Solution to Probability Homework 2

1. According to the Binomial Theorem, $(x+y)^n = \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} x^{n-k} y^k$. Therefore,

$$[(1-p)+p]^N = \sum_{m=0}^{N} \begin{pmatrix} N \\ m \end{pmatrix} (1-p)^{N-m} p^m.$$

Since the binomial distribution is given by $\text{Bin}(m|N,p) = \begin{pmatrix} N \\ m \end{pmatrix} p^m (1-p)^{N-m}$, it follows that

$$\sum_{m=0}^{N} \text{Bin}(m|N,p) = \sum_{m=0}^{N} \begin{pmatrix} N \\ m \end{pmatrix} p^m (1-p)^{N-m} = [(1-p)+p]^N = 1.$$

2. We show that for multivariate Gaussian, the conjugate prior on the mean of Gaussian is also Gaussian as the univariate case can be derived by setting the dimension to 1. Let $\boldsymbol{\mu} \in \mathbb{R}^B$, $\boldsymbol{\Sigma} \in \mathbb{R}^{B\times B}$ be the mean and covariance matrix for the Gaussian distribution

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{B/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Given data $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, the likelihood is

$$p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{BN/2}|\boldsymbol{\Sigma}|^{N/2}} e^{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}.$$

Define a prior for $\boldsymbol{\mu}$ as

$$p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = \frac{1}{(2\pi)^{B/2}|\boldsymbol{\Sigma}_0|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}.$$

Then the posterior is

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\Sigma},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) &\propto p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Sigma})\, p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\ &= \frac{1}{(2\pi)^{BN/2}|\boldsymbol{\Sigma}|^{N/2}(2\pi)^{B/2}|\boldsymbol{\Sigma}_0|^{1/2}} e^{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}. \end{aligned}$$

Since

$$\begin{aligned}
&\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\
=& C_1 - 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \mathbf{x}_i + \boldsymbol{\mu}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + N\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\
=& C_1 + \boldsymbol{\mu}^T \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \mathbf{x}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right) \\
=& \left[\boldsymbol{\mu} - \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \mathbf{x}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right)\right]^T \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right) \\
& \left[\boldsymbol{\mu} - \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \mathbf{x}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right)\right] + C_2
\end{aligned}$$

where C_1 and C_2 are constants involving $\boldsymbol{\Sigma}$, $\mathbf{x}_i$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, the posterior is also a Gaussian

$$p\left(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \propto e^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}-\boldsymbol{\mu}_1)},$$

where

$$\begin{aligned}
\boldsymbol{\mu}_1 &= \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \mathbf{x}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right), \\
\boldsymbol{\Sigma}_1 &= \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}.
\end{aligned}$$

3. Given prior $\text{Dir}\left(\mathbf{p}|\left(2,4,3\right)\right) \propto p_1 p_2^3 p_3^2$ where $\mathbf{p} = \left(p_1, p_2, p_3\right)^T$ and likelihood $\text{Mult}\left(\left(8,3,2\right)|\mathbf{p}\right) \propto p_1^8 p_2^3 p_3^2$, the posterior is

$$\begin{aligned}
f\left(\mathbf{p}|\left(8,3,2\right),\left(2,4,3\right)\right) &\propto \text{Mult}\left(\left(8,3,2\right)|\mathbf{p}\right) \text{Dir}\left(\mathbf{p}|\left(2,4,3\right)\right) \\
&= p_1^9 p_2^6 p_3^4.
\end{aligned}$$

4. The log-likelihood function of a Gaussian given independent samples $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ is

$$\begin{aligned}
\log p\left(X|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) &= \log \prod_{i=1}^{N} p\left(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{i=1}^{N} \log p\left(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
&= \sum_{i=1}^{N} \left\{-\frac{B}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right\} \\
&= -\frac{NB}{2}\log 2\pi - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right).
\end{aligned}$$

5. Let $p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$ where $\mathbf{x} = (x_1, ..., x_K)$. The log-likelihood function given independent samples $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ is

$$\log(p(X|\boldsymbol{\mu})) = \log\left(\prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu})\right) = \log\prod_{i=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{ik}} = \log\prod_{k=1}^{K}\mu_k^{m_k}$$

where $m_k = \sum_{i=1}^{N} x_{ik}$. So $\log(p(X|\boldsymbol{\mu})) = \sum_{k=1}^{K} m_k \log\mu_k$.

6. The objective function is

$$J(\{\mathbf{c}_m\}, \sigma) = \sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} w_m e^{-\frac{1}{2\sigma^2}\|\mathbf{x}_n - \mathbf{c}_m\|^2}\right)^2.$$

Taking the derivative of J with respect to $\mathbf{c}_m$ gives

$$\frac{\partial J}{\partial \mathbf{c}_m} = \frac{2w_m}{\sigma^2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} w_m K(\mathbf{x}_n, \mathbf{c}_m)\right) K(\mathbf{x}_n, \mathbf{c}_m)(\mathbf{c}_m - \mathbf{x}_n).$$

Taking the derivative of J with respect to σ, we have

$$\frac{\partial J}{\partial \sigma} = -2\sum_{n=1}^{N}\left\{\left(y_n - \sum_{m=1}^{M} w_m K(\mathbf{x}_n, \mathbf{c}_m)\right)\left(\sum_{m=1}^{M} w_m K(\mathbf{x}_n, \mathbf{c}_m)\|\mathbf{x}_n - \mathbf{c}_m\|^2\sigma^{-3}\right)\right\}.$$

The update formulas for $\mathbf{c}_m$ and σ are

$$\mathbf{c}_m \leftarrow \mathbf{c}_m - \Delta t\frac{\partial J}{\partial \mathbf{c}_m}$$
$$\sigma \leftarrow \sigma - \Delta t\frac{\partial J}{\partial \sigma}.$$

7. Let $\mathbf{z} = (x, y)^T$, $J(\mathbf{z}) = \mathbf{z}^T\mathbf{A}\mathbf{z}$ where $\mathbf{A} = \text{diag}(4, 9)$. Using Lagrange multiplier, we have

$$L(\mathbf{z}) = \mathbf{z}^T\mathbf{A}\mathbf{z} - \lambda\left(\mathbf{z}^T\mathbf{z} - 1\right).$$

Taking the derivative of L with respect to $\mathbf{z}$ and setting it to zero, we have

$$\frac{\partial L}{\partial \mathbf{z}} = 2\mathbf{A}\mathbf{z} - 2\lambda\mathbf{z} = 0,$$

i.e. $\mathbf{A}\mathbf{z} = \lambda\mathbf{z}$. Hence $\mathbf{z}$ is an eigenvector of $\mathbf{A}$ and λ is the corresponding eigenvalue. Plugging $\mathbf{A}\mathbf{z} = \lambda\mathbf{z}$ into the original $J(\mathbf{z})$, we have

$$J(\mathbf{z}) = \mathbf{z}^T\lambda\mathbf{z} = \lambda.$$

Hence the minimum of J subject to the constraint is the smallest eigenvalue of $\mathbf{A}$, i.e. 4, the $\mathbf{z}$ that minimizes J is the corresponding eigenvector, i.e. $\mathbf{z} = (1, 0)^T$, or $\mathbf{z} = (-1, 0)^T$.