(1.a)

$$let\ h_1 = [W_{11}, W_{21}],\ h_2 = [W_{12}, W_{22}],\ h_3 = [W_{13}, W_{23}]\ h_0 = [W_1, W_2, W_3]$$
$$x^{(i)} = [x_1^{(i)}, x_2^{(i)}]$$

forward:

$$let\ z_1^{(i)} = x_{(i)} h_1^T$$
$$z_2^{(i)} = x_{(i)} h_2^T$$
$$z_3^{(i)} = x_{(i)} h_3^T$$
$$a^{(i)} = g(z^{(i)})$$
$$a'^{(i)} = a^{(i)} h_0^T$$
$$o^{(i)} = g(a'^{(i)})$$
$$l = \frac{1}{m} \sum_{i=1}^{m} (o^{(i)} - y^{(i)})^2$$

backward:

$$\nabla_o l = \frac{2}{m} \sum_{i=1}^{m} o^{(i)}$$
$$\nabla_{a'} o = g(a'^{(i)})(1 - g(a'^{(i)}))$$
$$\nabla_{a_[2]} a' = h_0[2] = W_2'$$
$$\nabla_{z_2^{(i)}} a_{[2]} = g(z_2^{(i)})(1 - g(z_2^{(i)}))$$
$$\nabla_{W_{12}} = x_1^{(i)}$$
$$\nabla_{W_{12}} L = \frac{2}{m} \cdot \sum_{i=1}^{m} \cdot o^{(i)} \cdot g(a'^{(i)})(1 - g(a'^{(i)})) \cdot W_2' \cdot g(z_2^{(i)})(1 - g(z_2^{(i)})) \cdot x_1^{(i)}$$

(1.b)

The network can find three line to separate two types of point.

(1.c)

$$
\begin{aligned}
&W_{11}^{[2]}(W_{11}^{[1]}X_1 + W_{21}^{[1]}X_2 + W_{01}^{[1]}) \\
&+ W_{21}^{[2]}(W_{12}^{[1]}X_1 + W_{22}^{[1]}X_2 + W_{02}^{[1]}) \\
&+ W_{31}^{[2]}(W_{13}^{[1]}X_1 + W_{23}^{[1]}X_2 + W_{03}^{[1]}) + W_{01}^{[2]} \\
&= (W_{11}^{[2]}W_{11}^{[1]} + W_{21}^{[2]}W_{12}^{[1]} + W_{31}^{[2]}W_{13}^{[1]})X_1 \\
&+ (W_{11}^{[2]}W_{21}^{[1]} + W_{21}^{[2]}W_{22}^{[1]} + W_{31}^{[2]}W_{23}^{[1]})X_2 \\
&+ W_{11}^{[2]}W_{01}^{[1]} + W_{21}^{[2]}W_{02}^{[1]} + W_{31}^{[2]}W_{03}^{[1]} + W_{01}^{[2]} \\
&= aX_1 + bX_2 + C
\end{aligned}
$$

The network can only find a single line to separate two types of point, so it cannot reach 100% accuracy.

(2.a)

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

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

assume $f(x) = \frac{Q(x)}{P(x)}$ then

$$\sum_x P(x)\log\frac{Q(x)}{P(x)} \le \log\sum_x P(x)\frac{Q(x)}{P(x)} \le \log\sum_x Q(x) = 0$$

Because $-D_{\mathrm{KL}} \le 0$ then $D_{\mathrm{KL}} \ge 0$

(2.b)

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

$$
\begin{aligned}
rightside &= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)(\sum_y P(y \mid x)\log\frac{P(y \mid x)}{Q(y \mid x)}) \\
&= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)(\sum_y \frac{P(x,y)}{P(x)}\log\frac{P(x,y)}{P(x)}/\frac{Q(x,y)}{Q(x)}) \\
&= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)(\sum_y \frac{P(x,y)}{P(x)}\log\frac{P(x,y)}{Q(x,y)} * \frac{Q(x)}{P(x)}) \\
&= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)(\sum_y \frac{P(x,y)}{P(x)}\log\frac{P(x,y)}{Q(x,y)} + \sum_y \frac{P(x,y)}{P(x)}\log\frac{Q(x)}{P(x)}) \\
&= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)\log\frac{Q(x)}{P(x)} + \sum_x\sum_y \frac{P(x,y)}{P(x)}\log\frac{P(x,y)}{Q(x,y)} \\
&= leftside
\end{aligned}
$$

(2.c)

$$-D_{\mathrm{KL}} = \sum_x \hat{P} \log \frac{\hat{P}}{P_\theta} = \sum_x \hat{P} \log \hat{P} - \sum_x \hat{P} \log P_\theta$$

$$\arg\min_\theta D_{\mathrm{KL}}\left(\hat{P} \| P_\theta\right) = \arg\min_\theta (-\sum_x \hat{P} \log P_\theta)$$

$$= \arg\max_\theta \sum_x \hat{P} \log P_\theta$$

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

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

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

(3.a)

$$\mathbb{E}_{y \sim p(y)}[g(y)] = \int_{-\infty}^{\infty} p(y) g(y) dy$$

$$\mathbb{E}_{y \sim p(y;\theta)}\left[\nabla_{\theta'} \log p\left(y; \theta'\right)\big|_{\theta'=\theta}\right] = \int_{-\infty}^{\infty} p(y;\theta)(\nabla_{\theta'} \log p\left(y;\theta'\right) \ |_{\theta'=\theta}) dy$$

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

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

$$= \nabla_{\theta'} \int_{-\infty}^{\infty} \log p(y;\theta') \ |_{\theta'=\theta} \, dy = 0$$

(3.b)

$$\mathcal{I}(\theta) = \mathrm{Cov}_{y \sim p(y;\theta)}\left[\nabla_{\theta'} \log p\left(y;\theta'\right)\big|_{\theta'=\theta}\right]$$

$$= \mathbb{E}_{y \sim p(y;\theta)}\left[(\nabla_{\theta'} \log p(y;\theta')) - \mathbb{E}[\nabla_{\theta'} \log p(y;\theta')])(\nabla_{\theta'} \log p(y;\theta')) - \mathbb{E}[\nabla_{\theta'} \log p(y;\theta')])^\top\big|_{\theta'=\theta}\right]$$

$$= \mathbb{E}_{y \sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta') \nabla_{\theta'} \log p(y;\theta')^\top\big|_{\theta'=\theta}\right]$$

(3.c)

$$
\begin{aligned}
\mathcal{I}(\theta) &= \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta')\nabla_{\theta'} \log p(y;\theta')^\top\big|_{\theta'=\theta}\right] \\
&= \int p(y;\theta')\nabla_{\theta'} \log p(y;\theta')\nabla_{\theta'} \log p(y;\theta')^\top dy \\
&= \int p(y;\theta')(p(y;\theta')^{-1}\nabla_{\theta'} p(y;\theta'))(\nabla_{\theta'} p(y;\theta')^\top p(y;\theta')^{-\top})dy \\
&= \int \frac{\nabla_{\theta'} p(y;\theta') \cdot \nabla_{\theta'} p(y;\theta')^\top}{p(y;\theta')^\top} dy \\
&= \int p(y;\theta')\frac{\nabla_{\theta'} p(y;\theta') \cdot \nabla_{\theta'} p(y;\theta')^\top}{p(y;\theta')p(y;\theta')^\top} dy
\end{aligned}
$$

and because:

$$
\begin{aligned}
\nabla^2_{\theta'} \log p(y;\theta') &= \nabla_{\theta'} \frac{\nabla_{\theta'} p(y;\theta')}{p(y;\theta')} \\
&= \nabla_{\theta'}\left[\frac{1}{p(y;\theta')}\left[\nabla_{\theta'} p(y;\theta')\right]\right] \\
&= -\frac{\nabla_{\theta'} p(y;\theta')}{p(y;\theta')^2} \cdot \nabla_{\theta'} p(y;\theta') + \frac{\nabla^2_{\theta'} p(y;\theta')}{p(y;\theta')}
\end{aligned}
$$

so:

$$
\begin{aligned}
\int p(y;\theta')\nabla^2_{\theta'} \log p(y;\theta')dy &= \int -p(y;\theta')\frac{\nabla_{\theta'} p(y;\theta') \cdot \nabla_{\theta'} p(y;\theta')^\top}{p(y;\theta')^2} dy + \int \nabla^2_{\theta'} \log p(y;\theta')dy \\
&= \int -p(y;\theta')\frac{\nabla_{\theta'} p(y;\theta') \cdot \nabla_{\theta'} p(y;\theta')^\top}{p(y;\theta')p(y;\theta')^\top} dy
\end{aligned}
$$

then:

$$
\mathcal{I}(\theta) = \mathbb{E}_{y\sim p(y;\theta)}[-\nabla^2_{\theta'} p(y;\theta')|_{\theta'=\theta}]
$$

(3.d)

$$
\begin{aligned}
D_{\mathrm{KL}} &= P_\theta \log \frac{P_\theta}{P_{\theta+d}} = P_\theta \log P_\theta - P_\theta \log P_{\theta+d} \\
&= P_\theta \log P_\theta - \left[P_\theta \log P_\theta + d^\top \nabla_\theta (P_\theta \log P_\theta) + \frac{1}{2} d^\top \nabla_\theta^2 (P_\theta \log P_\theta) d \right] \\
&= -d^\top \nabla_\theta (P_\theta \log P_\theta) - \frac{1}{2} d^\top \nabla_\theta^2 (P_\theta \log P_\theta) d \\
&= -d^\top \mathbb{E}[\nabla_\theta \log P_\theta] + \frac{1}{2} d^\top \mathbb{E}[-\nabla_\theta^2 \log P_\theta] d \\
&= \frac{1}{2} d^\top \mathbb{E}[-\nabla_\theta^2 \log P_\theta] d \\
&= \frac{1}{2} d^\top \mathcal{I}(\theta) d
\end{aligned}
$$

(3.e)

$$
\mathcal{L}(d, \lambda) = \log p(y; \theta) + d^\top \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta} - \lambda [\frac{1}{2} d^\top \mathcal{I}(\theta) d - c]
$$

for d, we have:

$$
\begin{aligned}
\nabla_d \mathcal{L}(d, \lambda) &= \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta} - \frac{1}{2} \lambda [\mathcal{I}(\theta) d + \mathcal{I}(\theta)^\top d] = 0 \\
d &= \frac{\mathcal{I}(\theta)^{-1} \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta}}{\lambda} \\
\hat{d} &= \mathcal{I}(\theta)^{-1} \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta}
\end{aligned}
$$

for λ, we have:

$$
\begin{aligned}
\nabla_\lambda \mathcal{L}(d, \lambda) &= \frac{1}{2} d^\top \mathcal{I}(\theta) d - c = 0 \\
d^\top \mathcal{I}(\theta) d &= 2c
\end{aligned}
$$

then:

$$
\begin{aligned}
\left(\frac{\mathcal{I}(\theta)^{-1} \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta}}{\lambda} \right)^\top \mathcal{I}(\theta) \left(\frac{\mathcal{I}(\theta)^{-1} \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta}}{\lambda} \right) &= 2c \\
\lambda &= \sqrt{\frac{1}{2c} \left(\nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta} \right)^\top \mathcal{I}(\theta)^{-1} \left(\nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta} \right)} \\
d^* = \sqrt{\frac{2c}{\left(\nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta} \right)^\top \mathcal{I}(\theta)^{-1} \left(\nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta} \right)}} &* \mathcal{I}(\theta)^{-1} \nabla_{\theta'} \log P(y; \theta')|_{\theta'=\theta}
\end{aligned}
$$

(4.a)

$$d_{semi-sup}(\theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) + \alpha \sum_{i=1}^{\tilde{m}} \log p(\widetilde{x^{(i)}}, \widetilde{z^{(i)}}; \theta)$$

Because:

$$p(x^{(i)}, z^{(i)}; \theta^{(t+1)}) \geq p(x^{(i)}, z^{(i)}; \theta^{(t)})$$
$$\log p(\widetilde{x^{(i)}}, \widetilde{z^{(i)}}; \theta^{(t+1)}) \geq p(\widetilde{x^{(i)}}, \widetilde{z^{(i)}}; \theta^{(t)})$$

So:

$$d_{semi-sup}(\theta^{(t+1)}) \geq d_{semi-sup}(\theta^{(t)})$$

(4.b)

$$w_j^{(i)} = \frac{p(z^{(i)} = j) \cdot p(x^{(i)} | z^{(i)} = j; \mu_j, \Sigma_j)}{\sum_{i=1}^{m} p(z^{(i)} = j) \cdot p(x^{(i)} | z^{(i)} = j; \mu_j, \Sigma_j)}$$

(4.c)

$$set\ w_j^{(i)} = \widetilde{\mathbb{1}\{z^{(i)} = j\}}$$

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

$$\phi_l = \frac{1}{m + \alpha\tilde{m}} \left(\sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \widetilde{w_l^{(i)}} \right)$$

$$\mu_l = \frac{1}{\sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \widetilde{w_l^{(i)}}} \left(\sum_{i=1}^{m} w_l^{(i)} x^i + \alpha \sum_{i=1}^{\tilde{m}} \widetilde{w_l^{(i)}} x^i \right)$$