HW2 Answer

Problem 1

Denoted $\pi = (\pi_1, \cdots, \pi_K)$
The probability of one data point $\mathbf{x}_n$ is

$$p(\mathbf{x}_n, \mathbf{t}_n) = p(\mathbf{x}_n|\mathbf{t}_n)p(\mathbf{t}_n) = \prod_{k=1}^{K} (p(\mathbf{x}_n|C_k)\pi_k)^{t_n^k}$$

So the likelihood function is given by

$$p(\mathbf{x}_n, \mathbf{t}_n|\pi_k) = \prod_{n=1}^{N} \prod_{k=1}^{K} (p(\mathbf{x}_n|C_k)\pi_k)^{t_n^k}$$

and taking the logarithm, we get

$$\log p(\mathbf{x}_n, \mathbf{t}_n|\pi_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_n^k (\log p(\mathbf{x}_n|C_k) + \log \pi_k)$$

Note that $\log$ is denoted by natural log.
Now, we can formalize our maximize likelihood problem as an optimization problem:

$$\begin{aligned} \max_{\pi} \quad & \log p(\mathbf{x}_n, \mathbf{t}_n|\pi) \\ \text{subject to} \quad & \sum_{k=1}^{K} \pi_k = 1 \end{aligned}$$

By introducing a Lagrange multiplier λ and maximizing

$$\mathcal{L}(\pi, \lambda) = \log p(\mathbf{x}_n, \mathbf{t}_n|\pi) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

Taking the derivative with respect to π_k and setting it to 0, we have

$$\begin{aligned}\frac{\partial}{\partial \pi_k}\mathcal{L}(\pi,\lambda) &= \frac{\partial}{\partial \pi_k}\{\sum_{n=1}^{N}\sum_{k=1}^{K} t_n^k(\log p(\mathbf{x}_n|C_k)+\log \pi_k)+\lambda(\sum_{k=1}^{K}\pi_k - 1)\} \\ &= \frac{1}{\pi_k}\sum_{n=1}^{N} t_n^k + \lambda = 0 \\ &\Rightarrow \pi_k = -\frac{1}{\lambda}\sum_{n=1}^{N} t_n^k = -\frac{N_k}{\lambda}\end{aligned}$$

where N_k us the number of data points whose label is class k. Taking the derivative with respect to λ, we have

$$\begin{aligned}&\frac{\partial}{\partial \lambda}\mathcal{L}(\pi,\lambda) = \sum_{k=1}^{K}\pi_k - 1 = 0 \\ &\Rightarrow \sum_{k=1}^{K}\pi_k = 1\end{aligned}$$

Combining two equations, we get

$$\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K} -\frac{N_k}{\lambda} = -\frac{N}{\lambda} \Rightarrow \lambda = -N$$

Finally, we can put it back into our equation to solve π_k's. Thus, we have

$$\pi_k = \frac{N_k}{N}$$

Problem 2

1. Conseider $f(\mathbf{w}) = \mathbf{w}^T A\mathbf{w}$, and $f(\mathbf{w}+h) = (\mathbf{w}+h)^T A(\mathbf{w}+h)$
   Then,

$$\begin{aligned}f(\mathbf{w}+h) - f(\mathbf{w}) &= \mathbf{w}^T A\mathbf{w} + h^T A\mathbf{w} + \mathbf{w}^T Ah + h^T Ah - \mathbf{w}^T A\mathbf{w} \\ &= h^T A\mathbf{w} + \mathbf{w}^T Ah + h^T Ah \\ &= h^T(A^T\mathbf{w} + A\mathbf{w}) + h^T Ah \\ &= (A^T\mathbf{w} + A\mathbf{w})\cdot h + h^T Ah\end{aligned}$$

By definition, $\frac{\partial\, \mathbf{w}^T A \mathbf{w}}{\partial\, \mathbf{w}} = A^T \mathbf{w} + A\mathbf{w}$. In particular, A is a symmetric matrix i.e. $A = A^T$, then $\frac{\partial\, \mathbf{w}^T A \mathbf{w}}{\partial\, \mathbf{w}} = 2A\mathbf{w}$

2. Define $C = AB$. Note that $c_{ij} := \sum_{k=1}^m a_{ik} b_{kj}$, where c_{ij} is the i-th row and j-th columns of matrix C. i.e. the dot product of the i-th row of A and the j-th column of B.
   Then,

$$tr(AB) = tr(C) = \sum_{l=1}^{m} c_{ll} = \sum_{l=1}^{m} \sum_{k=1}^{m} a_{lk} b_{kl}$$

   Hence,

$$\frac{\partial\, tr(AB)}{\partial a_{ij}} = \frac{\partial\, \sum_{l=1}^m \sum_{k=1}^m a_{lk} b_{kl}}{\partial a_{ij}} = b_{ji}$$

3. Follow the hint

Problem 3

By assumption, we know

$$p\left(x_n | C_k\right) = \frac{1}{(2\pi)^{D/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\left(x_n - \boldsymbol{\mu}_k\right)^T \mathbf{\Sigma}^{-1}\left(x_n - \boldsymbol{\mu}_k\right)\right)$$
$$p(x_n, t_n) = p(x_n | t_n) p(t_n) = \prod_{k=1}^{K} (p(x_n | C_k)\pi_k)^{t_n^k}$$

where D is dimension of x.
The log-likelihood function is given by

$$\log p(x_n, t_n | \pi_k, \boldsymbol{\mu_k}, \mathbf{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_n^k (\log p(x_n | C_k) + \log \pi_k)$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} t_n^k (\log \pi_k + (-\frac{1}{2}\left(x_n - \boldsymbol{\mu}_k\right)^T \mathbf{\Sigma}^{-1}\left(x_n - \boldsymbol{\mu}_k\right)) - \frac{1}{2} \log|\mathbf{\Sigma}| - \frac{D}{2} \log 2\pi)$$

By introducing a Lagrange multiplier λ and maximizing

$$\mathcal{L}(\pi, \boldsymbol{\mu_k}, \boldsymbol{\Sigma}, \lambda) = \log p(x_n, t_n | \pi_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

Taking the derivative with respect to π_k and setting it to 0, we have

$$\pi_k = \frac{N_k}{N} (\text{ follow problem 1})$$

Taking the derivative with respect to $\boldsymbol{\mu_k}$ and setting it to 0, we have

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu_k}} \mathcal{L}(\pi, \boldsymbol{\mu_k}, \boldsymbol{\Sigma}, \lambda) &= \sum_{n=1}^{N} t_n^k \boldsymbol{\Sigma}^{-1} (x_n - \boldsymbol{\mu_k}) \\
&= \boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} t_n^k (x_n - \boldsymbol{\mu_k}) = 0
\end{aligned}$$

Since $\boldsymbol{\Sigma}^{-1}$ is positive definite, then

$$\begin{aligned}
\sum_{n=1}^{N} t_n^k (x_n - \boldsymbol{\mu_k}) &= \sum_{n=1}^{N} t_n^k x_n - \boldsymbol{\mu_k} \sum_{n=1}^{N} t_n^k = 0 \\
\Rightarrow \boldsymbol{\mu_k} &= \frac{\sum_{n=1}^{N} t_n^k x_n}{\sum_{n=1}^{N} t_n^k} = \frac{\sum_{n=1}^{N} t_n^k x_n}{N_k}
\end{aligned}$$

We rewrite the log-likelihood function:

$$\begin{aligned}
\mathcal{L}(\pi, \boldsymbol{\mu_k}, \boldsymbol{\Sigma}, \lambda) &= \sum_{n=1}^{N} \sum_{k=1}^{K} t_n^k (\{-\frac{1}{2} (x_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (x_n - \boldsymbol{\mu}_k)\}) - \frac{1}{2} \log |\boldsymbol{\Sigma}|) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_n^k (\{-\frac{1}{2} tr\{(x_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (x_n - \boldsymbol{\mu}_k)\})\} + \frac{1}{2} \log |\boldsymbol{\Sigma^{-1}}|) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_n^k (\{-\frac{1}{2} tr\{\boldsymbol{\Sigma}^{-1} (x_n - \boldsymbol{\mu}_k)^T (x_n - \boldsymbol{\mu}_k)\})\} + \frac{1}{2} \log |\boldsymbol{\Sigma^{-1}}|)
\end{aligned}$$

Taking the derivative with respect to $\boldsymbol{\Sigma^{-1}}$ and setting it to 0, we get

$$\begin{aligned}\frac{\partial}{\partial \mathbf{\Sigma^{-1}}}\mathcal{L}(\pi, \boldsymbol{\mu_k}, \mathbf{\Sigma}, \lambda) &= \frac{-1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_n^k\left(x_n-\boldsymbol{\mu}_k\right)\left(x_n-\boldsymbol{\mu}_k\right)^T - t_n^k\mathbf{\Sigma}^T \\ &= \frac{-1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} t_n^k\left(x_n-\boldsymbol{\mu}_k\right)\left(x_n-\boldsymbol{\mu}_k\right)^T - \sum_{k=1}^{K}\sum_{n=1}^{N} t_n^k\mathbf{\Sigma} = 0\end{aligned}$$

Hence,

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{n=1}^{N} t_n^k\mathbf{\Sigma} = \sum_{k=1}^{K}\sum_{n=1}^{N} t_n^k\left(x_n-\boldsymbol{\mu}_k\right)\left(x_n-\boldsymbol{\mu}_k\right)^T \\ &\Rightarrow N\mathbf{\Sigma} = \sum_{k=1}^{K} N_k\mathbf{\Sigma_k} \Rightarrow \mathbf{\Sigma} = \sum_{k=1}^{K}\frac{N_k}{N}\mathbf{\Sigma}_k\end{aligned}$$

Note:

1. By the invariance property of MLE, we take derivative of $\mathcal{L}$ w.r.t. Σ^{-1}
2. If you want to take derivative of $\mathcal{L}$ w.r.t. Σ, then you would apply the fact that $\frac{\partial\, tr\left(X^{-1}M\right)}{\partial X} = -\left(X^{-1}MX^{-1}\right)$

Problem 4

1.

$$\begin{aligned}\sum_{i=1}^{m}\left\|\boldsymbol{z}^i-\boldsymbol{z}\right\|^2 &= \sum_{i=1}^{m}\left\|\left(\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right)+\left(\overline{\boldsymbol{z}}-\boldsymbol{z}\right)\right\|^2 \\ &= \sum_{i=1}^{m}\left(\left\|\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right\|^2+\|\overline{\boldsymbol{z}}-\boldsymbol{z}\|^2+2\left(\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right)\cdot\left(\overline{\boldsymbol{z}}-\boldsymbol{z}\right)\right) \\ &= \sum_{i=1}^{m}\left\|\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right\|^2+\sum_{i=1}^{m}\|\overline{\boldsymbol{z}}-\boldsymbol{z}\|^2+2\sum_{i=1}^{m}\left(\boldsymbol{z}^i\cdot\overline{\boldsymbol{z}}-\boldsymbol{z}^i\cdot\boldsymbol{z}-\overline{\boldsymbol{z}}\cdot\overline{\boldsymbol{z}}+\overline{\boldsymbol{z}}\cdot\boldsymbol{z}\right) \\ &= \sum_{i=1}^{m}\left\|\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right\|^2+m\|\overline{\boldsymbol{z}}-\boldsymbol{z}\|^2+2(m\overline{\boldsymbol{z}}\cdot\overline{\boldsymbol{z}}-m\overline{\boldsymbol{z}}\cdot\boldsymbol{z}-m\overline{\boldsymbol{z}}\cdot\overline{\boldsymbol{z}}+m\overline{\boldsymbol{z}}\cdot\boldsymbol{z}) \\ &= \sum_{i=1}^{m}\left\|\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right\|^2+m\|\overline{\boldsymbol{z}}-\boldsymbol{z}\|^2 \\ &\geq \sum_{i=1}^{m}\left\|\boldsymbol{z}^i-\overline{\boldsymbol{z}}\right\|^2.\end{aligned}$$

2. It follows directly from the logic of the algorithm: $\mathcal{C}^t$ and $\mathcal{C}^{t+1}$ are different only if there is a point that finds a closer cluster center in $\boldsymbol{\mu}^t$ than the one assigned to it by $\mathcal{C}^t$:

$$L\left(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t\right) = \sum_{i=1}^{n} \left\| \boldsymbol{x}^i - \boldsymbol{\mu}^t_{C^{t+1}(i)} \right\|^2 < \sum_{i=1}^{n} \left\| \boldsymbol{x}^i - \boldsymbol{\mu}^t_{C^t(i)} \right\|^2 = L\left(\mathcal{C}^t, \boldsymbol{\mu}^t\right)$$

3. Use the result in (1):

$$\begin{aligned} L\left(\mathcal{C}^{t+1}, \boldsymbol{\mu}^{t+1}\right) &= \sum_{i=1}^{n} \left\| \boldsymbol{x}^i - \boldsymbol{\mu}^{t+1}_{C^{t+1}(i)} \right\|^2 \\ &= \sum_{k'=1}^{k} \sum_{i \in \{1,2,\ldots,n\}, \mathcal{C}^{t+1}(i)=k'} \left\| \boldsymbol{x}^i - \boldsymbol{\mu}^{t+1}_{C^{t+1}(i)} \right\|^2 \\ &\leq \sum_{k'=1}^{k} \sum_{i \in \{1,2,\ldots,n\}, \mathcal{C}^{t+1}(i)=k'} \left\| \boldsymbol{x}^i - \boldsymbol{\mu}^{t}_{C^{t+1}(i)} \right\|^2 (\text{by } 1) \\ &= \sum_{i=1}^{n} \left\| \boldsymbol{x}^i - \boldsymbol{\mu}^{t}_{C^{t+1}(i)} \right\|^2 \\ &= L\left(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t\right). \end{aligned}$$

4. Define the sequence $\{l_t\}$, where $l_t = L(\mathcal{C}^t, \boldsymbol{\mu}^t)$. By previous result, we have

$$l_t = L(\mathcal{C}^t, \boldsymbol{\mu}^t) \leq L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^{t+1}) = l_{t+1}$$

for all t. Hence, $\{l_t\}$ is a monotonic decreasing sequence.
Note that we apply **monotonic convergence theorem of sequence** to prove the sequence is convergence, which does not guarantee this algorithm could find the **global** minimum, just a **local** minimum.

5. There are at most k^N ways to partition N data points into k clusters. Then, this algorithm converges in finitely many steps.
   Note that the upper bound (k^N) may not tight.