Homework-5

1(a) Given

$$l(\theta) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{K} P(x^{(i)} \mid \mu_j, C_j)\, P(j)\right]$$

$$\frac{\partial}{\partial \mu_j} l(\theta) = \sum_{i=1}^{n} \frac{P(j) N(x^{(i)} \mid \mu_j, C_j) \nabla_{\mu_j} \left\{-\frac{1}{2}(x^{(i)} - \mu_j)^T C_j^{-1} (x^{(i)} - \mu_j)\right\}}{\sum_{j'=1}^{K} P(j') N(x^{(i)} \mid \mu_{j'}, C_{j'})}$$

$$= \sum_{i=1}^{n} P_{i,j} C_j^{-1} (x^{(i)} - \mu_j)$$

1(b) $$\frac{\partial}{\partial P(j)} l(\theta) = \sum_{i=1}^{n} \frac{N(x^{(i)} \mid \mu_j, C_j)}{\sum_{j'=1}^{K} P(j') N(x^{(i)} \mid \mu_{j'}, C_{j'})}$$

$$= \frac{1}{P(j)} \sum_{i=1}^{n} P_{i,j} \quad [\text{Given } P_{i,j} = P(j \mid x^{(i)})]$$

No, it will not be a valid probability distribution

1(c) $\frac{\partial}{\partial P(j)} l(\theta) = \frac{1}{P(j)} \sum_{i=1}^{n} P_{i,j}$

$$P(j) = \frac{\exp(w_j)}{\sum_{j'=1}^{K} \exp(w_{j'})}$$

using lagrange multiplier,

$$P(j) = \frac{\sum_{i=1}^{n} P_{i,j}}{\lambda}$$

Summing over j and normalizing,

$$P(j) = \frac{\sum_{i=1}^{n} P_{i,j}}{N}$$

$$\Rightarrow \frac{\sum_{i=1}^{n} P_{i,j}}{P(j)} = N$$

$$\therefore \frac{\partial}{\partial w_j} l(\theta) \propto \sum_{i=1}^{n} P_{i,j} - P(j)$$

Problem 1(d):

$$\frac{\partial}{\partial C_j} l(\theta) = \sum_{i=1}^{n} \frac{P(j) \nabla_{C_j} N(x^{(i)} | \mu_j, C_j)}{\sum_{j'=1}^{K} P(j') N(x^{(i)} | \mu_{j'}, C_{j'})}$$

Here $\nabla_{C_j} N(x | \mu_j, C_j) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|C_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T C_j^{-1}(x-\mu_j)\right\}$

$\nabla_{C_j} \cdot \{\nabla_{C_j}(-\frac{1}{2}(x-\mu_j) C_j^{-1}(x-\mu_j)) -$

$C_j^{-1} \nabla_{C_j} |C_j|\}$

$= N(x | \mu_j, C_j) \nabla(\log N(x | \mu_j, C_j))$

So, $$C_j = \frac{\sum_{i=1}^{n} P_{i,j} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} P_{i,j}}$$

No, the result won't stay positive.

1(e)

Advantages:

1) In EM-algorithm the proposed parameter values are always valid for example, probability masses between [0,1] sums to 1, which is not in the cases of gradient descent.

2) In EM-algorithm we don't have to calculate the likelihood to insure it has increased at every step which is not in the case while gradient descent.

3) EM method exploits structure of the objective and the variable involved in a manner that they are largely decoupled which allows good convergence rate than gradient descent.