_{(i)}(z_j)}$$

$$+ \alpha \sum_{i=1}^{\tilde{m}} \log \prod_{j=1}^{k} \left(\frac{1}{\sqrt{(2\pi)^k |\Sigma_k|}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_k) \Sigma_k^{-1} (x^{(i)} - \mu_k)\right) \phi_k \right)^{1\{\tilde{z}^{(i)} = k\}}$$

$$\text{ELBO}(Q, \theta) = \sum_{i=1}^{m} \sum_{l=1}^{k} Q_{(i)}(z_l) \left(-\frac{1}{2} \log |\Sigma_l| - \frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) + \log \phi_l\right) + \alpha \sum_{i=1}^{\tilde{m}} \sum_{l=1}^{k} 1\{\tilde{z}^{(i)} = l\} \left(-\frac{1}{2} \log |\Sigma_l| - \frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) + \log \phi_l\right) + C.$$

let $m' = m + \tilde{m}$

$M' = M \| \tilde{M}$

$$\therefore \text{ELBO}(Q, \theta) = \sum_{i=1}^{m'} \sum_{l=1}^{k} \beta_l^{(i)} \left(-\frac{1}{2} \log |\Sigma_l| - \frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) + \log \phi_l\right) + C$$

where $\beta_l^{(i)} = \begin{cases} Q_{(i)}(z_l) & i \leq m \\ \alpha 1\{\tilde{z}^{(i)} = l\} & i > m \end{cases}$

In every M step, we don't modify Q, so we have.

$$l(\theta) = \sum_{i=1}^{m'} \sum_{l} \beta_l^{(i)} \left(-\frac{1}{2} \log |\Sigma_l| - \frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) + \log \phi_l\right) + C$$

$$\frac{\partial l(\theta)}{\partial \phi_l} = \sum_{i=1}^{m'} \left(\beta_l^{(i)} \frac{1}{\phi_l} - \beta_k^{(i)} \frac{1}{\phi_k}\right)$$

where $\phi_k = 1 - \sum_{l=1}^{k-1} \phi_l$

let $\frac{\partial l(\theta)}{\partial \phi_l} = 0$ for $l = \{1, 2, \ldots, k-1\}$

$\forall_{l = \{1, 2, \ldots, k-1\}} \; \frac{\partial l(\theta)}{\partial \phi_l} = 0$

$$\sum_{l=1}^{k} \phi_l = 1.$$

let $\varepsilon_l = \sum_{i=1}^{m'} \beta_l^{(i)}$ which are constants, regarding to ϕ.

$$\forall_{l \in \{1, 2, \ldots, k-1\}} \; \varepsilon_l \frac{1}{\phi_l} - \varepsilon_k \frac{1}{\phi_k} = 0.$$

$$\therefore \phi_l = \frac{\varepsilon_l}{\varepsilon_k} \phi_k$$

$$\therefore \sum^{k} \phi_l = \sum \frac{\varepsilon_l}{\varepsilon_k} \phi_k = 1$$

$$\phi_k = \left(\sum_{l=1}^{k} \frac{\varepsilon_l}{\varepsilon_k}\right)^{-1}$$

Since $\sum_{l=1}^{k} \varepsilon_l = \sum_{l=1}^{k}\sum_{i=1}^{m'} \beta_l^{(i)}$

$$= \sum_{l=1}^{k}\sum_{i=1}^{m} \beta_l^{(i)} + \sum_{l=1}^{k}\sum_{i=m+1}^{m'} \beta_l^{(i)}$$

$$= \sum_{l=1}^{k}\sum_{i=1}^{m} Q_{(i)}(z_l) + \sum_{l=1}^{k}\sum_{i=1}^{\tilde m} \alpha 1\{\tilde z^{(i)}=l\}$$

$$= m + \alpha \tilde m$$

$$\therefore \phi_k = \frac{\varepsilon_k}{m+\alpha\tilde m}.$$

$$\therefore \phi_l = \frac{\varepsilon_l}{\varepsilon_k}\cdot\frac{\varepsilon_k}{m+\alpha\tilde m} = \frac{\varepsilon_l}{m+\alpha\tilde m}$$

$$\therefore \forall l\in\{1,\dots,k\}\quad \phi_l = \frac{\varepsilon_l}{m+\alpha\tilde m} = \boxed{\frac{\sum_{i=1}^{m} Q^{(i)}(z_l)+\sum_{i=1}^{\tilde m}\alpha 1\{\tilde z^{(i)}=l\}}{m+\alpha\tilde m}}$$

$$\cancel{\nabla_{\mu_l} l(\theta) = \sum_{i=1}^{m'} \beta_l^{(i)}\left(\Sigma_l^{-1}(x^{(i)}-\mu_l)\right) = 0}.$$

$$\nabla_{\mu_l} l(\theta) = \sum_{i=1}^{m'} \beta_l^{(i)}\left(\Sigma_l^{-1}(x^{(i)}-\mu_l)\right) = 0.$$

$$\cancel{\beta_l^{(i)}\Sigma_l^{-1}\left(\sum_{i=1}^{m'}(x^{(i)}-\mu_l)\right) = 0}.$$

~~where Σ_l^{-1} is full rank, so~~

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

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

$$\sum_{i=1}^{m'} \beta_l^{(i)}\Sigma_l^{-1} x^{(i)} = \sum_{i=1}^{m'}\beta_l^{(i)}\Sigma_l^{-1}\mu_l.$$

$$\mu_l = \frac{\sum_{i=1}^{m'}\beta_l^{(i)} x^{(i)}}{\sum_{i=1}^{m'}\beta_l^{(i)}}$$

$$= \frac{\sum_{i=1}^{m} Q_{(i)}(z_l)x^{(i)} + \alpha\sum_{i=1}^{\tilde m} 1\{\tilde z^{(i)}=l\}\tilde x^{(i)}}{\sum_{i=1}^{m} Q_{(i)}(z_l) + \alpha\sum_{i=1}^{\tilde m} 1\{\tilde z^{(i)}=l\}}.$$

According to note 8.

$$\Sigma_l = \frac{\sum_{i=1}^{m'}\beta_l^{(i)}(x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T}{\sum_{i=1}^{m'}\beta_l^{(i)}}.$$

$$= \boxed{\frac{\sum_{i=1}^{m} Q_{(i)}(z_l)(x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T + \alpha\sum_{i=1}^{\tilde m} 1\{\tilde z^{(i)}=l\}(\tilde x^{(i)}-\mu_l)(\tilde x^{(i)}-\mu_l)^T}{\sum_{i=1}^{m} Q_{(i)}(z_l) + \alpha\sum_{i=1}^{\tilde m} 1\{\tilde z^{(i)}=l\}}}$$

(f) i) Semi-supervised EM takes fewer iteration to converge.

ii) Semi-supervised EM is more stable.

iii) Semi-supervised is more accurate, and have higher quality.

5. b) We have $2^8\cdot2^8\cdot2^8 = 2^{24}$ colors before, and now we have $16 = 2^4$ colors, ~~so we have~~

The storage for a single rgb grain reduce from 24 bits to 4 bits, so we have compressed the image by around $\frac{24}{4} = 6$ times.