1. (**"Hard EM" for the k-means clustering algorithm**) Say we have observed data $\mathbf{x} = x_1, x_2, \cdots, x_N$, and we want to cluster this data into K clusters, with means $\mu_1, \cdots, \mu_K$. Let $\mathbf{z} = z_1, z_2, \cdots, z_N$ be our latent variables, which represent the cluster each data point belongs to. Each z_n is a vector, with $z_n^i = 1$ if x_n belongs to cluster i, and 0 otherwise.

   (a) ("E-step") Given an initial guess for the cluster means $\Theta = (\mu_1, \cdots, \mu_K)$, how can the cluster assignments z_n^i be updated?

   Solution: For a given data point x_n, we want to assign x_n to the cluster whose mean it is closest to (using Euclidean distance). Therefore:

   $$z_n^i = \begin{cases} 1, & \text{if} \quad i = \arg\min_j ||x_n - \mu_j||^2 \\ 0, & \text{otherwise.} \end{cases}$$

   (b) ("M-step") Given these new cluster assignments, how can the means be updated?

   Solution: The new mean of cluster i should be the average of all the points assigned to cluster i. Therefore:

   $$\mu_i^* = \frac{\sum_{n=1}^N z_n^i \cdot x_n}{\sum_{n=1}^N z_n^i}.$$

2. (**EM for a Gaussian mixture model**) Now our model for data generation will be a mixture of K Gaussians. To generate a data point, the ith Gaussian is chosen with probability π_i, then a data point is chosen according to the normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$. (Here we consider the one dimensional case, but this approach can easily be extended.)

   (a) What is the likelihood of our data?

   Solution: Let the parameters of this model be denoted $\Theta = \{\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \sigma_1, \cdots, \sigma_K\}$. Since each data point is generated independently, we can decompose the likelihood as:

   $$\mathbb{P}(\mathbf{x}|\Theta) = \prod_{n=1}^N \mathbb{P}(x_n|\Theta) = \prod_{n=1}^N \sum_{i=1}^K \mathbb{P}(x_n, z_n^i = 1|\Theta) = \prod_{n=1}^N \sum_{i=1}^K \mathbb{P}(z_n^i = 1|\Theta) \cdot \mathbb{P}(x_n|z_n^i = 1, \Theta).$$

   From our parameters, we know that $\mathbb{P}(z_n^i = 1|\Theta) = \pi_i$. Given Gaussian i, the likelihood of x_n is the associated normal distribution, denoted $\mathcal{N}(x_n|\mu_i, \sigma_i^2)$. Therefore:

   $$\mathbb{P}(\mathbf{x}|\Theta) = \prod_{n=1}^N \sum_{i=1}^K \pi_i \cdot \mathcal{N}(x_n|\mu_i, \sigma_i^2).$$

   (b) (E-step) Let τ_n^i be the posterior probability that x_n came from Gaussian i. Find a formula for τ_n^i.

   Solution: Using Bayes rule:

   $$\tau_n^i = \mathbb{P}(z_n^i = 1|x_n, \Theta) = \frac{\mathbb{P}(x_n|z_n^i = 1, \Theta) \cdot \mathbb{P}(z_n^i = 1|\Theta)}{\mathbb{P}(x_n|\Theta)} = \frac{\pi_i \cdot \mathcal{N}(x_n|\mu_i, \sigma_i^2)}{\sum_{j=1}^K \pi_j \cdot \mathcal{N}(x_n|\mu_j, \sigma_j^2)}.$$

(c) (M-step) Without taking derivatives, what should the updated parameters μ_i^*, σ_i^*, and π_i^* be?

Solution: The updated mean of Gaussian i should be an average of the data points, each weighted by the posterior probability they came from Gaussian i. To normalize we divide by the expected number of points generated by Gaussian i:

$$\mu_i^* = \frac{\sum_{n=1}^N \tau_n^i \cdot x_n}{\sum_{n=1}^N \tau_n^i}.$$

Similarly, we compute a weighted sample variance to find the updated σ_i^2's:

$$\sigma_i^{2*} = \frac{\sum_{n=1}^N \tau_n^i \cdot (x_n - \mu_i^*)^2}{\sum_{n=1}^N \tau_n^i}.$$

The updated weight for Gaussian i is the expected number of points it generated:

$$\pi_i^* = \frac{1}{N}\sum_{n=1}^N \tau_n^i.$$

(d) How does this fit into the equation for EM discussed during lecture?

Solution: We could have instead derived the E-step and M-step above using the EM equation. Let Θ_c be the current parameter values, and we want to find $\Theta^* = \arg\max_\Theta J(\Theta)$ where:

$$\begin{aligned}
J(\Theta) = \mathbb{E}[\log \mathbb{P}(\mathbf{x}, \mathbf{Z}|\Theta)|\mathbf{x}, \Theta_c] &= \mathbb{E}\left[\sum_{n=1}^N \log \mathbb{P}(x_n, Z_n|\Theta)\middle|\mathbf{x}, \Theta_c\right] \\
&= \mathbb{E}\left[\sum_{n=1}^N \log\left(\prod_{i=1}^K [\pi_i \cdot \mathcal{N}(x_n|\mu_i, \sigma_i^2)]^{Z_n^i}\right)\middle|\mathbf{x}, \Theta_c\right] \\
&= \mathbb{E}\left[\sum_{n=1}^N \sum_{i=1}^K Z_n^i[\log \pi_i + \log \mathcal{N}(x_n|\mu_i, \sigma_i^2)]\middle|\mathbf{x}, \Theta_c\right] \\
&= \sum_{n=1}^N \sum_{i=1}^K \mathbb{E}[Z_n^i|x_n, \Theta_c] \cdot [\log \pi_i + \log \mathcal{N}(x_n|\mu_i, \sigma_i^2)] \\
&= \sum_{n=1}^N \sum_{i=1}^K \mathbb{P}(z_n^i = 1|x_n, \Theta_c) \cdot [\log \pi_i + \log \mathcal{N}(x_n|\mu_i, \sigma_i^2)] \\
&= \sum_{n=1}^N \sum_{i=1}^K \tau_n^i \cdot [\log \pi_i + \log \mathcal{N}(x_n|\mu_i, \sigma_i^2)].
\end{aligned}$$

Now we can take the derivative of $J(\Theta)$ with respect to each one of our parameter types:

$$\frac{\partial J(\Theta)}{\partial \mu_i} = \sum_{n=1}^N \tau_n^i \left(\frac{-(x_n - \mu_i)}{\sigma_i^2}\right) = 0 \quad \Rightarrow \quad \mu_i^* = \frac{\sum_{n=1}^N \tau_n^i \cdot x_n}{\sum_{n=1}^N \tau_n^i}.$$

$$\frac{\partial J(\Theta)}{\partial \sigma_i} = \sum_{n=1}^{N} \tau_n^i \left(-\sigma_i + \frac{(x_n - \mu_i)^2}{\sigma_i} \right) = 0 \quad \Rightarrow \quad \sigma_i^{2*} = \frac{\sum_{n=1}^{N} \tau_n^i \cdot (x_n - \mu_i^*)^2}{\sum_{n=1}^{N} \tau_n^i}.$$

For the π_i terms, we need to add a Lagrange multiplier to ensure they sum to 1:

$$J'(\Theta) = \sum_{n=1}^{N} \sum_{i=1}^{K} \tau_n^i \cdot [\log \pi_i + \log \mathcal{N}(x_n | \mu_i, \sigma_i^2)] + \lambda \left(1 - \sum_{i=1}^{K} \pi_i \right).$$

Then we can take the derivative of $J'(\Theta)$ with respect to λ and each π_i:

$$\frac{\partial J'(\Theta)}{\partial \lambda} = 1 - \sum_{i=1}^{K} \pi_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{K} \pi_i = 1$$

$$\frac{\partial J'(\Theta)}{\partial \pi_i} = \sum_{n=1}^{N} \frac{\tau_n^i}{\pi_i} - \lambda = 0 \quad \Rightarrow \quad \pi_i = \frac{1}{\lambda} \sum_{n=1}^{N} \tau_n^i.$$

Solving, we obtain:

$$\sum_{i=1}^{K} \pi_i = \sum_{i=1}^{K} \frac{1}{\lambda} \sum_{n=1}^{N} \tau_n^i = \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{i=1}^{K} \tau_n^i = \frac{1}{\lambda} \sum_{n=1}^{N} 1 = 1 \quad \Rightarrow \quad \lambda^* = N,$$

and therefore:

$$\pi_i^* = \frac{1}{N} \sum_{n=1}^{N} \tau_n^i.$$