**ECE 645: Estimation Theory**
**Spring 2015**
**Instructor: Prof. Stanley H. Chan**

# Lecture 10: Expectation-Maximization Algorithm

(LaTeX prepared by Shaobo Fang)
May 4, 2015

This lecture note is based on ECE 645 (Spring 2015) by Prof. Stanley H. Chan in the School of Electrical and Computer Engineering at Purdue University.

## 1 Motivation

Consider a set of data points with their classes labeled, and assume that each class is a Gaussian as shown in Figure 1(a). Given this set of data points, finding the means of two Gaussian can be done easily by estimating the sample mean, as the class labels are known.

Now imagine that the classes are not labeled as shown in Figure 1(b). How should we determine the mean for each of the classes then? In order to solve this problem, we could use an iterative approach: first make a guess of the class label for each data point, then compute the means and update the guess of the class labels again. We repeat until the means converge.

The problem of estimating parameters in the absence of labels is known as unsupervised learning. There are many unsupervised learning methods. We will focus on the Expectation Maximization (EM) algorithm.

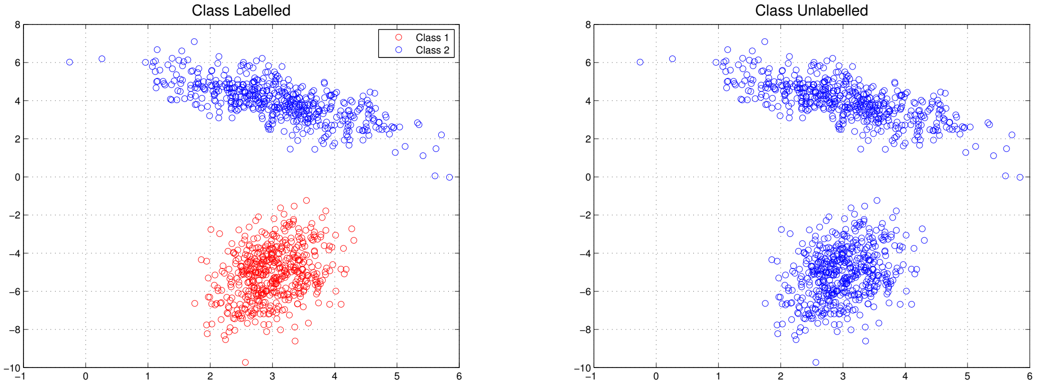

Figure 1: Estimation of parameters becomes trivial given the labelled classes

## 2 The EM-algorithm

### Notations

1. $\boldsymbol{Y}$, $\boldsymbol{y}$ observations. $\boldsymbol{Y}$ = random variable; $\boldsymbol{y}$ = realization of $\boldsymbol{Y}$.

2. $\boldsymbol{X}$, $\boldsymbol{x}$ complete data.

3. $\boldsymbol{Z}$, $\boldsymbol{z}$, missing data. Note that $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$.

4. $\boldsymbol{\theta}$: unknown deterministic parameter. $\boldsymbol{\theta}^{(t)}$: $t^{th}$ estimate of the $\boldsymbol{\theta}$ in the EM iteration.

5. $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the distribution of $\boldsymbol{Y}$ given $\boldsymbol{\theta}$.

6. $f(\boldsymbol{X}|\boldsymbol{\theta})$ is a random variable taking value of $f(\boldsymbol{X}|\boldsymbol{\theta})$ (Remember: $f(\cdot|\boldsymbol{\theta})$ is a function and thus we can put any argument into $f(\cdot|\boldsymbol{\theta})$ and evaluate its output.)

7. $\mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}}[g(\boldsymbol{X})] = \int g(\boldsymbol{x}) f_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) d\boldsymbol{x}$ is the conditional expectation of $g(\boldsymbol{X})$ given $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{\theta}$.

8. $\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{y}|\boldsymbol{\theta})$ is the log-likelihood. Note that $\ell(\boldsymbol{\theta})$ depends on $\boldsymbol{y}$.

### EM Steps

The EM-algorithm consists of two steps:

1. **E-step**: Given $\boldsymbol{y}$ and pretending for the moment that $\boldsymbol{\theta}^{(t)}$ is correct, formulate the distribution for the complete data $\boldsymbol{x}$:
$$f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}).$$
Then, we calculate the Q-function:
$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &\overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{X}|\boldsymbol{\theta})] \\ &= \int \log f(\boldsymbol{x}|\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{x} \end{aligned}$$

2. **M-step**: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with regard to $\boldsymbol{\theta}$:
$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

### Properties of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$

1. Ideally, if we have the distribution of the complete data $\boldsymbol{x}$, then finding the parameter can be done by maximizing $f(\boldsymbol{x}|\boldsymbol{\theta})$. However, the complete data is only a virtual thing we created to solved the problem. In reality we never know $\boldsymbol{x}$. All we know is its distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$, which depends on what we know about $\boldsymbol{x}$. So one way to handle this uncertainty is to compute the average. This average is the Q-function.

2. Another way of looking at $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. We can treat $\log f(\boldsymbol{X}|\boldsymbol{\theta})$ as a function of two variables $h(\boldsymbol{X}, \boldsymbol{\theta})$. Maximizing over $\boldsymbol{\theta}$ is problematic because it depends on $\boldsymbol{X}$. So by taking expectation $\mathbb{E}_{\boldsymbol{X}}[h(\boldsymbol{X}, \boldsymbol{\theta})]$ we can eliminate the dependency on $\boldsymbol{X}$.

3. $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ can be thought of a local approximation of the log-likelihood function $\ell(\boldsymbol{\theta})$: Here, by 'local' we meant that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ stays close to its previous estimate $\boldsymbol{\theta}^{(t)}$. In fact if $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$, then $\ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}^{(t)})$.

## 3 Estimating Mean with Partial Observation

Let us consider the first example of the EM algorithm. Suppose that we generated a sequence of $n$ random variables $Y_i \sim \mathcal{N}(\theta, \sigma^2)$ for $i = 1, \ldots, n$. Imagine that we have only observed $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_m]$ where $m < n$. How should we estimate $\theta$ based on $\boldsymbol{Y}$?

Intuitively, the estimated $\theta$ should be the sample mean of the $m$ observations $\widehat{\theta} = \frac{1}{m}\sum_{i=1}^{m} Y_i$. However, in this example we would like to derive the EM algorithm and see if the EM algorithm would match with our intuition.

**Solution:** To start the EM algorithm, we first need to specify the missing data and the complete data. In this problem, the missing data is $\boldsymbol{Z} = [Y_{m+1}, \ldots, Y_n]$, and the complete data is $\boldsymbol{X} = [\boldsymbol{Y}, \boldsymbol{Z}]$. The distribution of $\boldsymbol{X}$ is:
$$\log f(\boldsymbol{X}|\theta) = \frac{-n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(Y_i - \theta)^2}{2\sigma^2}. \tag{1}$$

Therefore, the Q function is

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &\overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\theta^{(t)}}[\log f(\boldsymbol{X}|\theta)] \\
&= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\theta^{(t)}}\left[\frac{-n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{m}\frac{(Y_i-\theta)^2}{2\sigma^2} - \sum_{i=m+1}^{n}\frac{(Y_i-\theta)^2}{2\sigma^2}\right] \\
&= \frac{-n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{m}\frac{(y_i-\theta)^2}{2\sigma^2} - \sum_{i=m+1}^{n}\frac{\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\theta^{(t)}}[(Y_i-\theta)^2]}{2\sigma^2}.
\end{aligned}
$$

The last expectation can be evaluated as

$$
\begin{aligned}
\mathbb{E}_{Y_i|\boldsymbol{Y},\theta^{(t)}}[(Y_i-\theta)^2] &= \mathbb{E}_{Y_i|\boldsymbol{Y},\theta^{(t)}}[Y_i^2 - 2Y_i\theta + \theta^2] \\
&= [(\theta^{(t)})^2 + \sigma^2 - 2\theta^{(t)}\theta + \theta^2].
\end{aligned}
$$

Therefore, the Q function is

$$
Q(\theta|\theta^{(t)}) = \frac{-n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{m}\frac{(y_i-\theta)^2}{2\sigma^2} - \frac{n-m}{2\sigma^2}[(\theta^{(t)})^2 + \sigma^2 - 2\theta^{(t)}\theta + \theta^2].
$$

In the M-step, we need to maximize the Q-function. To this end, we set

$$
\frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)}) = 0,
$$

which yields that

$$
\theta^{(t+1)} = \frac{\sum_{i=1}^{m} y_i + (n-m)\theta^{(t)}}{n}.
$$

It is not difficult to show that as $t \to \infty$, $\theta^{(t)} \to \theta^{(\infty)}$. Hence,

$$
\theta^{(\infty)} = \frac{\sum_{i=1}^{m} y_i}{n} + \left(1 - \frac{m}{n}\right)\theta^{(\infty)},
$$

which yields

$$
\theta^{(\infty)} = \frac{1}{m}\sum_{i=1}^{m} y_i.
$$

This result says that as the EM algorithm converges, the estimated parameter converges to the sample mean using the available $m$ samples, which is quite intuitive.

# 4 Gaussian Mixture With Known Mean And Variance

Our next example of the EM algorithm to estimate the mixture weights of a Gaussian mixture with known mean and variance. A Gaussian mixture is defined as

$$
f(y\,|\,\boldsymbol{\theta}) = \sum_{i=1}^{k}\theta_i\mathcal{N}(y\,|\,\mu_i,\sigma_i^2), \tag{2}
$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]$ is called the mixture weight. The mixture weight satisfies the condition that

$$
\sum_{i=1}^{k}\theta_i = 1.
$$

Our goal is to derive the EM-algorithm for $\boldsymbol{\theta}$.

**Solution**: We first need to define the missing data. For this problem, we observe that the observed data is $\boldsymbol{Y} = [y_1, y_2, \cdots, y_n]$. The missing data can be defined as the label for each $y_j$, so that $\boldsymbol{Z} = [Z_1, Z_2, \ldots, Z_n]$, with $Z_j \in \{1, \ldots, k\}$. Consequently, the complete data is $\boldsymbol{X} = [X_1, X_2, \cdots, X_n]$, where $X_j = (y_j, Z_j)$.

The distribution of the complete data can be computed as

$$f(x_j|\boldsymbol{\theta}) = f(y_j, z_j|\boldsymbol{\theta}) = \theta_{z_j}\mathcal{N}(y_j \,|\, \mu_{z_j}, \sigma^2_{z_j}),$$

Thus, the Q function is

$$\begin{aligned}
Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\boldsymbol{X}\,|,\ \boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left\{\log f(\boldsymbol{X}\,|,\ \boldsymbol{\theta})\right\} \\
&= \mathbb{E}_{\boldsymbol{Z}\,|,\ \boldsymbol{y},\boldsymbol{\theta}^{(t)}}\left\{\log f(\boldsymbol{Z},\boldsymbol{y}\,|,\ \boldsymbol{\theta})\right\} \\
&= \mathbb{E}_{\boldsymbol{Z}\,|,\ \boldsymbol{y},\boldsymbol{\theta}^{(t)}}\left\{\log \prod_{j=1}^{n} \theta_{z_j}\mathcal{N}(y_j\,|,\ \mu_{z_j}, \sigma^2_{z_j})\right\} \\
&= \sum_{j=1}^{n} \mathbb{E}_{Z_j|y_j,\boldsymbol{\theta}^{(t)}}\left\{\log\theta_{z_j} + \log\mathcal{N}(y_j\,|,\ \mu_{z_j}, \sigma^2_{z_j})\right\}.
\end{aligned}$$

The expectation can be evaluated as

$$\begin{aligned}
\mathbb{E}_{Z_j|y_j,\boldsymbol{\theta}^{(t)}}\{\log\theta_{z_j}\} &= \sum_{z_j} \log\theta_{z_j}\mathbb{P}(Z_j = z_j|y_j, \theta^{(t)}) \\
&= \sum_{i=1}^{k} \log\theta_i \underbrace{\mathbb{P}(Z_j = i|y_j, \theta^{(t)})}_{\overset{\text{def}}{=}\gamma_{ij}^{(t)}}.
\end{aligned}$$

By summing over all $j$'s, we can further define

$$\begin{aligned}
\gamma_i^{(t)} &= \sum_{j=1}^{n} \gamma_{ij}^{(t)} \\
&= \sum_{j=1}^{n} \mathbb{P}(Z_j = i\,|\,y_j, \boldsymbol{\theta}^{(t)}) \\
&= \sum_{j=1}^{n} \frac{\boldsymbol{\theta}_i^{(t)}\mathcal{N}(y_j\,|\,\mu_i, \sigma_i^2)}{\sum_{i=1}^{k}\boldsymbol{\theta}_i^{(t)}\mathcal{N}(y_j\,|\,\mu_i, \sigma_i^2)}
\end{aligned}$$

Therefore, the Q function becomes

$$\begin{aligned}
Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(t)}) &= \sum_{j=1}^{n}\sum_{i=1}^{k} \log\gamma_{ij}^{(t)}\theta_i + C \\
&= \sum_{i=1}^{k} \log\gamma_i^{(t)}\theta_i + C,
\end{aligned}$$

for some constant $C$ independent of $\boldsymbol{\theta}$. Maximizing over $\boldsymbol{\theta}$ yields

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \underset{\boldsymbol{\theta}}{\text{argmax}} \ \sum_{i=1}^{k}\gamma_i^{(t)}\log\theta_i \\
&= \frac{\gamma_i^{(t)}}{\sum_{i=1}^{k}\gamma_i^{(t)}},
\end{aligned}$$

where the last equality is due to Gibbs inequality. To summarize the EM algorithm is given in the algorithm below.

```
Data: Gaussian Mixture with known mean and variance
Result: Estimated $\boldsymbol{\theta}$
for $t = 1, \cdots$ do
    $\gamma_i^{(t)} = \sum_{j=1}^{n} \frac{\boldsymbol{\theta}_i^{(t)} \mathcal{N}(y_j | \mu_i, \sigma_i^2)}{\sum_{i=1}^{k} \boldsymbol{\theta}_i^{(t)} \mathcal{N}(y_j | \mu_i, \sigma_i^2)}$
    $\boldsymbol{\theta}_i^{(t)} = \frac{\gamma_i^{(t)}}{\sum_{i=1}^{k} \gamma_i^{(t)}}$
end
```

**Remark:** To solve $\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \sum_{i=1}^{k} \gamma_i^{(t)} \log \boldsymbol{\theta}_i$, we use the Gibbs inequality. Gibbs inequality states that for all $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\sum_{i=1}^{n} \alpha_i = 1$, $\sum_{i=1}^{n} \beta_i = 1$, $0 \leq \alpha_i \leq 1$ and $0 \leq \beta_i \leq 1$, it holds that

$$\sum_{i=1}^{n} \alpha_i \log \beta_i \leq \sum_{i=1}^{n} \alpha_i \log \alpha_i, \tag{3}$$

with the equality holds when $\alpha_i = \beta_i$ for all $i$. The proof of Gibbs inequality is due to the non-negativity of the KL-divergence which we will skip. What we want to show is that if we let

$$\alpha_i = \frac{\gamma_i^{(t)}}{\sum_{i=1}^{k} \gamma_i^{(t)}}, \qquad \beta_i = \theta_i,$$

then the equality holds when:

$$\theta_i = \frac{\gamma_i^{(t)}}{\sum_{i=1}^{k} \gamma_i^{(t)}},$$

which is the result we want.

# 5 Gaussian Mixture

Previously we have been working on Gaussian Mixtures with known mean and variance. However for most of the time it is likely neither mean nor variance is available for us. Thus, we are interested in deriving an EM-algorithm that would generally apply for any Gaussian mixture model with only observations available. Recall that a Gaussian mixture is defined as

$$f(\boldsymbol{y}_i | \boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i \mathcal{N}(y_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{4}$$

where $\boldsymbol{\theta} \overset{\text{def}}{=} \{(\pi_i \boldsymbol{\mu}_i \boldsymbol{\Sigma}_i)\}_{i=1}^{k}$ is the parameter, with $\sum_{i=1}^{k} \pi_i = 1$. Our goal is to derive the EM algorithm for learning $\boldsymbol{\theta}$.

**Solution**. We first specify the following data:

- **Observed Data:** $\boldsymbol{Y} = [\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n]$ with realizations $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]$;
- **Missing Data:** $\boldsymbol{Z} = [Z_1, \cdots, Z_n]$ with realizations $\boldsymbol{z} = [z_1, \cdots, z_n]$, where $z_j \in \{1, \cdots, k\}$;
- **Complete Data:** $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n]$ with realizations $\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ and $\boldsymbol{x}_j = (\boldsymbol{y}_j, z_j)$.

Accordingly, the distribution of the complete data is

$$f(\boldsymbol{y}_j, z_j | \boldsymbol{\theta}) = \pi_{z_j} \mathcal{N}(\boldsymbol{y}_j | \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})$$

Therefore, we can show that

$$\mathbb{P}(Z_j = i|\boldsymbol{y}_j, \boldsymbol{\theta}^{(t)}) = \frac{\pi_i^{(t)}\mathcal{N}(\boldsymbol{y}_j|\boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)})}{\sum_{i=1}^k \pi_i^{(t)}\mathcal{N}(\boldsymbol{y}_i|\boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)})}.$$

The Q function is

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}}\{\log f(\boldsymbol{X}|\boldsymbol{\theta})\} \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}}\{\log f(\boldsymbol{Z}, \boldsymbol{y}|\boldsymbol{\theta})\} \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}}\{\log(\prod_{j=1}^n \pi_{z_j}\mathcal{N}(y_j|\boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}))\} \\
&= \sum_{j=1}^n \mathbb{E}_{Z_j|y_j,\boldsymbol{\theta}^{(t)}}\{\log \pi_{z_j} - \frac{1}{2}\log|\boldsymbol{\Sigma}_{z_j}| - \frac{1}{2}(y_j - \boldsymbol{\mu}_{z_j})^T\boldsymbol{\Sigma}_{z_j}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_{z_j})\} + C \\
&= \sum_{j=1}^n\sum_{i=1}^k \mathbb{P}(Z_j = i|\boldsymbol{y}_i, \boldsymbol{\theta}^{(t)})\{\log \pi_i - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\} + C \\
&= \sum_{j=1}^n\sum_{i=1}^k \gamma_{ij}^{(t)}\{\log \pi_i - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\} + C,
\end{aligned}$$

where $C$ is a constant independent of $\boldsymbol{\theta}$.

The **Maximization** step is to solve the following optimization problem

$$\begin{array}{ll}
\underset{\boldsymbol{\theta}}{\text{maximize}} & Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\
\text{subject to} & \sum_{i=1}^k \pi_i = 1, \\
 & \pi_i > 0, \\
 & \boldsymbol{\Sigma}_i \succ 0.
\end{array} \tag{5}$$

For $\pi_i$, the maximization is

$$\begin{array}{ll}
\underset{\boldsymbol{\pi}}{\text{maximize}} & \sum_{i=1}^k\sum_{j=1}^n \gamma_{ij}^{(t)}\log \pi_i \\
\text{subject to} & \sum_{i=1}^k \pi_i = 1, \quad \pi_i > 0
\end{array} \tag{6}$$

The solution of this problem is

$$\pi_i^{(t+1)} = \frac{\sum_{j=1}^n \gamma_{ij}^{(t)}}{\sum_{i=1}^k\sum_{j=1}^n \gamma_{ij}^{(t)}} = \frac{\sum_{j=1}^n \gamma_{ij}^{(t)}}{n}. \tag{7}$$

For $\boldsymbol{\mu}_i$, the maximization can be reduced to solving the equation

$$\frac{\partial}{\partial\boldsymbol{\mu}_i}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = 0. \tag{8}$$

The left hand side is

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\mu}_i}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \frac{\partial}{\partial\boldsymbol{\mu}_i}\{\sum_{j=1}^n\sum_{i=1}^k \gamma_{ij}^{(t)}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\} \\
&= \boldsymbol{\Sigma}_i^{-1}(\sum_{j=1}^n \gamma_{ij}^{(t)}\boldsymbol{y}_j - \sum_{j=1}^n \gamma_{ij}^{(t)}\boldsymbol{\mu}_i).
\end{aligned}$$

Therefore,

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{\sum_{j=1}^n \gamma_{ij}^{(t)}\boldsymbol{y}_i}{\sum_{j=1}^n \gamma_{ij}^{(t)}} \tag{9}$$

For $\boldsymbol{\Sigma}_i$, the maximization is equivalent to solving

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = 0. \tag{10}$$

The left hand side is

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\Sigma}_i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= -\frac{1}{2}(\boldsymbol{\Sigma}_{j=1}^{n}\gamma_{ij}^{(t)})\frac{\partial \log|\boldsymbol{\Sigma}_i|}{\partial \boldsymbol{\Sigma}_i} - \frac{1}{2}\sum_{j=1}^{n}\gamma_{ij}^{(t)}\frac{\partial}{\partial \boldsymbol{\Sigma}_i}\{(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\} \\ &= -\frac{1}{2}(\sum_{i=1}^{n}\gamma_{ij}^{t})\boldsymbol{\Sigma}_i^{-1} + \frac{1}{2}\sum_{j=1}^{n}\gamma_{ij}^{(t)}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}.\end{aligned}$$

Therefore,

$$\boldsymbol{\Sigma}_i^{t+1} = \frac{\sum_{j=1}^{n}\gamma_{ij}^{(t)}(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(t+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(t+1)})^T}{\sum_{i=1}^{n}\gamma_{ij}^{t}}. \tag{11}$$

# 6 Bernoulli Mixture

Our next example is to consider a Bernoulli mixture model. To motivate this problem, let us imagine that we have a dataset of various items. Our goal is to see whether there is any relationship between the presence or absence of these items. For example, if the object **'A'** (e.g. a tree) was presented, there is some probability that the object **'B'** (e.g. a flower) is also presented. However if given certain object **'C'** (e.g. a dinosaur) presented it is unlikely to see the object **'D'** (e.g. a car, unless you are in `Jurassic Park`!)

To setup the problem let us first define some notations. We use $\boldsymbol{Y}^1, \cdots, \boldsymbol{Y}^N$ to denote $N$ images we have observed. In each image, there are at most $M$ items, so that $\boldsymbol{Y}^n = [Y_1^n, \cdots, Y_M^n]$ for $n = 1, \ldots, N$. Each entry in this vector is a Bernoulli random variable. Moreover, we define

$$\mathbb{P}\left(Y_i^n = 1 \,|\, Y_k^n = 1\right) \overset{\text{def}}{=} \theta_{ki}. \tag{12}$$

Therefore, the goal is to estimate the matrix

$$\boldsymbol{\Theta} = \begin{bmatrix}\theta_{11} & \ldots & \theta_{1M} \\ \vdots & \ddots & \vdots \\ \theta_{M1} & \ldots & \theta_{MM}\end{bmatrix} \tag{13}$$

from the observations $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^N$.

The general problem of estimating $\boldsymbol{\Theta}$ from $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^N$ is very difficult. Therefore, it is necessary to pose some assumptions on the problem. The assumption we make here is "semi-valid" from our daily experience. It is not completely true, but they are simple enough to provide us some computational solutions.

**Assumption 1.** CONDITIONAL INDEPENDENCE
We assume that the observations follow the conditional independence structure:

$$\mathbb{P}(Y_i^n = 1 \cap Y_j^n = 1 \,|\, Y_k^n = 1) = \mathbb{P}(Y_i^n = 1 \,|\, Y_k^n = 1) \cdot \mathbb{P}(Y_j^n = 1 \,|\, Y_k^n = 1). \tag{14}$$

Remark: Conditional independence is not the same as independence. For example, we let $A$ be the event that a puppy breaks a toy, $B$ be the event that a mother yells, and $C$ be the event that a child cries. Without knowing the relationship, it could be that the child cries because the mother yells. However, if we assume the conditional independence of $B$ and $C$ given $A$, then we know that the crying of the child and the yelling of the mother are both triggered by the dog, but not by each other.

## Individual Model

In order to understand the EM algorithm of Bernoulli Mixture, let us set $n$ fixed. Consequently,

$$\mathbb{P}(\boldsymbol{Y}^n = \boldsymbol{y}^n) = \sum_{m=1}^{M} \mathbb{P}(\boldsymbol{Y}^n = \boldsymbol{y}^n | \text{‘item } m \text{ is active’}) \underbrace{\mathbb{P}(\text{‘item } m \text{ is active’})}_{\overset{\text{def}}{=} \pi_m}.$$

Furthermore,

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}^n = \boldsymbol{y}^n \,|\, \text{‘item } m \text{ is active’}) &= \prod_{i=1}^{M} \theta_{mi}^{y_i^n}(1-\theta_{mi})^{1-y_i^n} \\ &\overset{\text{def}}{=} f_m(\boldsymbol{y}_n \,|\, \boldsymbol{\theta}_m),\end{aligned}$$

where $\boldsymbol{\theta}_m = [\theta_{m1}, \cdots, \theta_{mM}]$ is the $m$th row of $\boldsymbol{\Theta}$. Therefore,

$$P(\boldsymbol{Y}^n = \boldsymbol{y}^n) = \sum_{m=1}^{M} \pi_m \; f_m(\boldsymbol{y}^n \,|\, \boldsymbol{\theta}_m). \tag{15}$$

## EM Algorithm

Now, we will derive EM algorithm to estimate $\{\pi_1, \cdots, \pi_M\}$ and $\boldsymbol{\Theta}$. To start with, let us define the following types of data:

- Observed Data: $\boldsymbol{Y}^1, \cdots, \boldsymbol{Y}^N$;
- Missing Data: $Z_1, \cdots, Z_N$ with realizations $z_1, \cdots, z_N$ and $z_n \in \mathbb{R}^{1\times N}$;
- Complete Data: $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$, accordingly $\boldsymbol{x}_n = (\boldsymbol{y}^n, z_n)$.

The distribution of the complete data is

$$\mathbb{P}(\boldsymbol{Y}^n = \boldsymbol{y}^n, Z_n = z_n \,|\, \boldsymbol{\Theta}) = \pi_m f_m(\boldsymbol{y}^n | \boldsymbol{\theta}_m).$$

The distribution of the missing data conditioned on the observed data is

$$\mathbb{P}(Z_n = m \,|\, \boldsymbol{Y}^n = \boldsymbol{y}^n, \boldsymbol{\Theta}^{(t)}) = \frac{\pi_m^{(t)} f_m(\boldsymbol{y}_n \,|\, \boldsymbol{\theta}_m^{(t)})}{\sum_{m=1}^{M} \pi_m^{(t)} f_m(\boldsymbol{y}_n | \boldsymbol{\theta}_m^{(t)})}.$$

The $n$th Q function is

$$\begin{aligned}Q_n(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &\overset{\text{def}}{=} \mathbb{E}_{Z_n \,|\, \boldsymbol{y}_n, \boldsymbol{\Theta}^{(t)}}[\log f(\boldsymbol{X}_n | \boldsymbol{\Theta})] \\ &= \mathbb{E}_{Z_n \,|\, \boldsymbol{y}_n, \boldsymbol{\Theta}^{(t)}}[\log f(Z_n, \boldsymbol{y}_n | \boldsymbol{\Theta})] \\ &= \sum_{m=1}^{M} \log(\pi_m f_m(\boldsymbol{y}_n | \boldsymbol{\theta}_m^{(t)})) \underbrace{\mathbb{P}(Z_n = m \,|\, \boldsymbol{y}_n, \boldsymbol{\Theta}^{(t)})}_{\overset{\text{def}}{=} \gamma_{ij}^{(t)}} \\ &= \sum_{m=1}^{M} \gamma_{nm}^{(t)} \log(\pi_m f_m(\boldsymbol{y}_n | \boldsymbol{\theta}_m^{(t)})),\end{aligned}$$

where we can show that

$$\begin{aligned}\log(\pi_m f_m(\boldsymbol{y}_n \,|\, \boldsymbol{\theta}_m^{(t)})) &= \log \pi_m + \log \prod_{i=1}^{M} \theta_{mi}^{y_i^n}(1-\theta_{mi})^{1-y_i^n} \\ &= \log \pi_m + \sum_{i=1}^{M} y_i^n \log \theta_{mi} + (1-y_i^n)\log(1-\theta_{mi}).\end{aligned}$$

Therefore, overall Q-function is

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t}) = \sum_{n=1}^{M}\sum_{m=1}^{M}\gamma_{nm}^{(t)}\left[\log\pi_m + \sum_{i=1}^{M} y_i^n \log\theta_{mi} + (1-y_i^n)\log(1-\theta_{mi})\right]. \tag{16}$$

To maximize the Q function, we solve

$$\boldsymbol{\Theta}^{(t=1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}). \tag{17}$$

For a fixed $m$ and $i$, we have

$$\frac{\partial}{\partial\theta_{mi}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = \sum_{n=1}^{N}\gamma_{nm}^{(t)}\left[\frac{y_i^n}{\theta_{mi}} - \frac{1-y_i^n}{1-\theta_{mi}}\right].$$

Setting this to zero yields

$$\frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)} y_i^n}{\theta_{mi}} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)}(1-y_i^n)}{1-\theta_{mi}},$$

which is

$$\theta_{mi}^{(t+1)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)} y_i}{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}. \tag{18}$$

```
Data: EM Algorithm for Bernoulli Mixture Model
Result: Estimated $\Theta$ and $\pi_m$
for $t = 1, \cdots$ do
    $\gamma_{nm}^{(t)} = \frac{\pi_m^{(t)} f_m(\boldsymbol{y}^n|\boldsymbol{\theta}_m^{(t)})}{\sum_{m=1}^{M}\pi_m^{(t)} f_m(\boldsymbol{y}^n|\boldsymbol{\theta}_m^{(t)})}$
    $\theta_{mi}^{(t+1)} = \frac{\sum_{n=1}^{N}\gamma_{nm}^{(t)} y_i^n}{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}$
    $\pi_m^{(t+1)} = \frac{\gamma_{nm}^{(t)}}{\sum_{n=1}^{N}\gamma_{nm}^{(t)}}$
end
```

# 7 Convergence of EM

The convergence of EM algorithm is known to be local. What it means is that as the EM algorithm iterates, $\boldsymbol{\theta}^{(t+1)}$ will never be less likely than $\boldsymbol{\theta}^{(t)}$. This property is called the monotonicity of EM, which is the result of the following theorem.

**Theorem 1.**
Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random variables with parametric distribution controlled by a parameter $\boldsymbol{\theta} \in \Lambda$. Suppose that:

1. $\boldsymbol{X}$ does not depend on $\boldsymbol{\theta}$;

2. There exists a Markov relationship

$$\boldsymbol{\theta} \to \boldsymbol{X} \to \boldsymbol{Y}$$

i.e. $f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{x})$ for all $\boldsymbol{\theta} \in \Lambda$ and $\boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y} \in \mathcal{Y}$.

Then, for $\boldsymbol{\theta} \in \Lambda$ and $\boldsymbol{y} \in \mathcal{Y}$ such that $\mathcal{X}(y) \neq \emptyset$, we have:

$$\ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}^{(t)}) \ if \ Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}). \tag{19}$$

**Proof.**

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log f(\boldsymbol{y} \mid \boldsymbol{\theta}) \quad \text{(by \ definition)} \\
&= \log \int_{\mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{\theta}) d\boldsymbol{x} \quad \text{(marginalization, i.e., total probability)} \\
&= \log \int_{\mathcal{X}(\boldsymbol{y})} \frac{f(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{\theta})}{f(\boldsymbol{x} | \boldsymbol{y}, \boldsymbol{\theta}^{(t)})} f(\boldsymbol{x} | \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{x} \\
&= \log \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \frac{f(\boldsymbol{X}, \boldsymbol{y} | \boldsymbol{\theta})}{f(\boldsymbol{X} | \boldsymbol{y}, \boldsymbol{\theta})} \right] \\
&\geq \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \log \frac{f(\boldsymbol{X}, \boldsymbol{y} | \boldsymbol{\theta})}{f(\boldsymbol{X} | \boldsymbol{y}, \boldsymbol{\theta})} \right] \quad \text{(Jensen's \ Inequality)} \\
&= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \log \frac{f(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}) f(\boldsymbol{X} | \boldsymbol{\theta})}{\frac{f(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}^{(t)}) f(\boldsymbol{X} | \boldsymbol{\theta}^{(t)})}{f(\boldsymbol{y} | \boldsymbol{\theta}^{(t)})}} \right] \quad \text{(Baye's \ Rule)} \\
&= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \log \frac{f(\boldsymbol{y} | \boldsymbol{X}) f(\boldsymbol{X} | \boldsymbol{\theta}) f(\boldsymbol{y} | \boldsymbol{\theta}^{(t)})}{f(\boldsymbol{y} | \boldsymbol{X}) f(\boldsymbol{X} | \boldsymbol{\theta}^{(t)})} \right] \quad \text{(assumption 2)} \\
&= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \log \frac{f(\boldsymbol{X} | \boldsymbol{\theta}) f(\boldsymbol{y} | \boldsymbol{\theta}^{(t)})}{f(\boldsymbol{X} | \boldsymbol{\theta}^{(t)})} \right] \\
&= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} [\log f(\boldsymbol{X} | \boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \log f(\boldsymbol{X} | \boldsymbol{\theta}^{(t)}) \right] + \mathbb{E}_{\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}} \left[ \log f(\boldsymbol{y} | \boldsymbol{\theta}^{(t)}) \right] \\
&= Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t)}) + \underbrace{\log f(\boldsymbol{y} | \boldsymbol{\theta}^{(t)})}_{= \ell(\boldsymbol{\theta}^{(t)})}
\end{aligned}$$

Thus, $\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$. Hence if $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$, then $\ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}^{(t)})$.

□

# 8 Using Prior with EM

The EM algorithm can fail due to singularity of the log-likelihood function. For example, when learning a GMM with 10 components, the algorithm may decide that the most likely solution is for one of the Gaussians to only have one data point assigned to it. This could yield some bad result of having zero covariance.

To alleviate this problem, one can use the prior information about $\boldsymbol{\theta}$. In this case, we can modify the EM setp as

- **E-step**:
$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{X}|y,\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{X}|\boldsymbol{\theta})];$$

- **M-step**:
$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\text{argmax}} \;\; Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \underbrace{\log f(\boldsymbol{\theta})}_{\text{prior}}.$$

## Example

Assume that we have a GMM of $k$-components:

$$f(y_j|\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \mathcal{N}(y_j|\mu_i, \sigma^2). \tag{20}$$

Let us consider a constraint on $\mu_i$:

$$\mu_i = \mu + (i-1)\Delta\mu, \quad for\ i = 1, \cdots, k,$$

i.e. the means are equally spaced. (For details please refer to section 3.3 of `Gupta and Chen`.

**Priors:**

We assume the following priors:

1.
$$\sigma^2 \sim \text{inverse-gamma}\left(\frac{v}{2}, \frac{\epsilon^2}{2}\right).$$

   That is,

$$\begin{aligned} f(\sigma^2) &= \frac{(\frac{\xi^2}{2})^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}(\sigma^2)^{-\frac{v}{2}-1}\exp\left(-\frac{\xi^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-\frac{v+3}{2}}\exp\left(-\frac{\xi^2}{2\sigma^2}\right). \end{aligned}$$

2.
$$\Delta\mu \,|\, \sigma^2 \sim \mathcal{N}\left(\eta, \frac{\sigma}{\rho}\right).$$

   That is,

$$f(\Delta\mu \,|\, \sigma^2) \propto \exp\left(-\frac{(\Delta\mu - \eta)^2}{2(\frac{\sigma^2}{\rho})}\right).$$

Therefore, the joint distribution of the prior is:

$$f(\Delta\mu, \sigma^2) \propto (\sigma^2)^{-\frac{v+3}{2}}\exp\left\{-\frac{\xi^2 + l(\Delta - \eta)^2}{2\sigma^2}\right\}. \tag{21}$$

**Parameters**: $\boldsymbol{\theta} = (w_1, \cdots, w_k, \mu, \Delta\mu, \sigma^2)$. Our goal is to estimate $\boldsymbol{\theta}$.

**EM algorithm**:

First of all, we let

$$\gamma_{ij}^{(t)} = \frac{w_i^{(t)}\mathcal{N}(y_j|\mu_i^{(t)}, \sigma^{2(t)})}{\sum_{i=1}^k w_i^{(t)}\mathcal{N}(y_j|\mu_i^{(t)}, \sigma^{2(t)})}. \tag{22}$$

The EM steps can be derived as follows.

**The Expectation Step**

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{j=1}^n\sum_{i=1}^k \gamma_{ij}^{(t)}\log(w_i\mathcal{N}(y_j|\mu_i, \sigma^2)) \\ &= \sum_{j=1}^n\sum_{i=1}^k \gamma_{ij}^{(t)}\log(w_i\mathcal{N}(y_j|\mu + (i-1)\Delta\mu, \sigma^2)) \\ &= \sum_{j=1}^n\sum_{i=1}^k \gamma_{ij}^{(t)}\log w_i - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^n\sum_{i=1}^k \gamma_{ij}^{(t)}(y_j - \mu - (i-1)\Delta\mu)^2 \end{aligned}$$

**The Maximization Step**

$$
\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \log f(\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad \sum_{j=1}^{n}\sum_{i=1}^{k}\gamma_{ij}^{(t)}\log w_i - \frac{n+v+3}{2}\log\sigma^2 - \frac{\xi + l(\Delta\mu-\eta)^2}{2\sigma^2} \\
&\qquad - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\sum_{i=1}^{k}\gamma_{ij}^{(t)}(y_j - \mu - (i-1)\Delta\mu)^2 + C
\end{aligned}
$$

Thus,

$$
w_i^{(t+1)} = \frac{\sum_{j=1}^{n}\gamma_{ij}^{(t)}}{\sum_{i=1}^{k}\sum_{j=1}^{n}\gamma_{ij}^{(t)}},
$$

and

$$
\begin{aligned}
&\begin{cases} \frac{\partial}{\partial\mu}[Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \log f(\boldsymbol{\theta})] = 0 \\ \frac{\partial}{\partial\Delta\mu}[Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \log f(\boldsymbol{\theta})] = 0 \end{cases} \\
&\Rightarrow \begin{bmatrix} 1 & \sum_{i=1}^{k-1} w_{i+1}^{(t+1)} i \\ \sum_{i=1}^{k-1} w_{i+1}^{(t+1)} i & \sum_{i=1}^{k-1} w_{i+1}^{(t+1)} i^2 + \frac{l}{n} \end{bmatrix} \begin{bmatrix} \mu \\ \Delta\mu \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n} y_j \\ \frac{\rho\eta}{n} + \frac{1}{n}\sum_{j=1}^{n}\sum_{i=2}^{k}\gamma_{ij}^{(t)}(i-1)y_j \end{bmatrix}.
\end{aligned}
$$

The solution of $\mu$ and $\Delta\mu$ can be obtained by solving the linear system. Finally,

$$
\begin{aligned}
&\frac{\partial}{\partial\sigma^2}\left(Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \log f(\boldsymbol{\theta})\right) = 0 \\
&\Rightarrow \sigma^{2^{(t+1)}} = \frac{\xi^2 + l(\Delta\mu^{(t+1)} - \eta)^2 + \sum_{j=1}^{n}\sum_{i=1}^{k}\gamma_{ij}^{(t)}(y_j - \mu_i^{(t+1)})^2}{n+v+3}.
\end{aligned}
$$

# 9 MALAB Demo: EM Algorithm for Bernoulli Mixture

## 9.1 Synthesize The Data

```
function [ data_rand] = MakeData( DS, u_vec, p_mat )

cnt = 0;
for ii = 1:1:length(u_vec)
    N = DS*u_vec(ii);
    p_vec = p_mat(ii,:);
    %%
    for m = 1:1:length(p_vec)
        data_vec = randperm(N);
        th = N*p_mat(ii,m);
        for n = 1:1:N
            if data_vec(n) > th
                data_vec(n)= 0;
            else
                data_vec(n) = 1;
            end
        end
        data(cnt+1:cnt+N, m) = data_vec';
    end
    cnt = cnt + N;
end

%% Now randomly permutate the rows of the matrix
[row, column] = size(data);
row_vec = randperm(row);
```

```
for ii = 1:1:row
    randtemp = row_vec(ii);
    data_rand(ii,:) = data(randtemp,:);
end

end
```

## 9.2 Estimate the probability of a Vector Given Bernoulli Distribution

```
function [ p_b ] = Bernoulli_vec( p_vec, y_vec )
%% Calculate the probability of using the current Bernoulli Mixture
p_b = 1;
for ii = 1:1:length(p_vec)
    p_b = p_b*(p_vec(ii)^(y_vec(ii)))*((1-p_vec(ii))^(1-y_vec(ii)));
end

end
```

## 9.3 The Main Function for EM with Bernoulli Mixture

```
close all
clear all
clc
DS = input('Eneter the synthetized data size:');
u_vec = [1/4, 1/2, 1/4]
p_mat = [1, 0.4, 0.05;
    0.2, 1, 0.8;
    0.3, 0.7, 1]
data_rand = MakeData(DS, u_vec, p_mat);
T = input('Enter the desired number of iterations:');

%% Pick Initialization of parameters
u_initial = [1/4, 1/8, 5/8];
p_initial = [0.3, 0.2, 0.8;
    0.1, 0.8, 0.7;
    0.5, 0.15, 0.6];

M = length(u_initial);
N = size(data_rand, 1);
% Initiliaze the parameters
u = u_initial;
p = p_initial;

u_history = zeros(M,T);
p_history = zeros(M,M,T);

for t = 1:1:T
    for m = 1:1:M
        pi_m = u(m);
        p_vec = p(m,:);
        for n= 1:1:N
            y_vec = data_rand(n,:);
            %% Find the Hidden Variable, lambda
            numerator = pi_m*Bernoulli_vec(p_vec, y_vec); % Modle the Bernoulli Process
            denom = 0;
            for mm = 1:1:M
                p_vec_tmp = p(mm,:);
                denom = denom + u(mm)*Bernoulli_vec(p_vec_tmp, y_vec);
            end
            lambda(m,n) = numerator/denom;
        end
    end

    sum_lambda = sum(sum(lambda));

    %% Update mu
    for m = 1:1:M
        u(m) = sum(lambda(m,:))/sum_lambda;
    end

    %% Update P matrix
    for i = 1:1:M
        for m = 1:1:M
            p(m,i) = (sum(lambda(m,:).*data_rand(:,i)'))/(sum(lambda(m,:)));
        end
    end

    %% Save in history for each iteration to plot
    u_history(:,t) = u;
    p_history(:,:,t) = p;
end
disp('updated p and u:')
p
u

figure
hold on
grid on
for m = 1:1:M
    plot(u_history(m,:));
end
ylabel('Estimated \mu value', 'FontSize', 20)
xlabel('Iterations', 'FontSize', 20)
title('Convergence of \mu estimated for Mixture Number = 3', 'FontSize', 20)
for m = 1:1:M
        stem(T, u_vec(m));
end


figure
hold on
grid on
for ii = 1:1:M
    for jj = 1:1:M
        for t = 1:1:T
            tmp = p_history(ii,jj,t);
            plot_vec(t) = tmp;
        end
        plot(plot_vec)
    end
end
ylabel('Estimated P matrix values', 'FontSize', 20)
xlabel('Iterations', 'FontSize', 20)
title('Convergence of P matrix estimated for Mixture Number = 3', 'FontSize', 20)
for m = 1:1:M
    for n = 1:1:M
        stem(T, p_mat(m,n));
    end
end

for m = 1:1:M
    one_loc = find(abs(p(m,:) - 1) == min(abs(p(m,:) - 1)))
    p_final(one_loc,:) = p(m,:);
    u_final(one_loc) = u(m);
end

disp('After Automatic Sorting Based on Diagnals:')
p_final
u_final
```