CS 229: Assignment 3

Santiago de Buen

November 9, 2020

1 KL Divergence and Maximum Likelihood

(a) **Nonnegativity of KL divergence**

We want to prove that

$$\forall P, Q \qquad D_{KL}(P||Q) \geq 0$$

and

$$D_{KL}(P||Q) = 0 \qquad \text{if and only if} \qquad P = Q$$

The KL Divergence is defined as

$$
\begin{aligned}
D_{KL}(P||Q) &= \sum_{x\in\mathcal{X}} P(x)\text{log}\left(\frac{P(x)}{Q(x)}\right) \\
&= -\sum_{x\in\mathcal{X}} P(x)\text{log}\left(\frac{Q(x)}{P(x)}\right) \\
&= -\mathop{\mathbb{E}}_{x\sim P(x)}\left[\text{log}\left(\frac{Q(x)}{P(x)}\right)\right] \\
&\geq -\text{log}\left(\mathop{\mathbb{E}}_{x\sim P(x)}\left[\frac{Q(x)}{P(x)}\right]\right) && \text{by Jensen's inequality} \\
&= -\text{log}\left(\sum_{x\in\mathcal{X}}\frac{P(x)Q(x)}{P(x)}\right) \\
&= -\text{log}\left(\sum_{x\in\mathcal{X}} Q(x)\right) \\
&= -\text{log}(1) \\
&= 0
\end{aligned}
$$

We have proven nonnegativity

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

Furthermore, it is easy to see that this is only satisfied with equality when $P(x) = Q(x)$

$$\mathop{\mathbb{E}}_{x\sim P(x)}[\text{log}(1)] = \text{log}\left(\mathop{\mathbb{E}}_{x\sim P(x)}[1]\right) = 0$$

(b) **Chain rule for KL divergence**

We want to prove that

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

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

(c) KL and maximum likelihood

We want to prove that finding the maximum likelihood estimate for parameter θ is equivalent to finding P_θ with minimal KL divergence from $\hat{P}$.

$$
\text{arg}\ \min_\theta D_{KL}(\hat{P}||P_\theta) = \text{arg}\ \max_\theta \sum_{i=1}^{n} \text{log}P_\theta(x^{(i)}) \tag{1}
$$

Let's begin by using the definition of KL divergence.

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

Therefore, we have proven that maximum likelihood estimation is equivalent to minimizing the KL divergence. Note that we are able to eliminate $\hat{P}(x)\log(\hat{P}(x))$ because it does not depend on θ.

2 Semi-supervised EM

(a) **Convergence**

We want to show that the semi-supervised learning algorithm eventually converges. To do so, lets first start by describing the lower bound condition. Note that in this setting, the unsupervised and supervised segments are two concave functions. The addition is still a concave function, which allows us to apply Jensen's inequality to the semi-supervised setting.

$$\ell_{\text{semi-sup}}(\theta) \geq \sum_{i=1}^{n}\sum_{z^{(i)}} Q_i(z^{(i)})\log\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} + \alpha\sum_{i=1}^{\tilde{n}} \log P(\tilde{x}^{(i)},\tilde{z}^{(i)};\theta) \tag{2}$$

We want to show that

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

First, we would choose $Q_i^{(t)}(z^{(i)}) = P(x^{(i)},z^{(i)};\theta^{(t)})$ for every $i \in \{1,...,n\}$. That would make (2) hold with equality.

$$\ell_{\text{semi-sup}}(\theta^{(t)}) = \sum_{i=1}^{n}\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)})\log\frac{P(x^{(i)},z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} + \alpha\sum_{i=1}^{\tilde{n}} \log P(\tilde{x}^{(i)},\tilde{z}^{(i)};\theta^{(t)})$$

From (2), we know the following must hold:

$$\ell_{\text{semi-sup}}(\theta^{(t+1)}) \geq \text{ELBO}(x^{(i)}; Q_i^{(t+1)}, \theta^{(t+1)}) + \alpha \sum_{i=1}^{\tilde{n}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

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

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

The last inequality holds because $\theta^{(t+1)}$ is chosen to maximize (2). Therefore, we have proven convergence.

(b) **Semi-supervised E-Step**

The objective of the E-step is to model the probability of the latent variables being generated from different distributions. For the supervised data, we do not need to model the probability because we already know the ground truth. Therefore, we are only concerned with modeling the probability $P(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$ for the unsupervised part of our model. This means we only need to re-estimate the latent variables $z^{(i)}$.

From lecture notes, we know that

$$Q(z) = \frac{p(x, z; \theta)}{\sum_z p(x, z; \theta)}$$

Applying to our case:

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

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

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

At iteration $t + 1$, the M-Step can be computed as

$$w_{i,j}^{(t+1)} = Q_i^{(t+1)}(z^{(i)} = j) = P(z^{(i)} = j|x^{(i)}; \phi^{(t)}, \mu^{(t)}, \Sigma^{(t)}) \tag{3}$$

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

(c) **Semi-supervised M-Step**

In the M-step, we need to maximize likelihood by re-estimating the parameters of our model ϕ, μ, Σ. From question (a), we have

$$\ell_{\text{semi-sup}}(\theta^{(t)}) = \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{n}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})$$

Applying to mixture of Gaussians:

$$\begin{aligned} \ell_{\text{semi-sup}}(\phi^{(t)}, \mu^{(t)}, \Sigma^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{i,j}^{(t)} \log \frac{P(x^{(i)}|z^{(i)} = j; \mu^{(t)}, \Sigma^{(t)}) P(z^{(i)} = j; \phi^{(t)})}{w_{i,j}^{(t)}} + \\ \alpha \sum_{i=1}^{\tilde{n}} \log P(\tilde{x}^{(i)}|\tilde{z}^{(i)} = j; \mu^{(t)}, \Sigma^{(t)}) P(\tilde{z}^{(i)} = j; \phi^{(t)}) \end{aligned}$$

$$\begin{aligned} \ell_{\text{semi-sup}}(\phi^{(t)}, \mu^{(t)}, \Sigma^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{i,j}^{(t)} \log \left(\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \cdot \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)) \cdot \phi_j}{w_{i,j}^{(t)}} \right) + \\ \alpha \sum_{i=1}^{\tilde{n}} \log \left(\frac{1}{(2\pi)^{d/2}|\Sigma_{\tilde{z}^{(i)}}|^{1/2}} \cdot \exp(-\frac{1}{2}(\tilde{x}^{(i)} - \mu_{\tilde{z}^{(i)}})^T \Sigma_{\tilde{z}^{(i)}}^{-1} (\tilde{x}^{(i)} - \mu_{\tilde{z}^{(i)}})) \cdot \phi_{\tilde{z}^{(i)}} \right) \end{aligned} \tag{4}$$

Paramter μ_l :

$$\nabla_{\mu_l} \ell_{\text{semi-sup}}(\cdot) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{i,l}^{(t)} \nabla_{\mu_l} \left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) + \alpha \sum_{i=1}^{\tilde{n}} \nabla_{\mu_l} \left(-\frac{1}{2}(\tilde{x}^{(i)} - \mu_{\tilde{z}^{(i)}})^T \Sigma_{\tilde{z}^{(i)}}^{-1} (\tilde{x}^{(i)} - \mu_{\tilde{z}^{(i)}}) \right)$$

$$= -\frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} \nabla_{\mu_l} (-2x^{(i)} \Sigma_l^{-1} \mu_l + \mu_l^T \Sigma_l^{-1} \mu_l) - \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} \nabla_{\mu_l} (-2\tilde{x}^{(i)} \Sigma_l^{-1} \mu_l + \mu_l^T \Sigma_l^{-1} \mu_l)$$

$$= -\frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} (-2x^{(i)} \Sigma_l^{-1} + 2\Sigma_l^{-1} \mu_l) - \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} (-2\tilde{x}^{(i)} \Sigma_l^{-1} + 2\Sigma_l^{-1} \mu_l)$$

$$= \sum_{i=1}^{n} w_{i,l}^{(t)} (x^{(i)} \Sigma_l^{-1} - \Sigma_l^{-1} \mu_l) + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} (\tilde{x}^{(i)} \Sigma_l^{-1} - \Sigma_l^{-1} \mu_l)$$

Setting the derivative to zero and solving for μ_l:

$$\sum_{i=1}^{n} w_{i,l}^{(t)} \Sigma_l^{-1} \mu_l + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} \Sigma_l^{-1} \mu_l = \sum_{i=1}^{n} w_{i,l}^{(t)} x^{(i)} \Sigma_l^{-1} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} \tilde{x}^{(i)} \Sigma_l^{-1}$$

$$\mu_l \sum_{i=1}^{n} w_{i,l}^{(t)} + \mu_l \cdot \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} = \sum_{i=1}^{n} w_{i,l}^{(t)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}\tilde{x}^{(i)}$$

$$\mu_l^{(t)} = \frac{\sum_{i=1}^{n} w_{i,l}^{(t)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}\tilde{x}^{(i)}}{\sum_{i=1}^{n} w_{i,l}^{(t)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}} \tag{5}$$

Parameter Σ_l :

To simplify operations, let's first focus on the unsupervised part of equation (4). We will come back to the supervised part in a later stage.

$$\nabla_{\Sigma_l} \ell_{\text{unsup}} = -\sum_{i=1}^{n} \sum_{j=1}^{k} \nabla_{\Sigma_l} w_{i,j}^{(t)} \log(|\Sigma_j|^{\frac{1}{2}}) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \nabla_{\Sigma_l} w_{i,j}^{(t)} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)$$

Substituting $S = \Sigma^{-1}$ and recalling that $|X^{-1}| = \frac{1}{|X|}$ for invertible X.

$$\nabla_{S_l} \ell_{\text{unsup}} = \frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} \nabla_{S_l} \log(|S_l|) - \frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} \nabla_{S_l} (x^{(i)} - \mu_l)^T S_l (x^{(i)} - \mu_l)$$

Also recall that $\nabla_x |x| = |x|(x^{-1})^T$:

$$\nabla_{S_l} \ell_{\text{unsup}} = \frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} \frac{1}{|S_l|} |S_l| (S_l^{-1})^T - \frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} (x^{(i)} - \mu_l)^T (x^{(i)} - \mu_l)$$

Substituting $S = \Sigma^{-1}$ and simplifying:

$$\nabla_{\Sigma_l} \ell_{\text{unsup}} = \frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} \Sigma_l^T - \frac{1}{2} \sum_{i=1}^{n} w_{i,l}^{(t)} (x^{(i)} - \mu_l)^T (x^{(i)} - \mu_l) \tag{6}$$

Now focusing on the supervised part of equation (4):

$$\nabla_{\Sigma_l} \ell_{\text{sup}} = \nabla_{\Sigma_l} \left(\alpha \sum_{i=1}^{\tilde{n}} \log(|\Sigma_{\tilde{z}^{(i)}}^{-1}|^{1/2}) - \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} (\tilde{x}^{(i)} - \mu_{\tilde{z}^{(i)}})^T \Sigma_{\tilde{z}^{(i)}}^{-1} (\tilde{x}^{(i)} - \mu_{\tilde{z}^{(i)}}) \right)$$

Using the same substitution techniques as above:

$$\nabla_{S_l} \ell_{\text{sup}} = \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} \nabla_{S_l} \log(|S_l|) - \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} \nabla_{S_l} (\tilde{x}^{(i)} - \mu_l)^T S_l (\tilde{x}^{(i)} - \mu_l)$$

$$\nabla_{S_l} \ell_{\text{sup}} = \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} (S_l^{-1})^T - \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} (\tilde{x}^{(i)} - \mu_l)^T (\tilde{x}^{(i)} - \mu_l)$$

$$\nabla_{\Sigma_l} \ell_{\text{sup}} = \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} (\Sigma_l)^T - \frac{1}{2} \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\} (\tilde{x}^{(i)} - \mu_l)^T (\tilde{x}^{(i)} - \mu_l) \tag{7}$$

Putting together equations (6) and (7), and setting $\nabla_{\Sigma_l} = 0$:

$$\sum_{i=1}^{n} w_{i,l}^{(t)}\Sigma_l^T + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}\Sigma_l^T = \sum_{i=1}^{n} w_{i,l}^{(t)}(x^{(i)} - \mu_l)^T(x^{(i)} - \mu_l) + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}(\tilde{x}^{(i)} - \mu_l)^T(\tilde{x}^{(i)} - \mu_l)$$

Simplifying, we obtain

$$\Sigma_l^{(t)} = \frac{\sum_{i=1}^{n} w_{i,l}^{(t)}(x^{(i)} - \mu_l)(x^{(i)} - \mu_l)^T + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}(\tilde{x}^{(i)} - \mu_l)(\tilde{x}^{(i)} - \mu_l)^T}{\sum_{i=1}^{n} w_{i,l}^{(t)} + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}} \tag{8}$$

Parameter $\phi^{(t+1)}$:

Taking only the relevant terms from (4), we have

$$\nabla_{\phi_j}\ell(\cdot) = \nabla_{\phi_j}\left(\sum_{i=1}^{n}\sum_{j=1}^{k} w_{i,j}^{(t)}\log(\phi_j) + \alpha\sum_{i=1}^{\tilde{n}}\log(\phi_{\tilde{z}^{(i)}})\right)$$

But we have an additional constraint that $\sum_{j=1}^{k}\phi_j = 1$. We construct the Langrangian:

$$L(\phi) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{i,j}^{(t)}\log(\phi_j) + \alpha\sum_{i=1}^{\tilde{n}}\log(\phi_{\tilde{z}^{(i)}}) + \beta\left(\sum_{j=1}^{k}\phi_j - 1\right) \tag{9}$$

Taking the derivative of (9) with respect to ϕ:

$$\nabla_{\phi_l}L(\phi) = \sum_{i=1}^{n}\frac{w_{i,l}}{\phi_l} + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}\frac{1}{\phi_l} + \beta$$

Setting to zero and solving, we obtain

$$\phi_l = \frac{\sum_{i=1}^{n} w_{i,l} + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}}{-\beta}$$

Using the constraint that $\sum_{j=1}^{k}\phi_j = 1$, we can find β

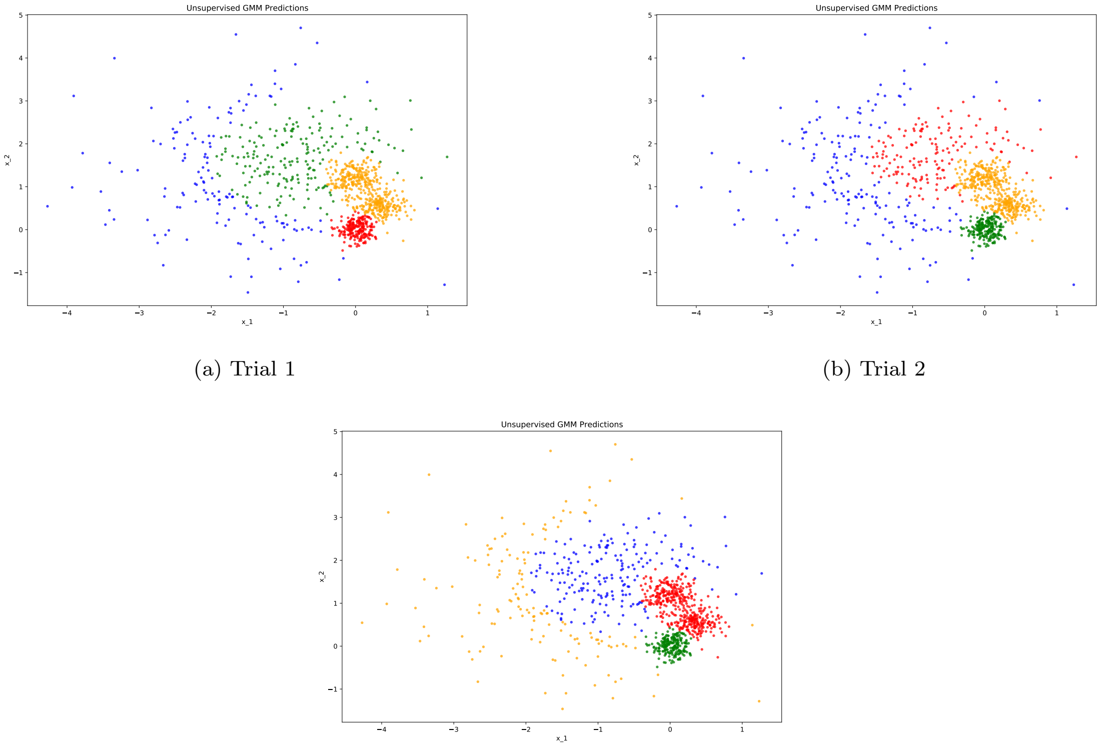
$$-\beta = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{i,j} + \alpha\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{k} 1\{\tilde{z}^{(i)} = l\} = n + \alpha\tilde{n}$$

Therefore, we have found the parameter update for ϕ

$$\phi_l^{(t)} = \frac{1}{n + \alpha\tilde{n}}\left(\sum_{i=1}^{n} w_{i,l}^{(t)} + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = l\}\right) \tag{10}$$

(d) Classical (Unsupervised) EM Implementation

The classical EM implementation produces the following results.



(a) Trial 1

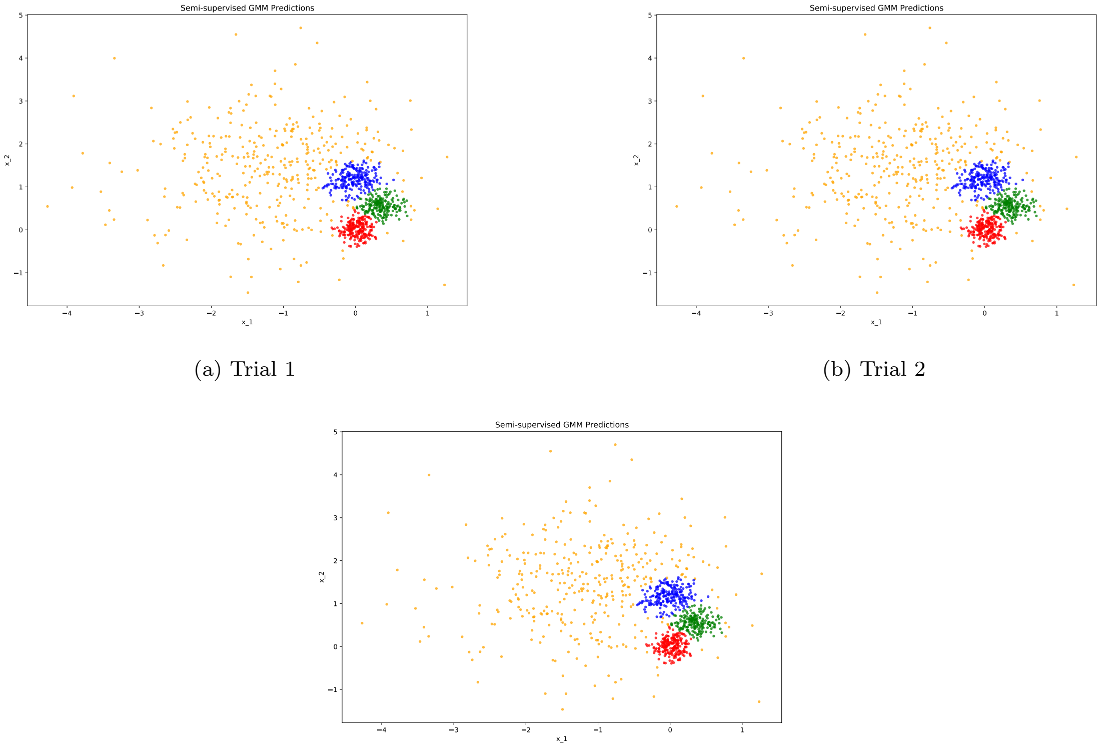
(b) Trial 2

(c) Trial 3

Figure 1: Clusters for unsupervised EM implementation

(e) **Semi-supervised EM Implementation**

The Semi-supervised EM implementation produces the following results.



(a) Trial 1

(b) Trial 2

(c) Trial 3

Figure 2: Clusters for semi-supervised EM implementation

(f) **Comparison of Unsupervised and Semi-supervised EM**

i. Number of iterations taken to converge

The semi-supervised EM model is able to converge in fewer iterations. This makes intuitive sense, as the model has some specific guidance on where to construct the Gaussian distributions, and it takes less time to reach convergence.

Trial	Unsupervised	Semi-supervised
1	160	20
2	164	30
3	126	32

Table 1: Number of iterations taken to converge

ii. Stability

The semi-supervised model is more stable than the unsupervised model. While there aren't drastic changes in the structure of the unsupervised EM clusters, there are a few points that change clusters with each different initialization. The EM model is subject to finding local optima, so this is expected. With the semi-supervised model, the clusters are nearly identical. It is more robust to different initialization.

iii. Quality of assignments

Because we know that the underlying data was generated with three low-variance Gaussian distributions and a fourth, high-variance Gaussian distribution, we can assert that the semi-supervised EM model generates higher quality clusters. In the unsupervised case, each cluster has a distinct variance whereas in the self-supervised case we can clearly see three low-variance clusters and one high-variance cluster.

3 Variational Inference in a Linear Gaussian Model

(a) **Exact marginal inference**

The process for generating x is $x = Wz + b + \delta$. We know that $p(x)$ must be a Gaussian distribution $\mathcal{N}(\nu, \Gamma)$. Our goal is to express ν and Γ as functions of (W, β, γ).

Finding ν first.

$$\begin{aligned}\nu &= \mathbb{E}[x] \\ &= \mathbb{E}[Wz + b + \delta] \\ &= W\mathbb{E}[z] + \mathbb{E}[b] + \mathbb{E}[\delta] \\ &= b\end{aligned}$$

Now for $\Gamma = Cov(x)$:

$$\begin{aligned}\Gamma &= \mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T \\ &= \mathbb{E}[(Wz + b + \delta)(Wz + b + \delta)^T] - \mathbb{E}[(Wz + b + \delta)]\mathbb{E}[(Wz + b + \delta)]^T \\ &= \mathbb{E}[(Wz + b + \delta)(z^TW^T + b^T + \delta^T)] - (\mathbb{E}[Wz] + \mathbb{E}[b] + \mathbb{E}[\delta])(\mathbb{E}[Wz] + \mathbb{E}[b] + \mathbb{E}[\delta])^T \\ &= \mathbb{E}[(Wz + b + \delta)(z^TW^T + b^T + \delta^T)] - bb^T \\ &= \mathbb{E}[Wzz^TW^T] + \mathbb{E}[Wzb^T] + \mathbb{E}[Wz\delta^T] + \mathbb{E}[bz^TW^T] + \mathbb{E}[bb^T] + \mathbb{E}[b\delta^T] + \mathbb{E}[\delta z^TW^T] + \mathbb{E}[\delta b^T] + \mathbb{E}[\delta\delta^T] - bb^T \\ &= W\mathbb{E}[zz^T]W^T + bb^T + \mathbb{E}[\delta\delta^T] - bb^T \\ &= WW^T + \gamma^2 I_d\end{aligned}$$

(b) **Understanding the ELBO**

The Evidence Lower Bound (ELBO) is defined as

$$\text{ELBO}(x;q) = \mathbb{E}_{z\sim q} \ln \frac{p(x,z)}{q(z)}$$

and the KL divergence as

$$D_{KL}(q||p) = \mathbb{E}_{z\sim q} \ln \frac{q(z)}{p(z)}$$

i. Prove the following

$$\text{ELBO}(x;q) = \mathbb{E}_{z\sim q} \ln p(x|z) - D_{KL}(q\,||\,p_z) \tag{11}$$

$$\begin{aligned}
\text{ELBO}(x;q) &= \mathbb{E}_{z\sim q} \ln \frac{p(x,z)}{q(z)} \\
&= \mathbb{E}_{z\sim q}(\ln p(x,z) - \ln q(z)) \\
&= \mathbb{E}_{z\sim q}(\ln p(x|z) + \ln p(z) - \ln q(z)) \\
&= \mathbb{E}_{z\sim q} \ln p(x|z) - \mathbb{E}_{z\sim q} \ln \frac{q(z)}{p(z)} \\
&= \mathbb{E}_{z\sim q} \ln p(x|z) - D_{KL}(q\,||\,p_z)
\end{aligned}$$

ii. Prove the following

$$\text{ELBO}(x;q) = \ln p(x) - D_{KL}(q\,||\,p_{z|x}) \tag{12}$$

$$\begin{aligned}
\text{ELBO}(x;q) &= \mathbb{E}_{z\sim q} \ln \frac{p(x,z)}{q(z)} \\
&= \mathbb{E}_{z\sim q}(\ln p(z|x) + \ln p(x) - \ln q(z)) \\
&= \mathbb{E}_{z\sim q} \ln p(x) - \mathbb{E}_{z\sim q} \ln \frac{q(z)}{p(z|x)} \\
&= \ln p(x) - D_{KL}(q\,||\,p_{z|x})
\end{aligned}$$

(c) **Optimizing the ELBO via gradient ascent**

i. Provide closed-form expressions for $\nabla_\mu D_{KL}(q||p_z)$ and $\nabla_\sigma D_{KL}(q||p_z)$

The KL divergence between q and p_z can be expressed as follows

$$D_{KL}(q||p_z) = \sum_{i=1}^{m} \left[-\ln \sigma_i + \frac{1}{2}(\sigma_i^2 + \mu_i^2 - 1) \right] \tag{13}$$

It is straightforward to derive the closed form expressions:

$$\nabla_\mu D_{KL}(q||p_z) = \mu \tag{14}$$

$$\nabla_\sigma D_{KL}(q||p_z) = -\left(\frac{1}{\sigma}\right) + \sigma = \frac{\sigma^2 - 1}{\sigma} \tag{15}$$

where μ and σ are vectors of size m.

ii. Provide closed-form expressions for $\nabla_\mu \ln p(x|\mu + \sigma \odot \epsilon)$ and $\nabla_\sigma \ln p(x|\mu + \sigma \odot \epsilon)$

First, lets determine the closed-form expression for $\nabla_z \ln p(x|z)$ in terms of (x, γ, W, b, z). We know that $x|z \sim \mathcal{N}(Wz + b, \gamma^2 I_d)$

$$p(x|z) = \frac{1}{(2\pi)^{d/2}|\gamma^2 I_d|^{1/2}} \exp\left(-\frac{1}{2}(x - (Wz + b))^T(\gamma^2 I_d)^{-1}(x - (Wz + b))\right)$$

Taking the derivative of the log-likelihood:

$$\begin{aligned}
\nabla_z \ln p(x|z) &= -\frac{1}{2}\nabla_z (x - Wz - b)^T(\gamma^2 I_d)^{-1}(x - Wz - b) \\
&= -\frac{1}{2}\nabla_z \left[\frac{1}{\gamma^2}\left(-x^T Wz - z^T W^T x + z^T W^T Wz + z^T W^T b + b^T Wz\right)\right] \\
&= -\frac{1}{2\gamma^2}(-2W^T x + 2W^T Wz + 2W^T b) \\
&= \frac{1}{\gamma^2}(W^T x - W^T Wz - W^T b)
\end{aligned}$$

Now expanding $z = \mu + \sigma \odot \epsilon$ and taking the derivative with respect to μ and σ,

$$\begin{aligned}
\nabla_\mu \ln p(x|\mu + \sigma \odot \epsilon) &= \frac{1}{\gamma^2}(W^T x - W^T Wz - W^T b) \\
&= \frac{1}{\gamma^2}(W^T x - W^T W(\mu + \sigma \odot \epsilon) - W^T b)
\end{aligned}$$

Now for σ,

$$\begin{aligned}\nabla_\sigma \ln p(x|\mu + \sigma \odot \epsilon) &= \frac{1}{\gamma^2}(W^T x - W^T W z - W^T b) \cdot \frac{\partial z}{\partial \sigma} \\ &= \frac{1}{\gamma^2}(W^T x - W^T W(\mu + \sigma \odot \epsilon) - W^T b) \cdot \epsilon\end{aligned}$$

Therefore,

$$\nabla_\mu \ln p(x|\mu + \sigma \odot \epsilon) = \frac{1}{\gamma^2}(W^T x - W^T W(\mu + \sigma \odot \epsilon) - W^T b) \tag{16}$$

$$\nabla_\sigma \ln p(x|\mu + \sigma \odot \epsilon) = \frac{1}{\gamma^2}(W^T x - W^T W(\mu + \sigma \odot \epsilon) - W^T b) \cdot \epsilon \tag{17}$$

(d) **Coding problem**

The output of lgm.py is

```
**********BEGIN EXPERIMENT**********
Experiment 1: Model-1 applied to observation x1
Model-1 loaded with W from W1.txt
Model-1 ln p(x1): -5.324831939364563
Model-1 ELBO(x1) using initial q: -17.442698224589353
Model-1 ELBO(x1) using optimized q: -10.313593083006968
**********END EXPERIMENT**********

**********BEGIN EXPERIMENT**********
Experiment 2: Model-2 applied to observation x2
Model-2 loaded with W from W2.txt
Model-2 ln p(x2): -5.324834488519816
Model-2 ELBO(x2) using initial q: -37.25719900317399
Model-2 ELBO(x2) using optimized q: -17.20566458893208
**********END EXPERIMENT**********
```

Unfortunately, the ELBO with the optimized q is not close to the log-likelihood. It is likely that I have a bug in the implementation. After much deliberation, I was unable to find the bug.