Problem 1

(a)

Model: $f_i = f(x_i) + \varepsilon$

Where we assume zero-mean i.i.d Gaussian nose: $\varepsilon \sim N(0, \sigma_i^2)$

So the predictive distribution given the model parameter is Gaussian: $p(f_i \mid x_i, w) = N(f(x_i), \sigma_i^2)$

Because the training data is nose free: $f_i = f(x_i)$ and $\varepsilon = 0$

We have $cov_{f_i} = 0$

(b)

Prior

$$p(f \mid \theta) = GP\big(f; \mu_\theta(x), K_\theta(x, x')\big)$$

The probability of observing the given data under the prior

$$p(y \mid X, \theta) = \int p(y \mid f)\, p(f \mid X, \theta)\, df$$

Where we have marginalized the unknown function values of f

$$p(y \mid X, \theta) = \int p(y \mid f)\, p(f \mid X, \theta)\, df$$

$$= \int (iid\ noise)(GP\ prior) df$$

$$= \int N(y; f, \sigma^2 I) N\big(f; \mu_\theta(X), K_\theta(X, X)\big)\, df$$

$$= N(y; \mu_{\theta(X)}, K_\theta(X, X) + \sigma^2 I)$$

Let $K_y = K_\theta(X, X) + \sigma^2 I$, the log-likelihood of the data under the Gaussian prior is

$$\log p(y \mid X, \theta) = -\frac{(y-\mu)^T K_y^{-1} (y-\mu)}{2} - \frac{\log\det K_y}{2} - \frac{N \log 2\pi}{2}$$

(c)

With Choleskey decomposition of $K_y = LL^T$ and $\alpha = L^T \setminus (L \setminus y)$ and $\mu = 0$

$$\log p(y \mid X, \theta) = -\frac{(y-\mu)^T K_y^{-1} (y-\mu)}{2} - \frac{\log\det K_y}{2} - \frac{N \log 2\pi}{2}$$

$$= -\frac{y^T\, (LL^T)^{-1} y}{2} - \frac{\log\det(LL^T)}{2} - \frac{N \log 2\pi}{2}$$

$$= -\frac{y^T \alpha}{2} - \sum_i \log L_{ii} - \frac{N \log 2\pi}{2}$$

(d)

The matrix derivation is

$$\frac{\partial}{\partial\theta}K^{-1} = -K^{-1}\frac{\partial K}{\partial\theta}K^{-1}$$

The log determinant derivation is

$$\frac{\partial}{\partial\theta}\log K = tr(K^{-1}\frac{\partial K}{\partial\theta})$$

The derivative of the log likelihood is then

$$\frac{\partial \log p(y\,|X,\theta)}{\partial\theta} = \frac{1}{2}y^T K_y^{-1}\frac{\partial K_y}{\partial\theta}K_y^{-1}y\; - \frac{1}{2}Tr(K_y^{-1}\frac{\partial K_y}{\partial\theta})$$

With Cholesky Decomposition $K_y = LL^T$

$$\partial K_y = \partial LL^T + L\partial L^T$$

$$\frac{\partial \log p(y\,|X,\theta)}{\partial\theta} = \frac{1}{2}y^T(LL^T)^{-1}\frac{\partial LL^T + L\partial L^T}{\partial\theta}\alpha - \frac{1}{2}Tr((LL^T)^{-1}\frac{\partial LL^T + L\partial L^T}{\partial\theta})$$

Problem 2

(a)

These two conditions are equivalent to each other

At certain iteration we have the cluster centers $\mu_\alpha = \frac{\sum_{i=1}^{n}[z_i]_\alpha x_i}{\sum_{i=1}^{n}[z_i]_\alpha}$

After next iteration

1. Assignment does not change: $z_i' = z_i \rightarrow \mu_\alpha' = \frac{\sum_{i=1}^{n}[z_i']_\alpha x_i}{\sum_{i=1}^{n}[z_i']_\alpha} = \frac{\sum_{i=1}^{n}[z_i]_\alpha x_i}{\sum_{i=1}^{n}[z_i]_\alpha} = \mu_\alpha$
2. Cluster center does not change: $\mu_\alpha' = \mu_\alpha$

   This suggest that no update happens at the reassignment stage because

   1. Any point assigned out of the current cluster must be replaced by a new point to keep the centroid unchanged, and they are the same point. That is, there is no new assignment
   2. All the points assigned out of the current cluster must have the same centroid as the current cluster. Then whichever cluster else the points are assigned to must have the same centroid. That is, there is no new assignment

   Therefore when the centroid does not change, it suggest that assignment does not change.

Therefore, there two conditions are equivalent to each other

(b)

Yes, the k-means algorithm always converge.

First, there are at most k^n ways to partition n data points into k clusters. At each iteration the algorithm generates a new clustering based on the old clustering. The assignment of the data points changes if and only if the data points find a closer cluster center.

Therefore we have:

1. If the new clustering is the same as the old clustering, then the next new clustering is the same because the assignment of data points are already to their closest centers and the cluster centers remained unchanged.
2. If the new clustering is different than the old clustering, the new clustering has a smaller D because one or more data points are assigned to closer cluster centers

That is, the algorithm search through finite space to find an optimal solution and hence it will eventually converge.

(c)

Yes, it is possible that k-means algorithm generates empty clusters.

At each iteration, the cluster centers are updated with new data point assignments. It is possible that at next iteration, all data points within one cluster c_i is now closer to other cluster centers and therefore

get assigned to other clusters, while no other data points is closer to c_i. This results in cluster c_i losing all its members and becoming an empty cluster.

(d)

Yes, it is possible to find non-convex clusters by the k-means algorithm.

Apply the kernel trick to kernelize the k-means algorithm. The inner product of x_i and x_j is replaced with a kernel function and the kernelized k-means algorithm can find non-convex clusters without explicitly computing the transformation from input space to feature space.

Problem 3

(a)

Probability of data point x from j^{th} distribution

$$P(\theta_j) = \pi_j$$

$$P(x|\theta_j) \sim N(\mu_j, \Sigma_j)$$

$$P(x, \theta_j) = P(x|\theta_j)P(\theta_j)$$

$$P(\theta_j|x) = \frac{P(x|\theta_j)P(\theta_j)}{P(x)} \propto P(x|\theta_j)P(\theta_j)$$

(b)

For any data point we have the likelihood

$$g(x|\Theta) = \sum_{j=1}^{k} \pi_j \, p(x|\theta_j)$$

For all the data points we have the likelihood

$$L(X|\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{k} \pi_j \, p(x_i|\theta_j)$$

(c)

E step:

From the problem 1 we have the likelihood

$$L(X, Z|\Theta) = \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{k} Z_j \left(\log \pi_i - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j (x_i - \mu_j) - \frac{n}{2}\log 2\pi\right)\right)$$

The distribution over Z is

$$T_{ij}^t = P(Z_i = j|X_i = x_i|\theta^t) = \frac{g(x_i|\theta_j^t)}{\sum_{all\ j} g(x_i|\theta_j^t)}$$

We have this Q function

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t}[\log L(X, Z|\Theta)] = E_{Z|X,\theta^t}\left[\log \prod_{i=1}^{N} L(x_i, z_i|\Theta)\right]$$

$$= E_{Z|X,\theta^t}\left[\sum_{i=1}^{N} \log L(x_i, z_i|\Theta)\right] = \sum_{i=1}^{N} E_{Z|X,\theta^t}[\log L(x_i, z_i|\Theta)]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} T_{ij}^{t}\left(\log \pi_i - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i-\mu_j)^T\Sigma_j(x_i-\mu_j) - \frac{n}{2}\log 2\pi\right)$$

Parameters is already known before the step except T_{ij} which is computed at the beginning of the step

M step:

Updating the values for parameters

$$\pi_i^{t+1} = \arg\max Q(\theta|\theta^t) = \arg\max_{\mathrm{j}}\left\{\left(\sum_{i=1}^{n} T_{ij}^{t}\right)\log \pi_i\right\}$$

$$\left(\mu_i^{t+1}, \Sigma_i^{t+1}\right) = \underset{\mu_i^{t+1},\Sigma_i^{t+1}}{\operatorname{argmax}} Q(\theta|\theta^t) = \underset{\mu_i^{t+1},\Sigma_i^{t+1}}{\operatorname{argmax}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ij}^{t}\left(-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i-\mu_j)^T\Sigma_j(x_i-\mu_j)\right)\right\}$$

Convergence:

$$E_{z|\theta^t,x}[\log L(X,Z|\Theta^t)] \le E_{z|\theta^{t-1},x}[\log L(X,Z|\Theta^{t-1})] + \varepsilon \to exit$$

(d)

Because we have the components as Bernoulli distributions

$$p(x_i|\theta_j) = \prod_{m=1}^{d}(\theta_{jm})^{x_{im}}(1-\theta_{jm})^{1-x_{im}}$$

With the density over Z is

$$T_{ij}^{t} = P(Z_i = j|X_i = x_i|\theta^t) = \frac{g(x_i|\theta_j^t)}{\sum_{all\ j} g(x_i|\theta_j^t)}$$

We have

$$\pi_i^{t+1} = \frac{\sum_{j=1}^{n} T_{ij}^{t}}{\sum_{i=1}^{n}\sum_{j=1}^{n} T_{ij}^{t}} = \frac{1}{n}\sum_{i=1}^{n} T_{ij}^{t}$$

$$\mu_i^{t+1} = \frac{\sum_{j=1}^{n} T_{ij}^{t}\, x_i}{\sum_{i=1}^{n}\sum_{j=1}^{n} T_{ij}^{t}}$$

$$\Sigma_i^{t+1} = \frac{\sum_{j=1}^{k} T_{ij}^{t}\left((x_j-\mu_i^{t+1})(x_j-\mu_i^{t+1})^T\right)}{\sum_{j=1}^{k} T_{ij}^{t}}$$

So the parameter μ_i and Σ_i are all weighted MLE with Bernoulli distributions