1. A Simple Neural Network.

(a) $z^{[1]} = W^{[1]} x + W_0^{[1]}$

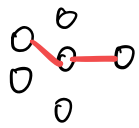
$h = \sigma(z^{[1]})$

$z^{[2]} = W^{[2]} h + W_0^{[1]}$

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

$l = \frac{1}{m} \sum_{i=0}^{m} (o^{(i)} - y^{(i)})^2 = \frac{1}{m} \sum J^{(i)}$

Sigmoid derivative: $\sigma' = \sigma(1-\sigma)$



$$\frac{\partial J}{\partial w_{1,2}^{[1]}} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2^{[1]}} \cdot \frac{\partial z_2^{[1]}}{\partial w_{1,2}^{[1]}}$$

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

$h_2 = w_{1,2}^{[1]} x_1 + w_{2,2}^{[1]} x_2 + w_{0,2}^{[1]}$

so the updating rule of $w_{1,2}^{[1]}$ is:

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

(b) Yes. Check the plot of the dataset, we can see the classifying boundary is a triangle, the three edges of triangle can be composed of three linear classifier.

Each neuron in the hidden layer can perform as a binary classifier. So it's possible to classify the dataset 100% right.

(c) It's not possible. If the activation function is linear,

$z^{[2]} = w^{[2]}(w^{[1]} x + w_0^{[1]}) + w_0^{[2]}$

$= w^{[2]} w^{[1]} x + w^{[2]} w_0^{[1]} + w_0^{[2]}$.

so the result is still linear. However, the dataset is not linearly separable.

2. KL Divergence and Maximum Likelihood

(a) Non-negativity: $\log \frac{P(x)}{Q(x)}$ is a convex function,

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

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

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

For a convex function f, according to Jensen's Inequality:

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

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

$$= -\log\left(\sum_{x \in X} Q(x)\right)$$

$$= 0.$$

so $D_{KL}(P \| Q) \geq 0$.

If $P=Q$, then $D_{KL}(P\|Q)=\sum_{x\in X} P(x)\cdot\log 1=0$.

If $D_{KL}(P\|Q)=0$, then $\sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)}=0$,

In the problem setting, $P(x)>0$ for $\forall x$. $\log\frac{P(x)}{Q(x)}=0$. $P=Q$.

So, $D_{KL}(P\|Q)=0$ iff $P=Q$.

(b) Chain Rule for KL Divergence.

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

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

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

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

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

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

(c) KL and Maximum likelihood.

proof: $\operatorname{argmin}_\theta D_{KL}(\hat{P}\|P_\theta)=\operatorname{argmin}_\theta\sum_{x\in X}\hat{P}(x)\log\frac{\hat{P}(x)}{P_\theta(x)}$

$$=\operatorname{argmin}_\theta\sum_{x\in X}\hat{P}(x)\log\hat{P}(x)-\hat{P}(x)\log P_\theta(x)$$

$$=\operatorname{argmin}_\theta\sum_{x\in X}-\hat{P}(x)\log P_\theta(x)$$

$$=\operatorname{argmax}_\theta\sum_{x\in X}\hat{P}(x)\log P_\theta(x)$$

$$=\operatorname{argmax}_\theta\sum_{x\in X}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\{x^{(i)}=x\}\right)\log P_\theta(x)$$

$$=\operatorname{argmax}_\theta\sum_{i=1}^{m}\log P_\theta(x^{(i)})$$

3. KL Divergence, Fisher Information, and the Natural Parameter

KL divergence is invariant to model parameterization (i), but the gradient w.r.t. the model parameter is not invariant to model parameterization (ii).

proof: (i) $y=mx+n$, $P(x)\,dx=P(y)\,dy$.

$$D_{KL}(P(x)\|Q(x))=\int_{x_a}^{x_b}P(x)\log\frac{P(x)}{Q(x)}\,dx$$

$$=\int_{y_a}^{y_b}P(y)\log\frac{P(y)\frac{dy}{dx}}{Q(x)\frac{dy}{dx}}\,dy$$

$$=\int_{y_a}^{y_b}P(y)\log\left(\frac{P(y)}{Q(y)}\right)dy=D_{KL}(P(y)\|Q(y))$$

(ii) eg. $D_{KL}(\hat{P}(x)\|P_\theta(x))$ is will change when updating parameter θ.

(a) Score function:

$$\begin{aligned}
\mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta}\right] &= \int_{-\infty}^{\infty} p(y;\theta)\,\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta}\,dy \\
&= \int_{-\infty}^{\infty} p(y;\theta)\nabla_{\theta}\log p(y;\theta)\,dy \\
&= \int_{-\infty}^{\infty} p(y;\theta)\cdot\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\,dy \\
&= \int_{-\infty}^{\infty} \nabla_\theta p(y;\theta)\,dy \\
&= \nabla_\theta \int_{-\infty}^{\infty} p(y;\theta)\,dy \\
&= \nabla_\theta 1\,dy = 0.
\end{aligned}$$

(b) Fisher Information.

X is real random vector,

$$\begin{aligned}
\mathrm{cov}(X,X) &= \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T] \\
&= \mathbb{E}[XX^T] - \mathbb{E}[X]\,\mathbb{E}[X]^T
\end{aligned}$$

$$\begin{aligned}
I(\theta) &= \mathrm{cov}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta}\right] \\
&= \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\,\nabla_{\theta'}\log p(y;\theta')^T\big|_{\theta'=\theta}\right] \\
&\quad - \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta}\right]\mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta}\right]^T \\
&= \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\,\nabla_{\theta'}\log p(y;\theta')^T\big|_{\theta'=\theta}\right]
\end{aligned}$$

(c) Fisher Information (alternative form)

$$\frac{\partial \log p(y;\theta)}{\partial \theta_i} = \frac{1}{p(y;\theta)}\,\frac{\partial p(y;\theta)}{\partial \theta_i}$$

$$\begin{aligned}
I(\theta)_{ij} &= \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\nabla_{\theta'}\log(y;\theta')^T\big|_{\theta'=\theta}\right]_{ij} \\
&= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{\partial \log p(y;\theta')}{\partial \theta_i}\cdot\frac{\partial \log p(y;\theta')^T}{\partial \theta_j}\right] \\
&= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{1}{(p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \log p(y;\theta)}{\partial\theta_i\partial\theta_j} &= \frac{\partial}{\partial\theta_j}\left(\frac{1}{p(y;\theta)}\cdot\frac{\partial p(y;\theta)}{\partial\theta_i}\right) \\
&= -\frac{1}{p(y;\theta)^2}\cdot\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j} + \frac{1}{p(y;\theta)}\cdot\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{y\sim p(y;\theta)}\left[-\nabla^2_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta}\right]_{ij} &= \mathbb{E}_{y\sim p(y)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j} - \frac{1}{p(y;\theta)}\cdot\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]_{ij} \\
&= \mathbb{E}_{y\sim p(y)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]_{ij} - \mathbb{E}_{y\sim p(y)}\left[\frac{1}{p(y;\theta)}\cdot\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]_{ij} \\
&= \mathbb{E}_{y\sim p(y)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]_{ij} - \int_{-\infty}^{\infty} p(y;\theta)\frac{1}{p(y;\theta)}\cdot\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}
\end{aligned}$$

$$= E_{y\sim p(y)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]_{ij} - \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int_{-\infty}^{\infty} p(y;\theta)dy$$

$$= E_{y\sim p(y)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\right]_{ij} = I(\theta)_{ij}$$

(d) Approxing D_{KL} with Fisher Information:

$$\log p(y;\tilde{\theta}) \approx \log p(y;\theta) + (\tilde{\theta}-\theta)^T \nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta} + \frac{1}{2}(\tilde{\theta}-\theta)^T(\nabla^2_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta})$$

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

$$= \log p(y;\theta) + d^T\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta} + \frac{1}{2}d^T(\nabla^2_{\theta'}\log p(y;\theta')\big|^{(\tilde{\theta}-\theta)}_{\theta'=\theta})d.$$

$$E_{y\sim p(y;\theta)}[\log p(y;\tilde{\theta})] = E_{y\sim p(y;\theta)}[\log p(y;\theta)] + 0 + \frac{1}{2}d^T E_{y\sim p(y;\theta)}[\nabla^2_\theta \log p(y;\theta)]d$$

$$D_{KL}(p(\theta)\|p(\theta+d)) = \int_{-\infty}^{\infty} p(y;\theta)\log\frac{p(y;\theta)}{p(y;\tilde{\theta})}$$

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

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

(e) Natural Gradient.

$d^* = \arg\max_d \ell(\theta+d)$ subject to $D_{KL}(p_\theta\|p_{\theta+d}) = c$.

construct Lagrangian

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

$$\approx \log p(y;\theta) + d^T\nabla_{\theta'}\log p(y;\theta')\big|_{\theta'=\theta} - \lambda[\frac{1}{2}d^T I(\theta)d - c]$$

$$\nabla_d\mathcal{L}(d,\lambda) \approx \nabla_{\theta'}\log(y;\theta')\big|_{\theta'=\theta} - \lambda I(\theta)d$$

$$= \frac{\nabla_{\theta'}p(y;\theta)\big|_{\theta'=\theta}}{p(y;\theta)} - \lambda I(\theta)d = 0. \qquad (a)$$

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

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

plug $\tilde{d}$ into (b)

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

$$= -\frac{1}{2\lambda^2 p(y;\theta)^2}\nabla_{\theta'}p(y;\theta)\big|_{\theta'=\theta}^T(I(\theta)^{-1})^T\nabla_{\theta'}p(y;\theta)\big|_{\theta'=\theta} + c$$

$=0$

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

plug λ back to δ.

$$\delta^* = \sqrt{\frac{2c}{\nabla_{\theta'} p(y;\theta)\big|_{\theta'=\theta}^T [I(\theta)]^{-1} \nabla_{\theta'} p(y;\theta)\big|_{\theta'=\theta}}} I(\theta)^{-1} \frac{\nabla_{\theta'} p(y;\theta)\big|_{\theta'=\theta}}{p(y;\theta)}$$

$$= \sqrt{\frac{2c}{\nabla_{\theta'} p(y;\theta)\big|_{\theta'=\theta}^T [I(\theta)]^{-1} \nabla_{\theta'} p(y;\theta)\big|_{\theta'=\theta}}} I(\theta)^{-1} \nabla_{\theta'} p(y;\theta)\big|_{\theta'=\theta}$$

(f) Relation with Newton's Method.

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

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

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

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

$$\theta := \theta + \delta$$

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

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

They're the same.

4. Semi-supervised EM.

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

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

$$\sum_i \log p(x^{(i)};\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

Jensen's Inequality

$$\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

E-step:

for each $i \in \{1, \dots m\}$, set

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^{(t)})$$

M-step:

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

(a) Convergence.

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

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

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

$$= \ell_{\text{unsup}}(\theta^{(t)}) + \alpha \ell_{\text{sup}}(\theta^{(t)})$$

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

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

(b) In E step, we need to reestimate $w_j^{(i)}$

For each i, j set

$$w_j^{(i)} := p(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma)$$

$$= \frac{p(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma)\, p(z^{(i)} = j \mid \phi)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma)\, p(z^{(i)} = l; \phi)}$$

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

(c) parameter ϕ, μ, Σ need to be updated.

In order to maximize log likelihood,

$$\ell_{\text{semi-sup}} = \ell_{\text{unsup}} + \alpha \ell_{\text{sup}}$$

$$= \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \phi, \mu, \Sigma)$$

$$= \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)} \mid z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi)$$

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

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

Similar to how we solve in unsupervised case, we construct the Lagrangian

ϕ: $\mathcal{L}(\phi) = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^{(i)} \log \phi_l + \beta\left(\sum_{l=1}^{k}\phi_l - 1\right) + \alpha\sum_{i=1}^{\tilde{m}}\sum_{l=1}^{k} 1\{\tilde{z}^{(i)}=l\}\log\phi_l$

$\nabla_{\phi_j}\mathcal{L}(\phi) = \sum_{i=1}^{m}\frac{w_j^{(i)}}{\phi_j} + \alpha\sum_{i=1}^{\tilde{m}}\frac{1\{\tilde{z}^{(i)}=j\}}{\phi_j} + \beta = 0, \quad \phi_j = -\frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}}{\beta}$

$\sum_{l=1}^{k}\phi_l = \frac{\sum_{i=1}^{m}\sum_{l=1}^{k} w_l^{(i)} + \alpha\sum_{i=1}^{\tilde{m}}\sum_{l=1}^{k} 1\{\tilde{z}^{(i)}=l\}}{-\beta} = \frac{m+\alpha\tilde{m}}{-\beta} = 1 \quad \therefore \beta = -(m+\alpha\tilde{m})$

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

μ_j: $\nabla_{\mu_j}\ell_{\text{semi-sup}} = \nabla_{\mu_j}\ell_{\text{unsup}} + \alpha\nabla_{\mu_j}\ell_{\text{sup}}$

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

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

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

Σ_j:

$\nabla_{\Sigma_j}\ell_{\text{semi-sup}} = \nabla_{\Sigma_j}\ell_{\text{unsup}} + \alpha\nabla_{\Sigma_j}\ell_{\text{sup}}$

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

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

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

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

$= 0$

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

(i) Semi-supervised EM algorithm requires less iteration than unsupervised EM.

(ii) Semi-supervised EM is more stable. The labeled data can efficiently guide unlabeled data to converge to the same color after different initialization.

(iii) The overall quality of semi-sup EM is better than unsup EM. EM sometimes failed to distinguish two close low-variance gaussian, but mix them into one gaussian instead.