Shanghai Jiao Tong University

Machine Learning PS: 3

Author: *LIUQiaoan 520030910220*

Course: *CS229: Machine Learning, Autumn 2018, Stanford*
Date: *August 3, 2022*

You can see each problem at Problem 1, Problem 2, Problem 3, Problem 4, Problem 5.

Problem 1

(a) **Note.** This sub-problem uses square loss function, instead of NLL.

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

The forwarding propogation step is:

$$\begin{aligned} z^{[1]} &= W^{[1]}x^{(i)} + W_0^{[1]} \\ a^{[1]} &= \sigma(z^{[1]}) \\ z^{[2]} &= W^{[2]}a^{[1]} + W_0^{[2]} \\ o^{(i)} = a^{[2]} &= \sigma(z^{[2]}) \end{aligned}$$

The back propogation step is:

$$\begin{aligned} \frac{\partial \mathcal{L}^{(\rangle)}}{\partial w_{1,2}^{[1]}} &= \frac{\partial \mathcal{L}^{(\rangle)}}{\partial o^{(i)}}\frac{\partial o^{(i)}}{\partial z^{[2]}}\frac{\partial z^{[2]}}{\partial a_2^{[1]}}\frac{\partial a_2^{[1]}}{\partial h_2}\frac{\partial h_2}{\partial w_{1,2}^{[1]}} \\ &= 2(o^{(i)} - y^{(i)}) \cdot o^{(i)}(1 - o^{(i)}) \cdot W_2^{[2]} \cdot h_2(1 - h_2) \cdot x_1^{(i)} \end{aligned}$$

(b) This is possible. From the dataset, we can see that there exists a triangle boundary which can seperate the blue circles and red crosses:

$$\begin{aligned} x_1 &> 0.5 \\ x_2 &> 0.5 \\ x_1 + x_2 &< 4 \end{aligned}$$

Then by the definition of step function, we can find the value of ws in hidden layer to satisfy 100% accuracy:

$$\begin{aligned} z_1 &= w_{1,1}^{[1]}x_1 + w_{2,1}^{[1]}x_2 + w_{0,1}^{[1]} = 2x_1 - 1 \\ \iff & w_{1,1}^{[1]} = 2, w_{2,1}^{[1]} = 0, w_{0,1}^{[1]} = -1 \end{aligned}$$

$$
\begin{aligned}
z_2 &= w_{1,2}^{[1]} x_1 + w_{2,2}^{[1]} x_2 + w_{0,2}^{[1]} = 2x_2 - 1 \\
\iff & w_{1,2}^{[1]} = 0, w_{2,2}^{[1]} = 2, w_{0,2}^{[1]} = -1
\end{aligned}
$$

$$
\begin{aligned}
z_3 &= w_{1,3}^{[1]} x_1 + w_{2,3}^{[1]} x_2 + w_{0,3}^{[1]} = 4 - x_1 - x_2 \\
\iff & w_{1,3}^{[1]} = -1, w_{2,3}^{[1]} = -1, w_{0,3}^{[1]} = 4
\end{aligned}
$$

For output layer:

$$
\begin{aligned}
z_1 &= w_1^{[2]} h_1 + w_2^{[2]} h_2 + w_3^{[2]} h_3 + w_4^{[2]} = h_1 + h_2 + h_3 - 3 \\
\iff & w_1^{[2]} = -1, w_2^{[2]} h_2 = -1, w_3^{[2]} = -1, w_4^{[2]} = 3
\end{aligned}
$$

(c) This is impossible. Since there is only one activation function to be step function (in th output layer), then we can only have one linear boundary, which is not enough for this problem since the datasets are not linearly seperable.

Problem 2

(a)
Sol1:

$$
\begin{aligned}
-D_{KL}(P||Q) &= \sum_x P \log \frac{Q}{P} \\
&= \mathbb{E}_{x\sim P} \log \frac{Q}{P} \\
&\leq \log \mathbb{E}_{x\sim P} \frac{Q}{P} \\
&= \log Q \\
&\leq \log 1 \\
&= 0
\end{aligned}
$$

Sol2:

$$
\begin{aligned}
-D_{KL}(P||Q) &= \sum_x P \log \frac{Q}{P} \\
&\leq \sum_x P(\frac{Q}{P} - 1) \quad \text{Use inequality } \log x \leq x - 1 \\
&= \sum_x (Q - P) \\
&= 0
\end{aligned}
$$

The second is to prove that $D_{KL}(P||Q) = 0$ iff $P = Q$. **Proof.**

Sufficiency: If $P = Q$, then $\log \frac{P}{Q} = 0$ for all $x \in \mathcal{X}$. Thus $D_{KL}(P||Q) = 0$.
Necessity. If $D_{KL}(P||Q) = 0$, then

$$\begin{aligned}
0 &= -D_{KL}(P||Q) \\
&= \sum_x P \log \frac{Q}{P} \\
&= \mathbb{E}_{x\sim P} \log \frac{Q}{P} \\
&\leq \log \mathbb{E}_{x\sim P} \frac{Q}{P} \\
&= \log Q \\
&\leq \log 1 \\
&= 0
\end{aligned}$$

Therefore, $\mathbb{E}_{x\sim P} \log \frac{Q}{P} = \log \mathbb{E}_{x\sim P} \frac{Q}{P}$. From the hint, we can get that $\frac{Q}{P} = \mathbb{E}_{x\sim P} \frac{Q}{P}$ with probability 1, i.e. $\frac{Q}{P}$ is a constant. Since P and Q are both probability distribution over X, then this constant must be 1. So $P = Q$.

(b)

$$\begin{aligned}
D_{KL}(P(X,Y)||Q(X,Y)) &= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)} \\
&= \sum_x \sum_y P(x,y) \log \frac{P(y|x)P(x)}{Q(y|x)Q(x)} \\
&= \sum_x \sum_y P(x,y) \log \frac{P(x)}{Q(x)} + \sum_x \sum_y P(x,y) \log \frac{P(y|x)}{Q(y|x)} \\
&= \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_x P(x) \log \frac{P(y|x)}{Q(y|x)} \\
&= D_{KL}(P(X)||Q(X)) + D_{KL}(P(Y|X)||Q(Y|X))
\end{aligned}$$

(c) We can write down $D_{KL}(\hat{P}||P_\theta)$ as:

$$\begin{aligned}
D_{KL}(\hat{P}||P_\theta) &= \sum_{x\in\mathcal{X}} \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)} \\
&= \sum_{x\in\mathcal{X}} \hat{P}(x) \log \hat{P}(x) - \sum_{x\in\mathcal{X}} \hat{P}(x) \log P_\theta(x)
\end{aligned}$$

Then

$$\begin{aligned}
\arg\min_\theta D_{KL}(\hat{P}||P_\theta) &= \arg\min_\theta -\sum_{x\in\mathcal{X}} \hat{P}(x)\log P_\theta(x) \\
&= \arg\max_\theta \sum_{x\in\mathcal{X}} \hat{P}(x)\log P_\theta(x) \\
&= \arg\max_\theta \sum_{x\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^{m} 1\left\{x^{(i)}=x\right\}\log P_\theta(x) \\
&= \arg\max_\theta \frac{1}{m}\sum_{i=1}^{m}\left[\sum_{x\in\mathcal{X}} 1\left\{x^{(i)}=x\right\}\log P_\theta(x)\right] \\
&= \arg\max_\theta \frac{1}{m}\sum_{i=1}^{m}\log P_\theta(x^{(i)}) \\
&= \arg\max_\theta \sum_{i=1}^{m}\log P_\theta(x^{(i)})
\end{aligned}$$

Problem 3

(a)

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

(b)

$$\begin{aligned}
\mathcal{I}(\theta) &= Cov_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right] \\
&= \mathbb{E}_{y\sim p(y;\theta)}[(\nabla_{\theta'}\log p(y;\theta') - \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\right]) \\
&\qquad (\nabla_{\theta'}\log p(y;\theta') - \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\right])^T|_{\theta'=\theta}] \\
&= \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}\right]
\end{aligned}$$

(c)

$$\frac{\partial \log p(y;\theta')}{\partial\theta'_j} = \frac{1}{p(y;\theta)}\frac{\partial p(y;\theta')}{\partial\theta'_j}$$

Then

$$\begin{aligned}-\nabla^2_{\theta'} \log p(y;\theta')_{ij} &= -\frac{\partial^2 \log p(y;\theta')}{\partial\theta'_i \partial\theta'_j} \\ &= \frac{1}{p(y;\theta')^2}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j} - \frac{1}{p(y;\theta')}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\end{aligned}$$

We take integral:

$$\begin{aligned}\mathbb{E}_{y\sim p(y;\theta)}\left[-\nabla^2_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right]_{ij} &= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta')^2}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\Big|_{\theta'=\theta}\right] \\ &\quad + \int_{-\infty}^{+\infty} p(y;\theta)\left[\frac{1}{p(y;\theta')}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\right]\Big|_{\theta'=\theta} dy \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta')^2}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\Big|_{\theta'=\theta}\right] + \int_{-\infty}^{+\infty}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\Big|_{\theta'=\theta} dy \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta')^2}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\Big|_{\theta'=\theta}\right] + \frac{\partial^2}{\partial\theta'_i\partial\theta'_j}\left[\int_{-\infty}^{+\infty} p(y;\theta')dy\right]\Big|_{\theta'=\theta} \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta')^2}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j}\Big|_{\theta'=\theta}\right]\end{aligned}$$

And since:

$$\begin{aligned}\mathcal{I}(\theta)_{ij} &= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{\partial \log p(y;\theta')}{\partial\theta'_i}\frac{\partial \log p(y;\theta')}{\partial\theta'_j}\Big|_{\theta'=\theta}\right] \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial p(y;\theta')}{\partial\theta'_i}\frac{\partial p(y;\theta')}{\partial\theta'_j}\Big|_{\theta'=\theta}\right] \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[-\nabla^2_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right]_{ij}\end{aligned}$$

We get that $\mathbb{E}_{y\sim p(y;\theta)}\left[-\nabla^2_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right] = \mathcal{I}(\theta)$.

(d) Set $f(\tilde{\theta}) = D_{KL}(p_\theta \| p_{\tilde{\theta}})$. From sub-problem (a), we know that

$$\begin{aligned}\nabla_{\tilde{\theta}} f(\tilde{\theta})|_{\tilde{\theta}=\theta} &= -\nabla_{\tilde{\theta}}\left[\sum_x p_\theta \log p_{\tilde{\theta}}\right]\Big|_{\tilde{\theta}=\theta} \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[\nabla_{\tilde{\theta}}\log p_{\tilde{\theta}}|_{\tilde{\theta}=\theta}\right] \\ &= 0\end{aligned}$$

From sub-problem (c), we know that

$$\begin{aligned}\nabla^2_{\tilde{\theta}} f(\tilde{\theta})|_{\tilde{\theta}=\theta} &= -\nabla^2_{\tilde{\theta}}\left[\sum_x p_\theta \log p_{\tilde{\theta}}\right]\Big|_{\tilde{\theta}=\theta} \\ &= \mathbb{E}_{y\sim p(y;\theta)}\left[-\nabla^2_{\tilde{\theta}}\log p_{\tilde{\theta}}|_{\tilde{\theta}=\theta}\right] \\ &= \mathcal{I}(\theta)\end{aligned}$$

Then we get:

$$\begin{aligned} f(\tilde{\theta}) &\approx f(\theta) + (\tilde{\theta}-\theta)^T \nabla_{\tilde{\theta}} f(\tilde{\theta})|_{\tilde{\theta}=\theta} + \frac{1}{2}(\tilde{\theta}-\theta)^T(\nabla^2_{\tilde{\theta}} f(\tilde{\theta})|_{\tilde{\theta}=\theta})(\tilde{\theta}-\theta) \\ &= D_{KL}(p_\theta \| p_\theta) + d^T \nabla_{\tilde{\theta}} f(\tilde{\theta})|_{\tilde{\theta}=\theta} + \frac{1}{2} d^T (\nabla^2_{\tilde{\theta}} f(\tilde{\theta})|_{\tilde{\theta}=\theta}) d \\ &= \frac{1}{2} d^T \mathcal{I} d \end{aligned}$$

(e)

$$\begin{aligned} \mathcal{L}(d,\lambda) &= f(d) - \lambda[g(d)-c] \\ &= \log p(y;\theta) + \frac{1}{p(y;\theta)} d^T \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta} - \lambda \left[\frac{1}{2} d^T \mathcal{I}(\theta) d - c\right] \end{aligned}$$

Take derivative and set to 0:

$$\begin{aligned} \nabla_d \mathcal{L}(d,\lambda) &= \frac{1}{p(y;\theta)} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta} - \lambda \mathcal{I}(\theta) d = 0 \\ \implies \tilde{d} &= \frac{1}{p(y;\theta)\lambda} \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta} \end{aligned}$$

$$\begin{aligned} \nabla_\lambda \mathcal{L}(d,\lambda) &= \frac{1}{2} d^T \mathcal{I}(\theta) d - c = 0 \\ \implies \frac{1}{2} \tilde{d}^T \mathcal{I}(\theta) \tilde{d} - c &= 0 \\ \implies \frac{1}{2p(y;\theta)^2 \lambda^2} \nabla_{\theta'} p(y;\theta')^T|_{\theta'=\theta} \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta} - c &= 0 \\ \implies \lambda^* &= \sqrt{\frac{\nabla_{\theta'} p(y;\theta')^T|_{\theta'=\theta} \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{2p(y;\theta)^2 c}} \end{aligned}$$

Plug λ^* into $\tilde{d}$:

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

(f) For natural gradient,

$$\begin{aligned} \mathcal{I}(\theta) &= \mathbb{E}_{y\sim p(y;\theta)} \left[-\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}\right] \\ &= \mathbb{E}_{y\sim p(y;\theta)} \left[-\nabla^2_{\theta'} l(\theta')|_{\theta'=\theta}\right] \\ &= \mathbb{E}_{y\sim p(y;\theta)} \left[-H\right] \end{aligned}$$

Then

$$\begin{aligned}\theta &:= \theta + \tilde{d} \\ &= \theta - \frac{1}{p(y;\theta)\lambda}\mathbb{E}_{y\sim p(y;\theta)}\left[H\right]^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}\end{aligned}$$

which is the same as Newton's method.

Problem 4

(a)

$$\begin{aligned}l_{semi-sup}(\theta^{(t+1)}) &\geq \sum_{i=1}^{m}(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})}) + \alpha(\sum_{i=1}^{\tilde{m}}\log p(\tilde{x}^{(i)},\tilde{z}^{(i)};\theta^{(t+1)})) \\ &\geq \sum_{i=1}^{m}(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})}) + \alpha(\sum_{i=1}^{\tilde{m}}\log p(\tilde{x}^{(i)},\tilde{z}^{(i)};\theta^{(t)})) \\ &= l_{semi-sup}(\theta^{(t)})\end{aligned}$$

The first inequality holds since

$$l_{semi-sup}(\theta) \geq \sum_{i=1}^{m}\left(\sum_{z^{(i)}} Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}\right) + \alpha\left(\sum_{i=1}^{\tilde{m}}\log p(\tilde{x}^{(i)},\tilde{z}^{(i)};\theta)\right)$$

choose $Q_i(z^{(i)}) = Q_i^{(t)}(z^{(i)})$ and $\theta = \theta^{(t+1)}$.

The second inequality holds since

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

(b) Latent variables: $z^{(i)}$. We want to update $w_j^{(i)} \triangleq Q_i(z^{(i)}=j)$

$$\begin{aligned}w_j^{(i)} &:= p(z^{(i)}|x^{(i)}|\phi,\mu.\Sigma) \\ &= \frac{p(x^{(i)}|z^{(i)};\mu_j,\Sigma_j)p(z^{(i)}=j;\phi)}{\sum_{j=1}^{k}p(x^{(i)}|z^{(i)};\mu_j,\Sigma_j)p(z^{(i)}=j;\phi)} \\ &= \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\phi_j}{\sum_{j=1}^{k}\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\phi_j}\end{aligned}$$

(c) Parameters: ϕ, μ, Σ.

First, we consider ϕ:

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

Take derivative and set to 0:

$$\begin{cases}\frac{\partial\mathcal{L}(\phi,\beta)}{\partial\phi_l} &= \sum_{i=1}^{m}\frac{w_l^{(i)}}{\phi_l}+\alpha\sum_{i=1}^{\tilde{m}}1\{\tilde{z}^{(i)}=l\}\frac{1}{\phi_l}-\beta=0\\ \frac{\partial\mathcal{L}(\phi,\beta)}{\partial\beta} &= \sum_{j=1}^{k}\phi_j-1=0\end{cases}$$

Then we get the expression of ϕ_l and plug it back to second equation:

$$\begin{aligned}&\phi_l=\frac{\sum_{i=1}^{m}w_l^{(i)}+\alpha\sum_{i=1}^{\tilde{m}}1\{\tilde{z}^{(i)}=l\}}{\beta}\\ \Longrightarrow&\sum_{l=1}^{k}\phi_l-1=\sum_{l=1}^{k}\frac{\sum_{i=1}^{m}w_l^{(i)}+\alpha\sum_{i=1}^{\tilde{m}}1\{\tilde{z}^{(i)}=l\}}{\beta}=\frac{m+\alpha\tilde{m}}{\beta}-1=0\\ \Longrightarrow&\beta=m+\alpha\tilde{m}\end{aligned}$$

As a result,

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

Second, we consider μ:

$$\mathcal{L}(\mu)=-\sum_{i=1}^{m}\sum_{j=1}^{k}w_j^{(i)}\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)-\alpha\sum_{i=1}^{\tilde{m}}\sum_{j=1}^{k}1\{\tilde{z}^{(i)}=j\}\frac{1}{2}(\tilde{x}^{(i)}-\mu_j)^T\Sigma_j^{-1}(\tilde{x}^{(i)}-\mu_j)$$

Take derivative and set to 0:

$$\begin{aligned}&\frac{\partial\mathcal{L}(\mu)}{\partial\mu_l}=\sum_{i=1}^{m}w_l^{(i)}\Sigma^{-1}(x^{(i)}-\mu_l)+\alpha\sum_{i=1}^{\tilde{m}}1\{\tilde{z}^{(i)}=l\}\Sigma^{-1}(\tilde{x}^{(i)}-\mu_l)=0\\ \Longrightarrow&\mu_l=\frac{\sum_{i=1}^{m}w_l^{(i)}x^{(i)}+\alpha\sum_{i=1}^{\tilde{m}}1\{\tilde{z}^{(i)}=l\}\tilde{x}^{(i)}}{\sum_{i=1}^{m}w_l^{(i)}+\alpha\sum_{i=1}^{\tilde{m}}1\{\tilde{z}^{(i)}=l\}}\end{aligned}$$

Third, we consider Σ:

$$\begin{aligned}\mathcal{L}(\Sigma)=&-\sum_{i=1}^{m}\sum_{j=1}^{k}w_j^{(i)}\left[\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)+\log|\Sigma|\right]\\ &-\alpha\sum_{i=1}^{\tilde{m}}\sum_{j=1}^{k}1\{\tilde{z}^{(i)}=j\}\left[\frac{1}{2}(\tilde{x}^{(i)}-\mu_j)^T\Sigma_j^{-1}(\tilde{x}^{(i)}-\mu_j)+\log|\Sigma|\right]\end{aligned}$$

Take derivative and set to 0:

$$\begin{aligned}\frac{\partial \mathcal{L}(\Sigma)}{\partial \Sigma_l} &= -\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left(\sum_{i=1}^{m} w_l^{(i)}(x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T\right)\Sigma^{-1} \\ &\quad - \frac{1}{2}\alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=l\}\Sigma^{-1} + \frac{1}{2}\alpha\Sigma^{-1}\left(\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=l\}(\tilde{x}^{(i)}-\mu_l)(\tilde{x}^{(i)}-\mu_l)^T\right)\Sigma^{-1} \\ &= -\frac{1}{2}\Sigma^{-1}\left(\sum_{i=1}^{m} w_l^{(i)} + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=l\}\right) \\ &\quad + \frac{1}{2}\Sigma^{-1}\left(\sum_{i=1}^{m} w_l^{(i)}(x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=l\}(\tilde{x}^{(i)}-\mu_l)(\tilde{x}^{(i)}-\mu_l)^T\right)\Sigma^{-1} \\ &= 0\end{aligned}$$

As a result,

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

Note.

$$\begin{aligned}0 = I' = (\Sigma\Sigma^{-1})' &= \Sigma'\Sigma^{-1} + \Sigma(\Sigma^{-1})' \\ \Longrightarrow (\Sigma^{-1})' &= -\Sigma^{-1}\Sigma'\Sigma^{-1}\end{aligned}$$

(f)

i Semi-supervised EM take less iterations to converge.

ii Semi-supervised EM seems to have more stability, since the assignments of unsupervised EM changes but Semi-supervised EM doesn't.

iii Semi-supervised EM has better overall quanlity, since the contours of different-color scatters are more like ellipses.

Problem 5

(b) Since we only use 16 colors to represent each pixel, which need 4 bits for computer to store. The factor of compression is $\frac{24}{4} = 6$.