1. (a)

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

$o^{(i)} = \sigma(w_1^{[2]} h_1^{(i)} + w_2^{[2]} h_2^{(i)} + w_3^{[2]} h_3^{(i)})$

$$\frac{\partial L}{\partial w_{1,2}^{[1]}} = \sum_{i=1}^{m} \frac{\partial L}{\partial o^{(i)}} \frac{\partial o^{(i)}}{\partial h_2^{(i)}} \frac{\partial h_2^{(i)}}{\partial w_{1,2}^{[1]}}$$

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

so $w_{1,2}^{[1]} := w_{1,2}^{[1]} - \alpha \frac{2}{m} w_2^{[2]} \sum_{i=1}^{m} (o^{(i)} - y^{(i)}) o^{(i)}(1-o^{(i)}) h_2^{(i)}(1-h_2^{(i)}) x_1^{(i)}$

where $h_2^{(i)} = \sigma(w_{1,2}^{[1]} x_1^{(i)} + w_{2,2}^{[1]} x_2^{(i)})$

(b)

Yes, it's possible. All the data of class 0 is inside a triangle, and all the examples outside the triangle is 1. The hidden layer consists of 3 neurals, each of which can be seemed as if the point is on the right side of the triangle. The 3 sides is approximately $x_1 = 0.5$, $x_2 = 0.5$, $x_1 + x_2 = 4$

(c) No, if $f(x) = x$,

$\vec{h} = W^{[1]} \vec{x}$ is a linear transformation, but the dataset is not linear seperable.

2.(a)

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

$$= E_{x\sim P(x)}\left[\log\frac{P(x)}{Q(x)}\right]$$

$$= E_{x\sim P(x)}\left[-\log\frac{Q(x)}{P(x)}\right]$$

$$\geq -\log\left[E_{x\sim P(x)}\frac{Q(x)}{P(x)}\right]$$

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

$$= -\log\sum_{x\in X} Q(x) = -\log 1 = 0$$

if and only if $\log\frac{P(x)}{Q(x)}$ is const,

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

which implies $\frac{P(x)}{Q(x)} = C$

because $\sum_{x\in X} P(x) = \sum_{x\in X} Q(x) = 1$

so it implies $P(x) = Q(x)$

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

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

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

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

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

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

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

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

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

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

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

3. (a)

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

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

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

$$= \nabla_\theta \int_{-\infty}^{+\infty} p(y;\theta)\,dy = \nabla_\theta 1 = 0$$

(b) $\mathrm{Cov}(X) = E[(X-\bar{X})(X-\bar{X})^T]$

so $I(\theta) = E_{y\sim p(y;\theta)}[(\nabla_\theta \log p(y;\theta) - E(\nabla_\theta \log p(y;\theta)))$

$(\nabla_\theta \log p(y;\theta) - E(\nabla_\theta \log p(y;\theta)))^T]$

$= E_{y\sim p(y;\theta)}[\nabla_\theta \log p(y;\theta)\,\nabla_\theta \log p(y;\theta)^T]$

(c) $I(\theta)_{(i,j)} = E_{y\sim p(y;\theta)}[\nabla_\theta \log p(y;\theta)_i \nabla_\theta \log p(y;\theta)_j]$

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

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

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

$$= \int_{-\infty}^{+\infty} p(y;\theta)\left(\frac{1}{p(y;\theta)}\frac{\partial p}{\partial \theta_i}\frac{\partial p}{\partial \theta_j} - \frac{\partial^2 p(y;\theta)}{\partial \theta_i \partial \theta_j}\frac{1}{p(y;\theta)}\right)dy$$

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

$$= I(\theta)_{ij}$$

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

(d) $\log p(y;\tilde{\theta}) = \log p(y;\theta)$

$+d^T \nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}$

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

$E_{y\sim p(y;\theta)}[\log p(y;\hat{\theta})] = E_{y\sim p(y;\theta)}[\log p(y;\theta)]$

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

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

so $E_{y\sim p(y;\theta)}[\log p(y;\theta)] - E_{y\sim p(y;\theta)}[\log p(y;\hat{\theta})]$

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

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

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

$d^* = \arg\max_d d^T \nabla_{\theta'} l(\theta')|_{\theta'=\theta}$

$\mathcal{L}(d,\lambda) = d^T \nabla_{\theta'} l(\theta')|_{\theta'=\theta} - \lambda\left(\frac{1}{2} d^T I(\theta)\,d - c\right)$

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

$d = \frac{1}{\lambda} I^{-1}(\theta)\,\nabla_{\theta'} l(\theta')|_{\theta'=\theta}$

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

$2\lambda^2 c = \nabla_{\theta'} l(\theta')|_{\theta'=\theta}^T I^{-1}(\theta)\,\nabla_{\theta'} l(\theta')|_{\theta'=\theta}$

$\lambda = \sqrt{\frac{\nabla_{\theta'} l(\theta')|_{\theta'=\theta}^T I^{-1}(\theta)\,\nabla_{\theta'} l(\theta')|_{\theta'=\theta}}{2c}}$

so $d^* = \sqrt{\frac{2c}{\nabla_\theta P_\theta^T I^{-1}(\theta)\,\nabla_\theta P_\theta}}\, I^{-1}(\theta)\,\nabla P_\theta$

(f) Newton's method: $\theta := \theta - H^{-1}\nabla_\theta l(\theta)$

Natural gradient:

$\theta = \theta + \frac{1}{\lambda} I^{-1}(\theta)\,\nabla_\theta l(\theta) = \theta - \frac{1}{\lambda}\epsilon(H^{-1})\,\nabla_\theta l(\theta))$

4. (a)

$$\ell_{semi-sup}(\theta^{(t+1)}) \geq \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i^{(t)} \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}} + \alpha \sum_{i=1}^{m} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

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

$$= \ell_{semi-sup}(\theta^{(t+1)})$$

(b) $w_j^{(i)} = P(z=j \mid x) = \dfrac{P(x \mid z=j) P(z=j)}{p(x)}$

$$= \frac{P(x \mid z=j) p(z=j)}{\sum_{l=1}^{k} P(x \mid z=l) P(z=l)}$$

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

(c) $ELBO(x^{(i)}, Q^{(i)}, \theta) =$

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{k} w_j \log \frac{P(x^{(i)} \mid z^{(i)}=j) P(z^{(i)}=j)}{w_j}\right)$$

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

$$= \sum_{i=1}^{m}\left(\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{|\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)\right) \phi_j}{w_j^{(i)}}\right)$$

$$+ \alpha \sum_{i=1}^{\tilde{m}} \log \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_{\tilde{z}^{(i)}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})^T \Sigma_{\tilde{z}^{(i)}}^{-1}(\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})\right)$$

$$\nabla_{\mu_j} ELBO_{un} = \nabla_{\mu_j} \sum_{i=1}^{m} -\frac{1}{2} w_j^{(i)} (x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)$$

$$= \sum_{i=1}^{m} w_j^{(i)} \Sigma_j^{-1}(x^{(i)}-\mu_j)$$

$$\nabla_{\mu_j} ELBO_{sup} = \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\} \Sigma_j^{-1}(\tilde{x}^{(i)}-\mu_j)$$

$$\nabla_{\mu_j} ELBO$$

$$= \Sigma_j^{-1}\left(\left(\sum_{i=1}^{m} w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\} \tilde{x}^{(i)}\right) - \mu_j\left(\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\right)\right)$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\} \tilde{x}^{(i)}}{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}}$$

$$\nabla_{\Sigma_j} ELBO_{un}$$

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

$$= \sum_{i=1}^{m} w_j^{(i)}\left(-\frac{1}{2}\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T \Sigma_j^{-1}\right)$$

$$\nabla_{\Sigma_j} ELBO_{sup}$$

$$= \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\left(-\frac{1}{2}\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T \Sigma_j^{-1}\right)$$

$$\nabla_{\Sigma_j} ELBO =$$

$$-\frac{1}{2}\left(\sum_{i=1}^{m} w_j^{(i)} \Sigma_j^{-1} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\} \Sigma_j^{-1}\right)$$

$$+ \frac{1}{2}\left(\sum_{i=1}^{m} w_j \Sigma_j^{-1}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T \Sigma_j^{-1} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\} \Sigma_j^{-1}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T \Sigma_j^{-1}\right)$$

$$= 0$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}(\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}}$$

$$\nabla_{\phi_j} ELBO_{un} = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j}$$

$$\nabla_{\phi_j} ELBO_{sup} = \sum_{i=1}^{\tilde{m}} 1\{z^{(i)} = j\} \frac{1}{\phi_j}$$

$$\sum_{j=1}^{k} \phi_j = 1$$

so $\mathcal{L}(\phi, \lambda) = ELBO - \lambda\left(\sum_{j=1}^{k} \phi_j - 1\right)$

$$\nabla \mathcal{L}_{\phi_j} = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \alpha \sum_{i=1}^{\tilde{m}} 1\{z^{(i)} = j\} \frac{1}{\phi_j} - \lambda = 0$$

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

$$\sum_{j=1}^{k} \phi_j = 1$$

$$\lambda = \sum_{j=1}^{k} \left(\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{z^{(i)} = j\} \right)$$

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

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

(f)

(i) for semi-supervised EM, it takes less time to converge

(ii) semi-supervised EM is more stable. For a unsupervised EM, it's possible to get different results with different initialization, but for semi-supervised EM, it always generate the same results

(iii) Semi-supervised Em has higher quality because it has lower variance for 3 distribution and higher variance for another distribution

5. (b)

$$\text{factor} = \frac{\log_2 16}{24} = \frac{4}{24} = \frac{1}{6}$$