This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.b

The original image is storing data with 8 bits for each color so to represent a pixel it needs 3 * 8=24 bits. In the compressed image, we only need 4 bits (16 colors) to represent a pixel as we have 16 clusters.
So the image is compressed by factor of 6

2.a

$$
\begin{aligned}
\ell(\theta^{(t+1)}) &= \alpha \ell_{\textsf{sup}}(\theta^{(t+1)}) + \ell_{\textsf{unsup}}(\theta^{(t+1)}) && \text{Definition} \\
&\geq \alpha \ell_{\textsf{sup}}(\theta^{(t+1)}) + \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} && \text{Jensen's inequality} \\
&\geq \\
&\geq \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)}) + \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} \\
&\geq \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)}) + \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \\
&= \alpha \ell_{\textsf{sup}}(\theta^{(t)}) + \ell_{\textsf{unsup}}(\theta^{(t)}) && \text{Definition} \\
\ell_{\textsf{semi-sup}}(\theta^{(t+1)}) &\geq \ell_{\textsf{semi-sup}}(\theta^{(t)})
\end{aligned}
$$

2.b

From Lecture notes:
Latent variables are $z^{(i)}s$ meaning they are hidden/unobserved
E Step is given as follows:
$w_j^{(i)} := p(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$

Using Baye's rule we can write this as:
$p(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$
$= \frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)}=l;\mu,\Sigma)p(z^{(i)}=l;\phi)}$

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

2.c

List the parameters which need to be re-estimated in the M-step:

In order to simplify derivation, it is useful to denote

$$w_j^{(i)} = Q_i^{(t)}(z^{(i)} = j),$$

and

$$\tilde{w}_j^{(i)} = \begin{cases} \alpha & \tilde{z}^{(i)} = j \\ 0 & \text{otherwise.} \end{cases}$$

We further denote $S = \Sigma^{-1}$, and note that because of chain rule of calculus, $\nabla_S \ell = 0 \Rightarrow \nabla_\Sigma \ell = 0$. So we choose to rewrite the M-step in terms of S and maximize it w.r.t S, and re-express the resulting solution back in terms of Σ.

Based on this, the M-step becomes:

$$\begin{aligned}
\phi^{(t+1)}, \mu^{(t+1)}, S^{(t+1)} &= \arg\max_{\phi,\mu,S} \sum_{i=1}^{n} \sum_{j=1}^{k} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, S)}{Q_i^{(t)}(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x^{(i)}}, \tilde{z^{(i)}}; \phi, \mu, S) \\
&= \\
&\arg\max_{\phi,\mu,S} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log(\frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T S_j (x^{(i)} - \mu_j))\phi_j}{w_j^{(i)}}) + \\
&\sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \tilde{w}_j^{(i)} \log \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(\tilde{x}^{(i)} - \mu_j)^T S_j (\tilde{x}^{(i)} - \mu_j))\phi_j
\end{aligned}$$

Now, calculate the update steps by maximizing the expression within the argmax for each parameter (We will do the first for you).

ϕ_j: We construct the Lagrangian including the constraint that $\sum_{j=1}^{k} \phi_j = 1$, and absorbing all irrelevant terms into constant C:

$$\begin{aligned}
\mathcal{L}(\phi, \beta) &= C + \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \tilde{w}_j^{(i)} \log \phi_j + \beta \left(\sum_{j=1}^{k} \phi_j - 1 \right) \\
\nabla_{\phi_j} \mathcal{L}(\phi, \beta) &= \sum_{i=1}^{n} w_j^{(i)} \frac{1}{\phi_j} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \frac{1}{\phi_j} + \beta = 0 \\
&\Rightarrow \phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{-\beta} \\
\nabla_\beta \mathcal{L}(\phi, \beta) &= \sum_{j=1}^{k} \phi_j - 1 = 0 \\
&\Rightarrow \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{-\beta} = 1 \\
&\Rightarrow -\beta = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \right) \\
\Rightarrow \phi_j^{(t+1)} &= \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{\sum_{j=1}^{k} \left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \right)}
\end{aligned}$$

$$= \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{n + \alpha\tilde{n}}$$

μ_j: Next, derive the update for μ_j. Do this by maximizing the expression with the argmax above with respect to μ_j.

First, calculate the gradient with respect to μ_j:

$$\nabla_{\mu_j} = \sum_{i=1}^{n} w_j^{(i)}(S_j)(x^{(i)} - \mu_j) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}(S_j)(\tilde{x}^{(i)} - \mu_j))$$

Next, set the gradient to zero and solve for μ_j:

$$\begin{aligned}
0 &= \\
&\sum_{i=1}^{n} w_j^{(i)}(S_j)(x^{(i)} - \mu_j) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}(S_j)(\tilde{x}^{(i)} - \mu_j)) \\
&\sum_{i=1}^{n} w_j^{(i)}(S_j)(x^{(i)} - \mu_j) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}(S_j)(\tilde{x}^{(i)} - \mu_j)) \\
\mu_j &= \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \tilde{x}^{(i)}}{\sum_{i=1}^{n} w_j^{(i)} \ + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}
\end{aligned}$$

$\mathbf{\Sigma}_j$: Finally, derive the update for Σ_j via S_j. Again, Do this by maximizing the expression with the argmax above with respect to S_j.

.

First, calculate the gradient with respect to S_j:

$$\begin{aligned}
\nabla_{S_j} &= \nabla_{S_j}(\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log(\frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T S_j (x^{(i)} - \mu_j))\phi_j}{w_j^{(i)}}))+ \\
&\nabla_{S_j}(\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{k} \tilde{w}_j^{(i)} \log \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(\tilde{x}^{(i)} - \mu_j)^T S_j (\tilde{x}^{(i)} - \mu_j))\phi_j) \\
&= \sum_{i=1}^{n} w_j^{(i)}(-\frac{1}{2}(S_j) + \frac{1}{2}(x^{(i)} - \mu_j)^T(x^{(i)} - \mu_j))(S_j^{-1})+ \\
&\sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}(-\frac{1}{2}(S_j) + \frac{1}{2}(\tilde{x}^{(i)} - \mu_j)^T(\tilde{x}^{(i)} - \mu_j))(S_j^{-1}))
\end{aligned}$$

Next, set the gradient to zero and solve for S_j:

$$\begin{aligned}
0 &= \\
&\sum_{i=1}^{n} w_j^{(i)}(-\frac{1}{2}(S_j) + \frac{1}{2}(x^{(i)} - \mu_j)^T(x^{(i)} - \mu_j))(S_j^{-1})+ \\
&\sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}(-\frac{1}{2}(S_j) + \frac{1}{2}(\tilde{x}^{(i)} - \mu_j)^T(\tilde{x}^{(i)} - \mu_j))(S_j^{-1})) \\
\Sigma_j &= \frac{\sum_{i=1}^{n} w_j^{(i)}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}(\tilde{x}^{(i)} - \mu_j)(\tilde{x}^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)} \ + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}
\end{aligned}$$

This results in the final set of update expressions:

$$\phi_j := \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{n + \alpha\tilde{n}}$$

$$\mu_j := \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \tilde{x}^{(i)}}{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

$$\Sigma_j := \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} (\tilde{x}^{(i)} - \mu_j)(\tilde{x}^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

2.f

i.
Unsupervised EM took a lot more iterations to converge as compared to Semi-Supervised EM.
Unsupervised EM took almost 1000 of iteration to converge.
Semi-Supervised EM took approximately 50-60 iterations.

ii.
The assignments by unsupervised EM were random with different random initializations.
The assignments by semi-supervised EM were same or roughly the same.
Semi-supervised EM are more stable than unsupervised EM.

iii.
The pictures of semi-supervised EM have nearly 3 same low-variance Gaussian distributions, and 1 high-variance Gaussian distribution.
The pictures of unsupervised EM have four Gaussian distributions with different variances.
The overall quality of assignments by semi-supervised EM are higher than unsupervised EM.