CS 229, Fall 2018

Problem Set #3 Solutions: Deep Learning & Unsupervised learning

---

1.

(a)

$$
\begin{aligned}
\frac{\partial l}{\partial w_{1,2}^{[1]}} &= \frac{\partial l}{\partial o} \cdot \frac{\partial o}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_{1,2}^{[1]}} \\
&= \frac{2}{m} \sum_{i=1}^{m} (o^{(i)} - y^{(i)}) \cdot o^{(i)} (1 - o^{(i)}) w_2^{[2]} \cdot h_2^{(i)} (1 - h_2^{(i)}) x_1^{(i)} \\
&= \frac{2}{m} w_2^{[2]} \sum_{i=1}^{m} (o^{(i)} - y^{(i)}) o^{(i)} (1 - o^{(i)}) h_2^{(i)} (1 - h_2^{(i)}) x_1^{(i)}
\end{aligned}
$$

$$
h_2^{(i)} = \sigma\left(w_{0,2}^{[1]} + x_1^{(i)} w_{1,2}^{[1]} + x_2^{(i)} w_{2,2}^{[1]}\right)
$$

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

(b)

It is possible. The three neurons in the hidden layer can be viewed as three independent linear classifiers, whose decision boundaries are the three sides of the triangle ($x_1 = 0.5$, $x_2 = 0.5$ and $x_1 + x_2 = 4$). The output is also a linear classifier, which differentiate the point is inside or outside the triangle.

(c)

It's not possible. When we adopt linear function for the hidden layer and step function for the output, the entire neuron network can be viewed as one linear classifier (not three). Because the dataset is not linearly separable, so it's impossible to achieve 100% accuracy.

2.

(a)

$$
\begin{aligned}
D_{\mathrm{KL}}(P\|Q) &= \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} \\
&= -\sum_{x\in\mathcal{X}} P(x)\log\frac{Q(x)}{P(x)} \\
&= E\Big[-\log\frac{Q(x)}{P(x)}\Big] \\
&\geq -\log E\Big[\frac{Q(x)}{P(x)}\Big] \\
&= -\log\Big(\sum_{x\in\mathcal{X}} P(x)\frac{Q(x)}{P(x)}\Big) \\
&= -\log\sum_{x\in\mathcal{X}} Q(x) \\
&= -\log 1 \\
&= 0
\end{aligned}
$$

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

If $D_{\mathrm{KL}}(P\|Q)=0$, then $\frac{Q(x)}{P(x)} = E\Big[\frac{Q(x)}{P(x)}\Big] = 1$, namely $P=Q$.

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

(b)

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

(c)

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

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

3.

(a)

$$\nabla_\theta \log p(y;\theta) = \frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}$$

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

(b)

$$\begin{aligned}
\mathrm{Cov}[X] &= E[(X-E[X])(X-E[X])^T] \\
&= E[XX^T] \quad \text{when } E[X]=0
\end{aligned}$$

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

(c)

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

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

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

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

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

(d)

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

$$\begin{aligned}
\log p(y;\tilde{\theta}) &\approx \log p(y;\theta) + (\tilde{\theta}-\theta)^T \nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta} + \frac{1}{2}(\tilde{\theta}-\theta)^T \left(\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}\right)(\tilde{\theta}-\theta) \\
&= \log p(y;\theta) + d^T \nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta} + \frac{1}{2}d^T \left(\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}\right) d
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\tilde{\theta})] &= \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta)] + \frac{1}{2}d^T \mathbb{E}_{y\sim p(y;\theta)}[\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]d \\
&= \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta)] + \frac{1}{2}d^T \mathcal{I}(\theta) d
\end{aligned}$$

$$\begin{aligned}
D_{\mathrm{KL}}(p_\theta \| p_{\theta+d}) &= D_{\mathrm{KL}}(p_\theta \| p_{\tilde{\theta}}) \\
&= \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta)] - \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\tilde{\theta})] \\
&\approx \frac{1}{2}d^T \mathcal{I}(\theta) d
\end{aligned}$$

(e)

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

$$\begin{aligned}
\ell(\theta+d) &\approx \ell(\theta) + d^T \nabla_{\theta'} \ell(\theta')|_{\theta'=\theta} \\
&= \log p(y;\theta) + d^T \nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta} \\
&= \log p(y;\theta) + d^T \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)}
\end{aligned}$$

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta+d}) \approx \frac{1}{2}d^T \mathcal{I}(\theta) d$$

$$\begin{aligned}
\mathcal{L}(d,\lambda) &= \ell(\theta+d) - \lambda\Big[D_{\mathrm{KL}}(p_\theta \| p_{\theta+d}) - c\Big] \\
&\approx \log p(y;\theta) + d^T \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)} - \lambda\Big[\frac{1}{2}d^T \mathcal{I}(\theta) d - c\Big]
\end{aligned}$$

$$\nabla_d \mathcal{L}(d,\lambda) \approx \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)} - \lambda \mathcal{I}(\theta) d = 0$$

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

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

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

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

(f)

Newton's method

$$\theta := \theta - H^{-1}\nabla_\theta \ell(\theta)$$

Natural gradient

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

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

4.

(a)

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

(b)

Latent variables: $z^{(i)}$

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

(c)

Parameters: ϕ, μ, Σ

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

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

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

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

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

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

$$\begin{aligned}\nabla_{\mu_j}\ell_{\text{unsup}} &= \sum_{i=1}^{m} w_j^{(i)}\Sigma_j^{-1}(x^{(i)}-\mu_j)\\ &= \Sigma_j^{-1}\Big(\sum_{i=1}^{m} w_j^{(i)}x^{(i)} - \mu_j\sum_{i=1}^{m} w_j^{(i)}\Big)\end{aligned}$$

$$\begin{aligned}\nabla_{\mu_j}\ell_{\text{sup}} &= \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Sigma_j^{-1}(\tilde{x}^{(i)}-\mu_j)\\ &= \Sigma_j^{-1}\Big(\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\tilde{x}^{(i)} - \mu_j\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Big)\end{aligned}$$

$$\begin{aligned}\nabla_{\mu_j}\ell_{\text{semi-sup}} &= \nabla_{\mu_j}\ell_{\text{unsup}} + \alpha\nabla_{\mu_j}\ell_{\text{sup}}\\ &= \Sigma_j^{-1}\Big[\Big(\sum_{i=1}^{m} w_j^{(i)}x^{(i)} - \mu_j\sum_{i=1}^{m} w_j^{(i)}\Big) + \alpha\Big(\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\tilde{x}^{(i)} - \mu_j\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Big)\Big]\\ &= \Sigma_j^{-1}\Big[\Big(\sum_{i=1}^{m} w_j^{(i)}x^{(i)} + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\tilde{x}^{(i)}\Big) - \mu_j\Big(\sum_{i=1}^{m} w_j^{(i)} + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Big)\Big]\\ &= 0\end{aligned}$$

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

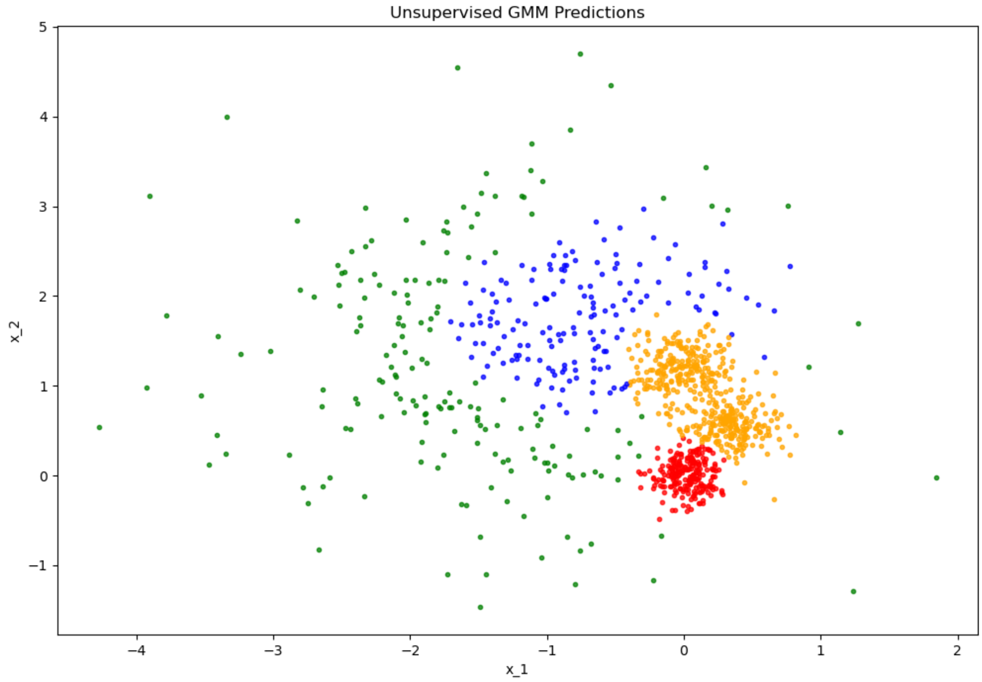
$$\nabla_{\Sigma_j}\ell_{\text{unsup}} = -\frac{1}{2}\sum_{i=1}^{m} w_j^{(i)}\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}\Big(\sum_{i=1}^{m} w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T\Big)\Sigma_j^{-1}$$

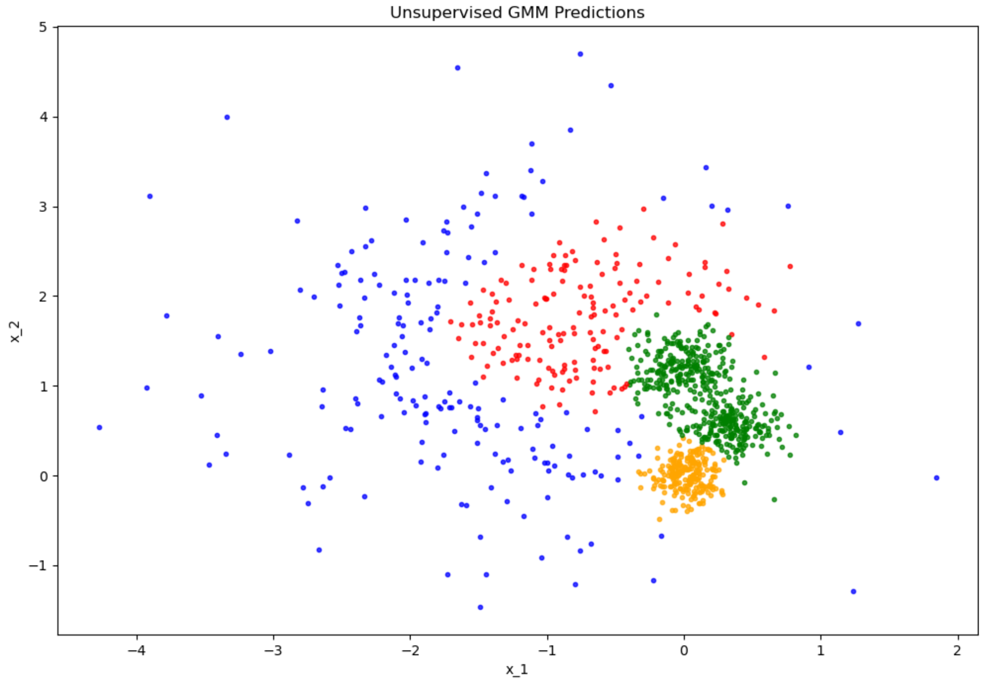
$$\nabla_{\Sigma_j}\ell_{\text{sup}} = -\frac{1}{2}\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}\Big(\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}(\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T\Big)\Sigma_j^{-1}$$

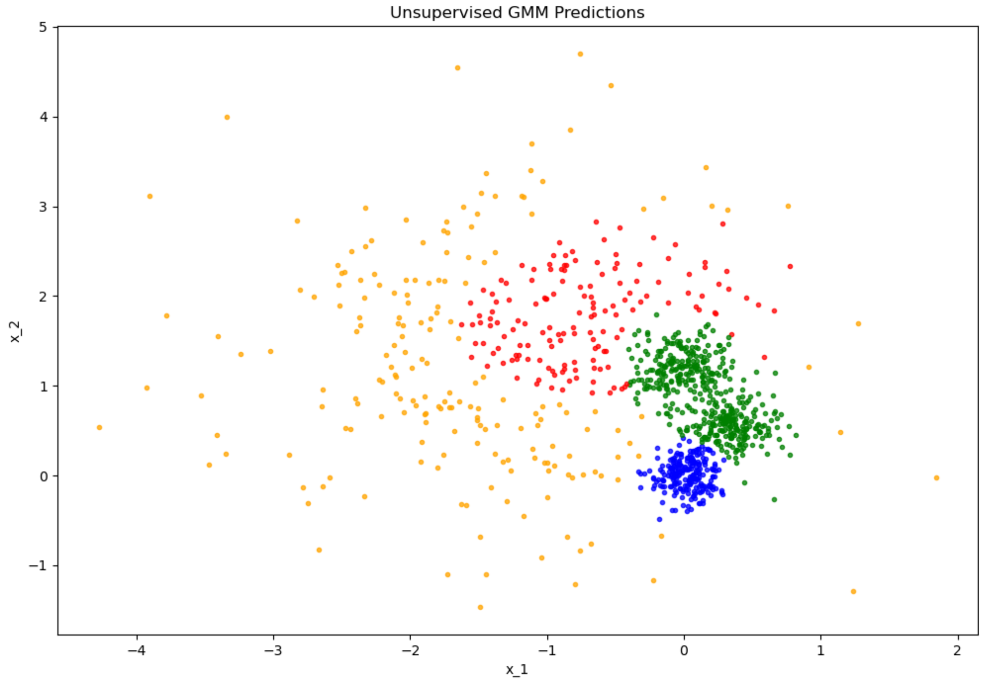
$$\begin{aligned}\nabla_{\Sigma_j}\ell_{\text{semi-sup}} &= \nabla_{\Sigma_j}\ell_{\text{unsup}} + \alpha\nabla_{\Sigma_j}\ell_{\text{sup}}\\ &= -\frac{1}{2}\sum_{i=1}^{m} w_j^{(i)}\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}\Big(\sum_{i=1}^{m} w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T\Big)\Sigma_j^{-1}\\ &\quad -\frac{1}{2}\alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Sigma_j^{-1} + \frac{1}{2}\alpha\Sigma_j^{-1}\Big(\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}(\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T\Big)\Sigma_j^{-1}\\ &= -\frac{1}{2}\Sigma_j^{-1}\Big(\sum_{i=1}^{m} w_j^{(i)} + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}\Big)\\ &\quad +\frac{1}{2}\Sigma_j^{-1}\Big(\sum_{i=1}^{m} w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T + \alpha\sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)}=j\}(\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T\Big)\Sigma_j^{-1}\\ &= 0\end{aligned}$$

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

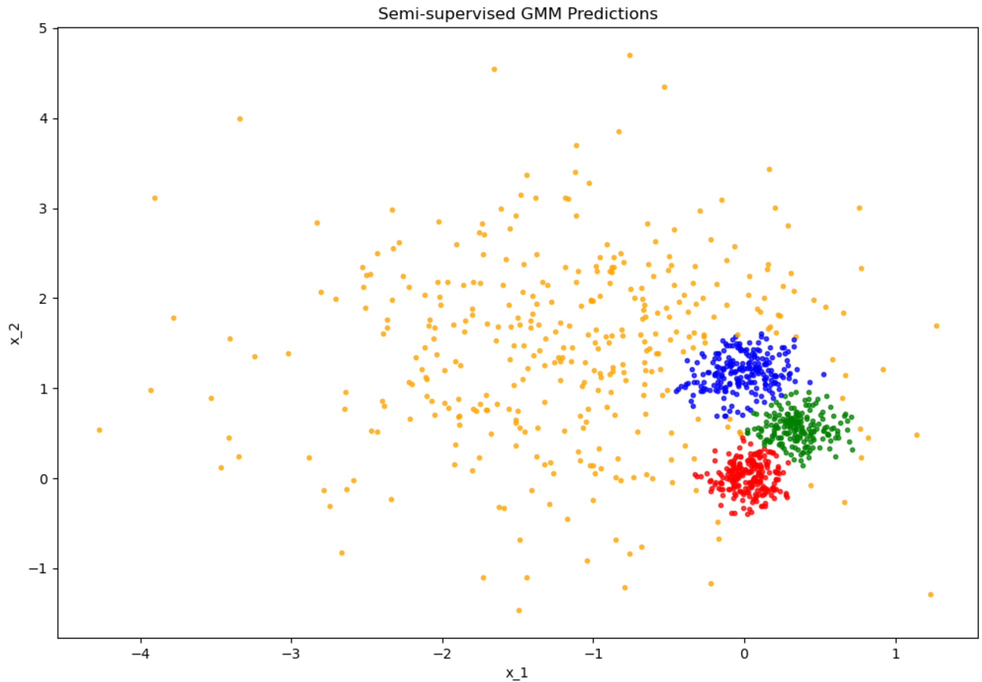
(d)

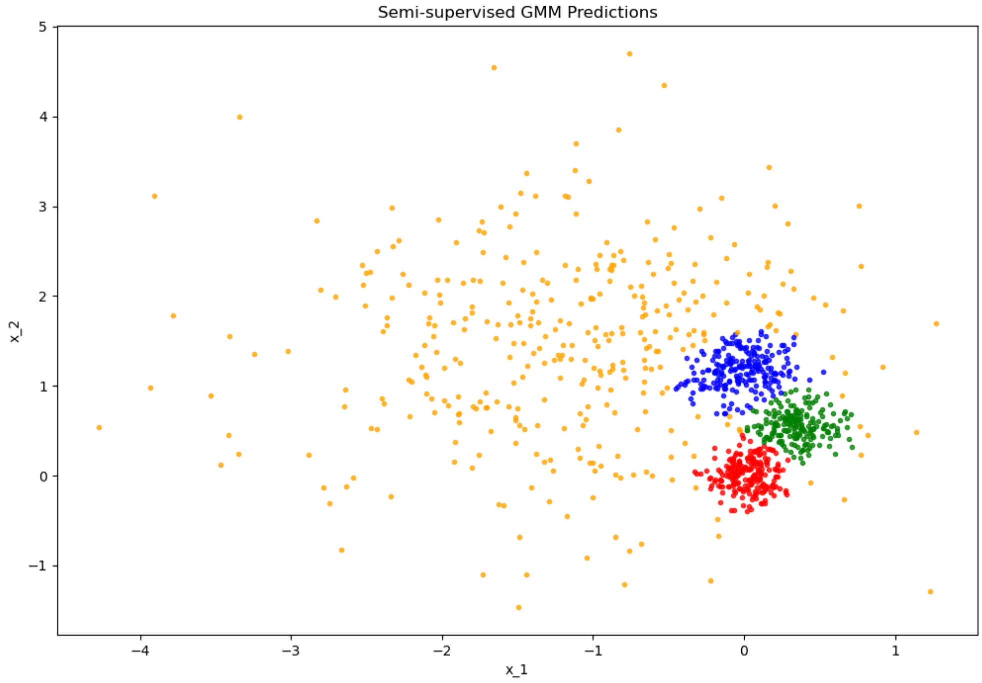

Unsupervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2

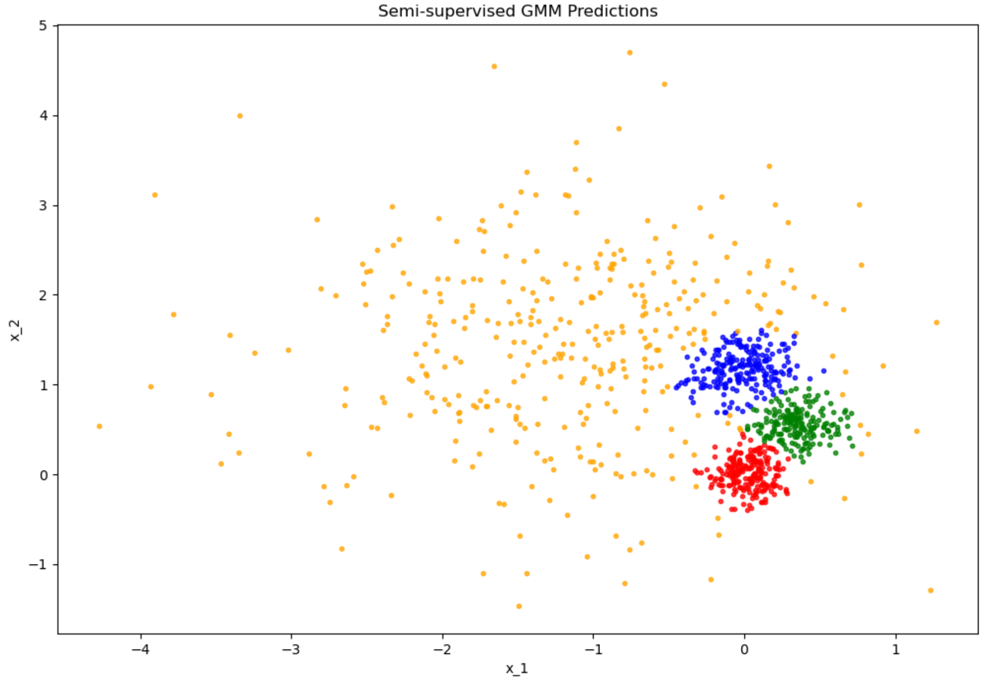

Unsupervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2


Unsupervised GMM Predictions
x_1
x_2

(e)


Semi-supervised GMM Predictions
x_1
x_2





(f)

i.

Semi-supervised EM take much less iterations to converge than unsupervised EM.

Nearly 50 iterations for semi-supervised EM and more than 1000 for unsupervised EM.

ii.

Semi-supervised EM are more stable than unsupervised EM.

The assignments by unsupervised EM are random with different random initializations.

But the assignments by semi-supervised EM are the same.

iii.

The overall quality of assignments by semi-supervised EM are higher than unsupervised EM.

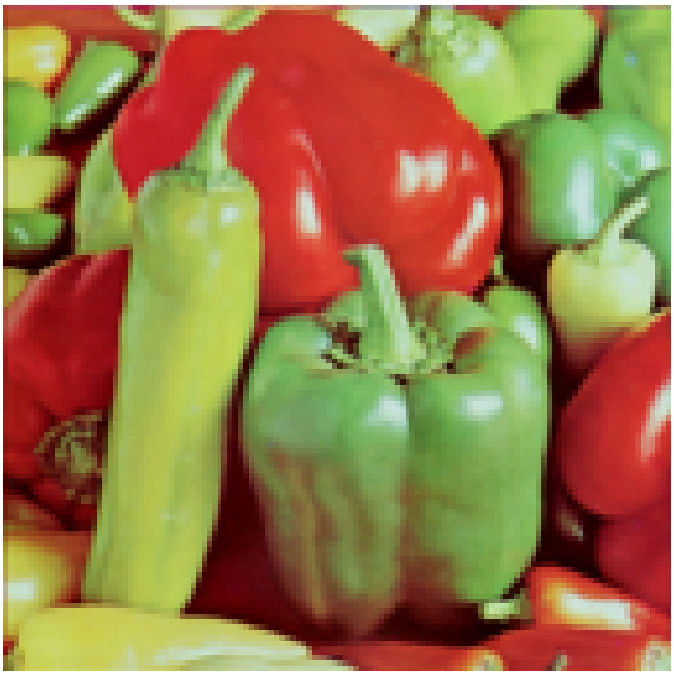
In the pictures of semi-supervised EM, there are three nearly the same low-variance Gaussian distributions, and a high-variance Gaussian distribution.

In the pictures of unsupervised EM, there are four Gaussian distributions which variances are different.

5.

(a)

Original small image



Original large image

Updated large image

(b)

In the original image, we need $3 \times 8 = 24$ bits to represent a pixel.

In the compressed image, we only need 4 bits (16 colors) to represent a pixel.

So the image are compressed by factor 6.