Lecture Stat 461-561
Expectation-Maximization Algorithm

AD

January 2008

Maximum Likelihood Estimation

- Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, ..., X_n)$. Then given that $\mathbf{X} = \mathbf{x}$ is observed, the likelihood function is given by $f(\mathbf{x}|\theta) = L(\theta|\mathbf{x})$ and $l(\theta) = \log L(\theta|\mathbf{x})$.

Maximum Likelihood Estimation

- Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, ..., X_n)$. Then given that $\mathbf{X} = \mathbf{x}$ is observed, the likelihood function is given by $f(\mathbf{x}|\theta) = L(\theta|\mathbf{x})$ and $l(\theta) = \log L(\theta|\mathbf{x})$.
- The Maximum Likelihood Estimate (MLE) is defined by

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, l(\theta).$$

Maximum Likelihood Estimation

- Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, ..., X_n)$. Then given that $\mathbf{X} = \mathbf{x}$ is observed, the likelihood function is given by $f(\mathbf{x}|\theta) = L(\theta|\mathbf{x})$ and $l(\theta) = \log L(\theta|\mathbf{x})$.
- The Maximum Likelihood Estimate (MLE) is defined by

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} l(\theta).$$

- Under regularity assumptions, the MLE is

Maximum Likelihood Estimation

- Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, ..., X_n)$. Then given that $\mathbf{X} = \mathbf{x}$ is observed, the likelihood function is given by $f(\mathbf{x}|\theta) = L(\theta|\mathbf{x})$ and $l(\theta) = \log L(\theta|\mathbf{x})$.
- The Maximum Likelihood Estimate (MLE) is defined by

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, l(\theta).$$

- Under regularity assumptions, the MLE is
  - consistent, i.e. $\widehat{\theta}_n \xrightarrow{P} \theta_*$ where θ_* is the true value

Maximum Likelihood Estimation

- Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, ..., X_n)$. Then given that $\mathbf{X} = \mathbf{x}$ is observed, the likelihood function is given by $f(\mathbf{x}|\theta) = L(\theta|\mathbf{x})$ and $l(\theta) = \log L(\theta|\mathbf{x})$.
- The Maximum Likelihood Estimate (MLE) is defined by

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} l(\theta).$$

- Under regularity assumptions, the MLE is
  - consistent, i.e. $\widehat{\theta}_n \xrightarrow{P} \theta_*$ where θ_* is the true value
  - asymptotically efficient, i.e. the MLE has the smallest variance, at least for large samples.

Standard Numerical Methods

- Assume $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$, then under regularity assumptions the MLE are the values that solve

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0 \text{ for } i = 1, ..., p. \tag{1}$$

Standard Numerical Methods

- Assume $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$, then under regularity assumptions the MLE are the values that solve

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0 \text{ for } i = 1, ..., p. \tag{1}$$

- Example: if $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ then

$$L(\theta | \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \theta)^2}{2}\right)$$

and

$$\frac{dl(\theta)}{d\theta} = 0 \Leftrightarrow \sum_{i=1}^{n} (x_i - \theta) = 0 \Rightarrow \widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

One can check $\left.\frac{d^2 l(\theta)}{d\theta^2}\right|_{\widehat{\theta}} < 0$, hence it is a maximum.

Standard Numerical Methods

- Assume $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$, then under regularity assumptions the MLE are the values that solve

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0 \text{ for } i = 1, ..., p. \tag{1}$$

- Example: if $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ then

$$L(\theta | \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \theta)^2}{2}\right)$$

and

$$\frac{dl(\theta)}{d\theta} = 0 \Leftrightarrow \sum_{i=1}^{n} (x_i - \theta) = 0 \Rightarrow \widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

One can check $\left.\frac{d^2 l(\theta)}{d\theta^2}\right|_{\widehat{\theta}} < 0$, hence it is a maximum.

- In more complex examples, we cannot solve (1) easily and we need to rely on numerical methods.

Newton-Raphson Optimization

- Iterative algorithm to find the MLE of $\theta = (\theta_1, ..., \theta_p)^T$.

Newton-Raphson Optimization

- Iterative algorithm to find the MLE of $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$.
- Assume $\theta^{(k)}$ is the estimate at iteration k then

$$\frac{\partial l(\theta)}{\partial \theta} \approx \frac{\partial l\left(\theta^{(k)}\right)}{\partial \theta} + \frac{\partial^2 l\left(\theta^{(k)}\right)}{\partial \theta \partial \theta^{\mathsf{T}}} \left(\theta - \theta^{(k)}\right).$$

Newton-Raphson Optimization

- Iterative algorithm to find the MLE of $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$.
- Assume $\theta^{(k)}$ is the estimate at iteration k then

$$\frac{\partial l(\theta)}{\partial \theta} \approx \frac{\partial l\left(\theta^{(k)}\right)}{\partial \theta} + \frac{\partial^2 l\left(\theta^{(k)}\right)}{\partial \theta \partial \theta^{\mathsf{T}}} \left(\theta - \theta^{(k)}\right).$$

- By writing

$$\begin{aligned} g(\theta) &= \left(\frac{\partial l(\theta)}{\partial \theta_1}, \frac{\partial l(\theta)}{\partial \theta_2}, ..., \frac{\partial l(\theta)}{\partial \theta_p}\right)^{\mathsf{T}}, \\ H(\theta) &= \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^{\mathsf{T}}} = \left(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_i}\right) \end{aligned}$$

this means that

$$0 = g(\theta) \approx g\left(\theta^{(k)}\right) + H\left(\theta^{(k)}\right)\left(\theta - \theta^{(k)}\right)$$

Newton-Raphson Optimization

- Iterative algorithm to find the MLE of $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$.
- Assume $\theta^{(k)}$ is the estimate at iteration k then

$$\frac{\partial l(\theta)}{\partial \theta} \approx \frac{\partial l\left(\theta^{(k)}\right)}{\partial \theta} + \frac{\partial^2 l\left(\theta^{(k)}\right)}{\partial \theta \partial \theta^{\mathsf{T}}} \left(\theta - \theta^{(k)}\right).$$

- By writing

$$\begin{aligned} g(\theta) &= \left(\frac{\partial l(\theta)}{\partial \theta_1}, \frac{\partial l(\theta)}{\partial \theta_2}, ..., \frac{\partial l(\theta)}{\partial \theta_p}\right)^{\mathsf{T}}, \\ H(\theta) &= \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^{\mathsf{T}}} = \left(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_i}\right) \end{aligned}$$

  this means that

$$0 = g(\theta) \approx g\left(\theta^{(k)}\right) + H\left(\theta^{(k)}\right)\left(\theta - \theta^{(k)}\right)$$

- This suggests

$$\theta^{(k+1)} = \theta^{(k)} - H\left(\theta^{(k)}\right)^{-1} g\left(\theta^{(k)}\right)$$

Newton-Raphson Optimization

- Iterative algorithm to find the MLE of $\theta = (\theta_1, ..., \theta_p)^{\mathsf{T}}$.
- Assume $\theta^{(k)}$ is the estimate at iteration k then

$$\frac{\partial l(\theta)}{\partial \theta} \approx \frac{\partial l\left(\theta^{(k)}\right)}{\partial \theta} + \frac{\partial^2 l\left(\theta^{(k)}\right)}{\partial \theta \partial \theta^{\mathsf{T}}} \left(\theta - \theta^{(k)}\right).$$

- By writing

$$\begin{aligned} g(\theta) &= \left(\frac{\partial l(\theta)}{\partial \theta_1}, \frac{\partial l(\theta)}{\partial \theta_2}, ..., \frac{\partial l(\theta)}{\partial \theta_p}\right)^{\mathsf{T}}, \\ H(\theta) &= \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^{\mathsf{T}}} = \left(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_i}\right) \end{aligned}$$

  this means that

$$0 = g(\theta) \approx g\left(\theta^{(k)}\right) + H\left(\theta^{(k)}\right)\left(\theta - \theta^{(k)}\right)$$

- This suggests

$$\theta^{(k+1)} = \theta^{(k)} - H\left(\theta^{(k)}\right)^{-1} g\left(\theta^{(k)}\right)$$

Difficulties with Newton-Raphson

- It may prove difficult to calculate the Hessian matrix.

Difficulties with Newton-Raphson

- It may prove difficult to calculate the Hessian matrix.
- At each iteration, it requires computing a new Hessian matrix.

Difficulties with Newton-Raphson

- It may prove difficult to calculate the Hessian matrix.
- At each iteration, it requires computing a new Hessian matrix.
- Depending on the initial value, the method may converge or diverge.

Difficulties with Newton-Raphson

- It may prove difficult to calculate the Hessian matrix.
- At each iteration, it requires computing a new Hessian matrix.
- Depending on the initial value, the method may converge or diverge.
- Quasi-Newton methods have been proposed to mitigate these problems: it avoids calculating the Hessian $H\left(\theta^{(k)}\right)$ and step widths can be introduced to accelerate convergence/prevent divergence.

Quasi-Newton maximization

- Determine a search direction vector $d_k = H_k^{-1} g\left(\theta^{(k)}\right)$.

Quasi-Newton maximization

- Determine a search direction vector $d_k = H_k^{-1} g\left(\theta^{(k)}\right)$.
- Determine the optimum step width λ_k maximizing $l\left(\theta^{(k)} + \lambda d_k\right)$.

Quasi-Newton maximization

- Determine a search direction vector $d_k = H_k^{-1} g\left(\theta^{(k)}\right)$.
- Determine the optimum step width λ_k maximizing $l\left(\theta^{(k)} + \lambda d_k\right)$.
- Set $\theta^{(k+1)} = \theta^{(k)} + \lambda_k d_k$ and $y_k = g\left(\theta^{(k+1)}\right) - g\left(\theta^{(k)}\right)$.

Quasi-Newton maximization

- Determine a search direction vector $d_k = H_k^{-1} g\left(\theta^{(k)}\right)$.
- Determine the optimum step width λ_k maximizing $l\left(\theta^{(k)} + \lambda d_k\right)$.
- Set $\theta^{(k+1)} = \theta^{(k)} + \lambda_k d_k$ and $y_k = g\left(\theta^{(k+1)}\right) - g\left(\theta^{(k)}\right)$.
- Obtain an estimate of $H\left(\theta^{(k+1)}\right)^{-1}$ by using
  Davidson-Fletcher-Powell (DFP) or
  Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm

$$H_{k+1}^{-1} = H_k^{-1} + \frac{s_k s_k^{\mathsf{T}}}{s_k^{\mathsf{T}} y_k} - \frac{H_k^{-1} y_k y_k^{\mathsf{T}} H_k^{-1}}{y_k^{\mathsf{T}} H_k^{-1} y_k},$$

$$H_{k+1}^{-1} = H_k^{-1} + \frac{s_k y_k^{\mathsf{T}} H_k^{-1}}{s_k^{\mathsf{T}} y_k} - \frac{H_k^{-1} y_k s_k^{\mathsf{T}}}{s_k^{\mathsf{T}} y_k} + \left(1 + \frac{y_k H_k^{-1} y_k^{\mathsf{T}}}{s_k^{\mathsf{T}} y_k}\right) \frac{s_k s_k^{\mathsf{T}}}{s_k^{\mathsf{T}} y_k}$$

where $s_k = \theta^{(k+1)} - \theta^{(k)}$.

- The algorithm is initialized with $\theta^{(0)}$ and H_0^{-1} where H_0^{-1} is picked at the identity matrix, an appropriately scale diagonal matrix or an approximate values of $H\left(\theta^{(0)}\right)^{-1}$.

- The algorithm is initialized with $\theta^{(0)}$ and H_0^{-1} where H_0^{-1} is picked at the identity matrix, an appropriately scale diagonal matrix or an approximate values of $H\left(\theta^{(0)}\right)^{-1}$.
- In situations where $g(\theta)$ is also difficult to compute, it can be computed approximately numerically.

- The algorithm is initialized with $\theta^{(0)}$ and H_0^{-1} where H_0^{-1} is picked at the identity matrix, an appropriately scale diagonal matrix or an approximate values of $H\left(\theta^{(0)}\right)^{-1}$.
- In situations where $g(\theta)$ is also difficult to compute, it can be computed approximately numerically.
- *Example*: Consider $X_i \overset{\text{i.i.d.}}{\sim} f\left(x|\,\mu,\tau^2\right)$ where

$$f\left(x|\,\mu,\tau^2\right) = \frac{1}{\pi}\frac{\tau}{(y-\mu)^2+\tau^2}$$

so for n observations

$$\begin{aligned}\frac{\partial l\left(\mu,\tau^2\right)}{\partial \mu} &= 2\sum_{i=1}^{N}\frac{X_i-\mu}{(X_i-\mu)^2+\tau^2}\\ \frac{\partial l\left(\mu,\tau^2\right)}{\partial \tau^2} &= \frac{n}{2\tau^2}-\sum_{i=1}^{N}\frac{1}{(X_i-\mu)^2+\tau^2}\end{aligned}$$

- Numerical results for $\theta^{(0)} = \left(\mu^{(0)}, \tau^{2(0)}\right)^{\mathsf{T}} = (0,1)^{\mathsf{T}}$.

k	$\mu^{(k)}$	$\tau^{2(k)}$	$-l\left(\theta^{(k)}\right)$	$\frac{\partial l\left(\theta^{(k)}\right)}{\partial \mu}$	$\frac{\partial l\left(\theta^{(k)}\right)}{\partial \tau^2}$
0	0.000	1.000	48.126	-0.839	-1.098
1	0.231	1.302	47.874	0.188	-0.144
2	0.180	1.357	47.865	-0.046	-0.040
3	0.189	1.379	47.865	0.002	-0.001
4	0.189	1.380	47.865	-0.001	0.000
5	0.189	1.380	47.865	0.000	0.000

- Numerical results for $\theta^{(0)} = \left(\mu^{(0)}, \tau^{2(0)}\right)^{\mathsf{T}} = (0,1)^{\mathsf{T}}$.

k	$\mu^{(k)}$	$\tau^{2(k)}$	$-l\left(\theta^{(k)}\right)$	$\frac{\partial l\left(\theta^{(k)}\right)}{\partial \mu}$	$\frac{\partial l\left(\theta^{(k)}\right)}{\partial \tau^2}$
0	0.000	1.000	48.126	-0.839	-1.098
1	0.231	1.302	47.874	0.188	-0.144
2	0.180	1.357	47.865	-0.046	-0.040
3	0.189	1.379	47.865	0.002	-0.001
4	0.189	1.380	47.865	-0.001	0.000
5	0.189	1.380	47.865	0.000	0.000

- The quasi-Newton method converges in 5 iterations here.

- Numerical results for $\theta^{(0)} = \left(\mu^{(0)}, \tau^{2(0)}\right)^{\mathsf{T}} = (0,1)^{\mathsf{T}}$.

k	$\mu^{(k)}$	$\tau^{2(k)}$	$-l\left(\theta^{(k)}\right)$	$\frac{\partial l\left(\theta^{(k)}\right)}{\partial \mu}$	$\frac{\partial l\left(\theta^{(k)}\right)}{\partial \tau^2}$
0	0.000	1.000	48.126	-0.839	-1.098
1	0.231	1.302	47.874	0.188	-0.144
2	0.180	1.357	47.865	-0.046	-0.040
3	0.189	1.379	47.865	0.002	-0.001
4	0.189	1.380	47.865	-0.001	0.000
5	0.189	1.380	47.865	0.000	0.000

- The quasi-Newton method converges in 5 iterations here.
- Many statisticians are not big fans of such methods and prefer using the Expectation-Maximization algorithm. (Note: statisticians should follow optimization courses).

Expectation-Maximization Algorithm

- Although the EM algorithm does not apply to all models, it is powerful and elegant: one of the most popular algorithms in statistics.

Expectation-Maximization Algorithm

- Although the EM algorithm does not apply to all models, it is powerful and elegant: one of the most popular algorithms in statistics.
- It is well-suited to so-called **missing data problems** where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$$

where

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{x}|\theta)\, d\mathbf{x}.$$

Expectation-Maximization Algorithm

- Although the EM algorithm does not apply to all models, it is powerful and elegant: one of the most popular algorithms in statistics.
- It is well-suited to so-called **missing data problems** where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$$

where

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{x}|\theta)\, d\mathbf{x}.$$

- We call

Expectation-Maximization Algorithm

- Although the EM algorithm does not apply to all models, it is powerful and elegant: one of the most popular algorithms in statistics.
- It is well-suited to so-called **missing data problems** where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$$

where

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{x}|\theta)\, d\mathbf{x}.$$

- We call
  - **Y** the incomplete data (i.e. the observed data) and **X** the missing data or the augmented data

Expectation-Maximization Algorithm

- Although the EM algorithm does not apply to all models, it is powerful and elegant: one of the most popular algorithms in statistics.
- It is well-suited to so-called **missing data problems** where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$$

where

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{x}|\theta)\, d\mathbf{x}.$$

- We call
  - $\mathbf{Y}$ the incomplete data (i.e. the observed data) and $\mathbf{X}$ the missing data or the augmented data
  - $(\mathbf{X}, \mathbf{Y})$ the complete data (although remember that $\mathbf{X}$ is not observed!).

Expectation-Maximization Algorithm

- Although the EM algorithm does not apply to all models, it is powerful and elegant: one of the most popular algorithms in statistics.
- It is well-suited to so-called **missing data problems** where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$$

where

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{x}|\theta)\, d\mathbf{x}.$$

- We call
  - $\mathbf{Y}$ the incomplete data (i.e. the observed data) and $\mathbf{X}$ the missing data or the augmented data
  - $(\mathbf{X}, \mathbf{Y})$ the complete data (although remember that $\mathbf{X}$ is not observed!).
- More generally, we can have $\mathbf{Y} = h(\mathbf{X})$.

- We introduce the function $Q : \Theta \times \Theta \longrightarrow \mathbb{R}$

$$Q\left(\theta, \theta'\right) = \mathbb{E}\left[\log f\left(\mathbf{y}, \mathbf{x} \middle| \theta\right) \middle| \mathbf{y}, \theta'\right] = \int \log f\left(\mathbf{y}, \mathbf{x} \middle| \theta\right) \ . \ f\left(\mathbf{x} \middle| \mathbf{y}, \theta'\right) d\mathbf{x}.$$

- We introduce the function $Q : \Theta \times \Theta \longrightarrow \mathbb{R}$

$$Q\left(\theta, \theta'\right) = \mathbb{E}\left[\log f\left(\mathbf{y}, \mathbf{x} \middle| \theta\right) \middle| \mathbf{y}, \theta'\right] = \int \log f\left(\mathbf{y}, \mathbf{x} \middle| \theta\right) \ . \ f\left(\mathbf{x} \middle| \mathbf{y}, \theta'\right) d\mathbf{x}.$$

- The EM algorithm is an iterative algorithm defined at iteration $k+1$ by

$$\theta^{(k+1)} = \underset{\theta \in \Theta}{\arg\max}\ Q\left(\theta, \theta^{(k)}\right). \tag{2}$$

- We introduce the function $Q : \Theta \times \Theta \longrightarrow \mathbb{R}$

$$Q\left(\theta, \theta'\right) = \mathbb{E}\left[\log f\left(\mathbf{y}, \mathbf{x} \middle| \theta\right) \middle| \mathbf{y}, \theta'\right] = \int \log f\left(\mathbf{y}, \mathbf{x} \middle| \theta\right) \ . \ f\left(\mathbf{x} \middle| \mathbf{y}, \theta'\right) d\mathbf{x}.$$

- The EM algorithm is an iterative algorithm defined at iteration $k+1$ by

$$\theta^{(k+1)} = \underset{\theta \in \Theta}{\arg\max}\ Q\left(\theta, \theta^{(k)}\right). \tag{2}$$

- **Theorem**: The sequence $\left\{\theta^{(k)}\right\}$ defined by (2) satisfies

$$L\left(\theta^{(k+1)} \middle| \mathbf{y}\right) \geq L\left(\theta^{(k)} \middle| \mathbf{y}\right).$$

- Example: Let

$$Y_i \overset{\text{i.i.d.}}{\sim} \theta m(y) + (1-\theta) h(y)$$

where $\theta \in [0,1]$ is unknown whereas $m(y)$ and $h(y)$ are known pdfs.

- Example: Let

$$Y_i \overset{\text{i.i.d.}}{\sim} \theta m(y) + (1-\theta) h(y)$$

where $\theta \in [0,1]$ is unknown whereas $m(y)$ and $h(y)$ are known pdfs.

- We have

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta) = \prod_{i=1}^{n} (\theta m(y_i) + (1-\theta) h(y_i)).$$

- Example: Let

$$Y_i \overset{\text{i.i.d.}}{\sim} \theta m(y) + (1-\theta) h(y)$$

where $\theta \in [0,1]$ is unknown whereas $m(y)$ and $h(y)$ are known pdfs.

- We have

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta) = \prod_{i=1}^{n} (\theta m(y_i) + (1-\theta) h(y_i)).$$

- Associate to each data Y_i the latent variable $X_i \in \{1,2\}$ which indicates from which distribution Y_i has been drawn, i.e.

$$Y_i|X_i = 1 \sim m(y) \text{ and } Y_i|X_i = 2 \sim h(y)$$

and $\Pr(X_i = 1) = 1 - \Pr(X_i = 2) = \theta$. Thus we have

$$\begin{aligned} f(\mathbf{y}, \mathbf{x}|\theta) &= \prod_{\{i:x_i=1\}} \theta m(y_i) \prod_{\{i:x_i=2\}} (1-\theta) h(y_i) \\ &= \theta^{n_1} (1-\theta)^{n_2} \prod_{\{i:x_i=1\}} m(y_i) \prod_{\{i:x_i=2\}} h(y_i) \end{aligned}$$

where $n_j = \sum_{i=1}^{n} \delta_j(x_i)$ with $\delta_b(a) = 1$ if $a = b$ and zero otherwise.

- From the expression of $f\left(\mathbf{y},\mathbf{x}\middle|\,\theta\right)$, it follows that

$$f\left(\mathbf{x}\middle|\,\mathbf{y},\theta'\right)=\prod_{i=1}^{n}f\left(x_i\middle|\,y_i,\theta'\right)$$

where

$$\begin{aligned}
f\left(x_i=1\middle|\,y_i,\theta'\right) &= 1-f\left(x_i=2\middle|\,y_i,\theta'\right)\\
&= \frac{f\left(x_i=1,y_i,\theta'\right)}{f\left(x_i=1,y_i,\theta'\right)+f\left(x_i=2,y_i,\theta'\right)}\\
&= \frac{\theta' m\left(y_i\right)}{\theta' m\left(y_i\right)+\left(1-\theta'\right)h\left(y_i\right)}.
\end{aligned}$$

$$
\begin{aligned}
Q\left(\theta,\theta'\right) &= \sum_{\mathbf{x}\in\{1,2\}^n} \log f\left(\mathbf{x},\mathbf{y}\middle|\,\theta\right) \;.\; f\left(\mathbf{x}\middle|\,\mathbf{y},\theta'\right) \\
&= \sum_{\mathbf{x}\in\{1,2\}^n} \left[\left(\sum_{i=1}^{n}\delta_1\left(x_i\right)\left(\log\theta+\log m\left(y_i\right)\right)\right)\right. \\
&\quad \left.+\left(\sum_{i=1}^{n}\delta_2\left(x_i\right)\right)\left(\log\left(1-\theta\right)+\log h\left(y_i\right)\right)\right]\prod_{i=1}^{n} f\left(x_i\middle|\,y_i,\theta'\right) \\
&= \sum_{i=1}^{n} f\left(x_i=1\middle|\,y_i,\theta'\right)\left(\log\theta+\log m\left(y_i\right)\right) \\
&\quad +\sum_{i=1}^{n} f\left(x_i=2\middle|\,y_i,\theta'\right)\left(\log\left(1-\theta\right)+\log h\left(y_i\right)\right).
\end{aligned}
$$

- We have

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q\left(\theta, \theta^{(k)}\right).$$

- We have

$$\theta^{(k+1)} = \underset{\theta \in \Theta}{\arg\max}\ Q\left(\theta, \theta^{(k)}\right).$$

- We have

$$\frac{dQ\left(\theta, \theta^{(k)}\right)}{d\theta} = \frac{\sum_{i=1}^{n} f\left(x_i = 1 | y_i, \theta^{(k)}\right)}{\theta} - \frac{\sum_{i=1}^{n} f\left(x_i = 2 | y_i, \theta^{(k)}\right)}{1-\theta}.$$

By solving $\frac{dQ\left(\theta,\theta^{(k)}\right)}{d\theta} = 0$ we obtain

$$\begin{aligned}
\theta^{(k+1)} &= \frac{\sum_{i=1}^{n} f\left(x_i = 1 | y_i, \theta^{(k)}\right)}{\sum_{i=1}^{n} f\left(x_i = 1 | y_i, \theta^{(k)}\right) + \sum_{i=1}^{n} f\left(x_i = 2 | y_i, \theta^{(k)}\right)} \\
&= \frac{\sum_{i=1}^{n} \theta^{(k)} m\left(y_i\right)}{\sum_{i=1}^{n} \theta^{(k)} m\left(y_i\right) + \sum_{i=1}^{n} \left(1-\theta^{(k)}\right) h\left(y_i\right)}.
\end{aligned}$$

- We have
$$\theta^{(k+1)} = \underset{\theta \in \Theta}{\arg\max}\ Q\left(\theta, \theta^{(k)}\right).$$
- We have
$$\frac{dQ\left(\theta, \theta^{(k)}\right)}{d\theta} = \frac{\sum_{i=1}^{n} f\left(x_i = 1 | y_i, \theta^{(k)}\right)}{\theta} - \frac{\sum_{i=1}^{n} f\left(x_i = 2 | y_i, \theta^{(k)}\right)}{1-\theta}.$$

  By solving $\frac{dQ\left(\theta,\theta^{(k)}\right)}{d\theta} = 0$ we obtain
$$\begin{aligned}\theta^{(k+1)} &= \frac{\sum_{i=1}^{n} f\left(x_i = 1 | y_i, \theta^{(k)}\right)}{\sum_{i=1}^{n} f\left(x_i = 1 | y_i, \theta^{(k)}\right) + \sum_{i=1}^{n} f\left(x_i = 2 | y_i, \theta^{(k)}\right)} \\ &= \frac{\sum_{i=1}^{n} \theta^{(k)} m\left(y_i\right)}{\sum_{i=1}^{n} \theta^{(k)} m\left(y_i\right) + \sum_{i=1}^{n} \left(1-\theta^{(k)}\right) h\left(y_i\right)}.\end{aligned}$$
- This updating equation is simple to implement and it is guaranteed that $L\left(\theta^{(k+1)} \middle| \mathbf{y}\right) \geq L\left(\theta^{(k)} \middle| \mathbf{y}\right)$.

- Finite mixture of Gaussians

$$Y_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{p} \pi_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

where $\pi_k \geq 0$ and $\sum_{k=1}^{p} \pi_k = 1$.

- Finite mixture of Gaussians

$$Y_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{p} \pi_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

  where $\pi_k \geq 0$ and $\sum_{k=1}^{p} \pi_k = 1$.
- This simple mixture model is widely used in practice for density estimation and clustering.

- Finite mixture of Gaussians

$$Y_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{p} \pi_k \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

  where $\pi_k \geq 0$ and $\sum_{k=1}^{p} \pi_k = 1$.
- This simple mixture model is widely used in practice for density estimation and clustering.
- Given n observations $\mathbf{Y} = \mathbf{y} = (y_1, ..., y_n)$, we are interested in the MLE estimate of $\theta = \left(\pi_1, ..., \pi_p, \mu_1, ..., \mu_p, \sigma_1^2, ..., \sigma_p^2\right)$ where

$$g\left(\mathbf{y}|\,\theta\right) = \prod_{i=1}^{n} \left(\sum_{k=1}^{p} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{\left(y_i - \mu_k\right)^2}{2\sigma_k^2} \right) \right).$$

- Similarly to the previous lecture, we associate to each Y_i a latent variable $X_i \in \{1, ..., p\}$ such that

$$\Pr(X_i = j) = \pi_j \text{ and } Y_i | (X_i = j) \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right).$$

It follows that

$$f(\mathbf{x}, \mathbf{y} | \theta) = \prod_{i=1}^{n} \frac{\pi_{x_i}}{\sqrt{2\pi\sigma_{x_i}^2}} \exp\left(-\frac{\left(y_i - \mu_{x_i}\right)^2}{2\sigma_{x_i}^2}\right)$$

and

$$f(\mathbf{x} | \theta, \mathbf{y}) = \prod_{i=1}^{n} f(x_i | \theta, y_i)$$

where

$$f(x_i = j | \theta, y_i) = \frac{\pi_j / \sigma_j \exp\left(-\left(y_i - \mu_j\right)^2 / \left(2\sigma_j^2\right)\right)}{\sum_{m=1}^{p} \pi_m / \sigma_m \exp\left(-\left(y_i - \mu_m\right)^2 / \left(2\sigma_m^2\right)\right)}.$$

- We have

$$
\begin{aligned}
Q\left(\theta,\theta^{(k)}\right) &= \sum_{\mathbf{x}} \log f\left(\mathbf{x},\mathbf{y}|\,\theta\right)\ .\ f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right) \\
&= cst + \sum_{\mathbf{x}} \sum_{i=1}^{n} \left(\log \pi_{x_i} - \frac{1}{2}\log \sigma^2_{x_i} - \frac{\left(y_i-\mu_{x_i}\right)^2}{2\sigma^2_{x_i}} \right) \prod_{i=1}^{n} f\left(x_i|\,\theta^{(k)},y_i\right) \\
&= cst + \sum_{m=1}^{p} \left[\left(\log \pi_m - \frac{1}{2}\log \sigma^2_m\right) \left(\sum_{i=1}^{n} f\left(x_i = m|\,\theta^{(k)},y_i\right)\right) \right. \\
&\left. - \sum_{i=1}^{n} \frac{(y_i-\mu_m)^2}{2\sigma^2_m} f\left(x_i = m|\,\theta^{(k)},y_i\right) \right].
\end{aligned}
$$

- We have to maximize $Q\left(\theta, \theta^{(k)}\right)$ over a constrained set, i.e. $\sum_{m=1}^{p} \pi_m = 1$. We introduce a Lagrange multiplier λ and maximize instead

$$Q\left(\theta, \theta^{(k)}\right) + \lambda\left(\sum_{m=1}^{p} \pi_m - 1\right).$$

- We have to maximize $Q\left(\theta, \theta^{(k)}\right)$ over a constrained set, i.e. $\sum_{m=1}^{p} \pi_m = 1$. We introduce a Lagrange multiplier λ and maximize instead

$$Q\left(\theta, \theta^{(k)}\right) + \lambda\left(\sum_{m=1}^{p} \pi_m - 1\right).$$

- We have

$$\frac{\partial Q\left(\theta,\theta^{(k)}\right)}{\partial \sigma_m^2} = -\frac{\left(\sum_{i=1}^{n} f\left(x_i=m|\theta^{(k)},y_i\right)\right)}{2\sigma_m^2} + \sum_{i=1}^{n} \frac{\left(y_i-\mu_m\right)^2}{2\sigma_m^4} f\left(x_i = m|\, \theta^{(k)}, y_i\right)$$

$$2\sigma_m^2 \frac{\partial Q\left(\theta,\theta^{(k)}\right)}{\partial \mu_m} = \mu_m \sum_{i=1}^{n} f\left(x_i = m|\, \theta^{(k)}, y_i\right) - \sum_{i=1}^{n} y_i f\left(x_i = m|\, \theta^{(k)}, y_i\right).$$

- It follows that

$$\mu_m^{(k+1)} = \frac{\sum_{i=1}^n y_i f\left(x_i = m \middle| \theta^{(k)}, y_i\right)}{\sum_{i=1}^n f\left(x_i = m \middle| \theta^{(k)}, y_i\right)},$$

$$\sigma_m^{2(k+1)} = \frac{\sum_{i=1}^n \left(y_i - \mu_m^{(k+1)}\right)^2 f\left(x_i = m \middle| \theta^{(k)}, y_i\right)}{\sum_{i=1}^n f\left(x_i = m \middle| \theta^{(k)}, y_i\right)}.$$

- It follows that

$$\mu_m^{(k+1)} = \frac{\sum_{i=1}^n y_i f\left(x_i = m|\, \theta^{(k)}, y_i\right)}{\sum_{i=1}^n f\left(x_i = m|\, \theta^{(k)}, y_i\right)},$$

$$\sigma_m^{2(k+1)} = \frac{\sum_{i=1}^n \left(y_i - \mu_m^{(k+1)}\right)^2 f\left(x_i = m|\, \theta^{(k)}, y_i\right)}{\sum_{i=1}^n f\left(x_i = m|\, \theta^{(k)}, y_i\right)}.$$

- We have

$$\frac{\partial Q\left(\theta, \theta^{(k)}\right)}{\partial \pi_m} = \frac{\left(\sum_{i=1}^n f\left(x_i = m|\, \theta^{(k)}, y_i\right)\right)}{\pi_m} + \lambda,$$

$$\frac{\partial Q\left(\theta, \theta^{(k)}\right)}{\partial \lambda} = \sum_{m=1}^p \pi_m - 1 \Rightarrow \pi_m^{(k+1)} = \frac{1}{N}\sum_{i=1}^n f\left(x_i = m|\, \theta^{(k)}, y_i\right)$$

- Although the EM is widely used to compute the MLE for finite mixture of Gaussians, we have

$$g\left(\mathbf{y}\middle|\,\theta\right) \longrightarrow \infty \text{ when } \exists \sigma_i \longrightarrow 0.$$

- Although the EM is widely used to compute the MLE for finite mixture of Gaussians, we have

$$g\left(\mathbf{y}\middle|\,\theta\right) \longrightarrow \infty \text{ when } \exists\sigma_i \longrightarrow 0.$$

- Hence, the true MLE estimate does not give sensible results. The EM is used to find a sensible local maximum of $g\left(\mathbf{y}\middle|\,\theta\right)$.

- Although the EM is widely used to compute the MLE for finite mixture of Gaussians, we have

$$g\left(\mathbf{y}\middle|\,\theta\right) \longrightarrow \infty \text{ when } \exists \sigma_i \longrightarrow 0.$$

- Hence, the true MLE estimate does not give sensible results. The EM is used to find a sensible local maximum of $g\left(\mathbf{y}\middle|\,\theta\right)$.
- A way to circumvent this problem consists of introducing a prior distribution $p\left(\theta\right)$ on θ and to use the EM to maximize $g\left(\mathbf{y}\middle|\,\theta\right) p\left(\theta\right)$. We will discuss this Bayesian approach later.

- A (somewhat) important generalization of the EM consists of considering cases where you are interested in maximizing with respect to θ the likelihood function

$$L\left(\theta|\,\mathbf{y}\right) = g\left(\mathbf{y}|\,\theta\right) \text{ where } g\left(\mathbf{y}|\,\theta\right) \text{ follows from } f\left(\mathbf{z}|\,\theta\right).$$

- A (somewhat) important generalization of the EM consists of considering cases where you are interested in maximizing with respect to θ the likelihood function

$$L\left(\theta|\,\mathbf{y}\right) = g\left(\mathbf{y}|\,\theta\right) \text{ where } g\left(\mathbf{y}|\,\theta\right) \text{ follows from } f\left(\mathbf{z}|\,\theta\right).$$

- We call

- A (somewhat) important generalization of the EM consists of considering cases where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta) \text{ where } g(\mathbf{y}|\theta) \text{ follows from } f(\mathbf{z}|\theta).$$

- We call
  - $\mathbf{Y}$ the incomplete data (i.e. the observed data) and $\mathbf{Z}$ are the complete data.

- A (somewhat) important generalization of the EM consists of considering cases where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta) \text{ where } g(\mathbf{y}|\theta) \text{ follows from } f(\mathbf{z}|\theta).$$

- We call
  - $\mathbf{Y}$ the incomplete data (i.e. the observed data) and $\mathbf{Z}$ are the complete data.
  - We can have $\mathbf{Y}$ as a subset of $\mathbf{Z}$, i.e. $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, or $\mathbf{Y} = h(\mathbf{Z})$ where h is a many-to-one mapping.

- A (somewhat) important generalization of the EM consists of considering cases where you are interested in maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta) \text{ where } g(\mathbf{y}|\theta) \text{ follows from } f(\mathbf{z}|\theta).$$

- We call
  - $\mathbf{Y}$ the incomplete data (i.e. the observed data) and $\mathbf{Z}$ are the complete data.
  - We can have $\mathbf{Y}$ as a subset of $\mathbf{Z}$, i.e. $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, or $\mathbf{Y} = h(\mathbf{Z})$ where h is a many-to-one mapping.
- The EM proceeds as follows

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} \int \log f(\mathbf{z}|\theta) \cdot f\left(\mathbf{z}|\mathbf{y}, \theta^{(k)}\right) d\mathbf{z}$$

- Example: Consider the following genetic linkage model where observations

$$(Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

- Example: Consider the following genetic linkage model where observations

$$(Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

- The observed likelihood function is given by

$$g(\mathbf{y}|\,\theta) \propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}.$$

- Example: Consider the following genetic linkage model where observations

$$(Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

- The observed likelihood function is given by

$$g(\mathbf{y}|\theta) \propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}.$$

- Introduce the artificial missing data (X_1, X_2) such that $Y_1 = X_1 + X_2$ and define

$$\begin{aligned} \mathbf{Z} &= (Z_1, ..., Z_5) = (X_1, X_2, Y_2, Y_3, Y_4) \\ &\sim \mathcal{M}\left(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right). \end{aligned}$$

In this case we have $\mathbf{Y} = h(\mathbf{Z})$; i.e. $(Y_1, Y_2, Y_3, Y_4) = (Z_1 + Z_2, Z_3, Z_4, Z_5)$. This equation defines

$$f(\mathbf{z}|\theta) \propto (1-\theta)^{y_2+y_3} \theta^{x_2+y_4}$$

and clearly we have $\sum_{z_1,z_2} f(\mathbf{z}|\theta)\, \delta_{z_1+z_2}(y_1)\, \delta_{z_3}(y_2)\, \delta_{z_4}(y_3)\, \delta_{z_5}(y_4) = g(\mathbf{y}|\theta)$.

- We have $f\left(\mathbf{z}|\,\mathbf{y},\theta'\right)=f\left(x_1,x_2|\,\mathbf{y},\theta'\right)\delta_{y_2}\left(z_3\right)\delta_{y_3}\left(z_4\right)\delta_{y_4}\left(z_5\right)$ where

$$\begin{aligned} f\left(x_1,x_2|\,\mathbf{y},\theta'\right) &= f\left(x_1,x_2|\,y_1,\theta'\right) \\ &= \mathcal{M}\left(y_1;\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta'}{4}},\frac{\frac{\theta'}{4}}{\frac{1}{2}+\frac{\theta'}{4}}\right). \end{aligned}$$

Now we have

$$\begin{aligned} &Q\left(\theta,\theta^{(k)}\right)=\sum_{x_1,x_2}\left[cst+\left(y_2+y_3\right)\log\left(1-\theta\right)+\left(x_2+y_4\right)\log\theta\right] \\ &\times f\left(x_1,x_2|\,y_1,\theta^{(k)}\right) \\ &=cst+\left(y_2+y_3\right)\log\left(1-\theta\right)+\left(\mathbb{E}\left(X_2|\,y_1,\theta^{(k)}\right)+y_4\right)\log\theta \end{aligned}$$

where $\mathbb{E}\left(X_2|\,y_1,\theta^{(k)}\right)=y_1\frac{\theta^{(k)}}{2+\theta^{(k)}}$, thus

$$\begin{aligned} &\frac{dQ\left(\theta,\theta^{(k)}\right)}{d\theta}=0\Leftrightarrow-\frac{\left(y_2+y_3\right)}{\left(1-\theta\right)}+\frac{y_1\frac{\theta^{(k)}}{2+\theta^{(k)}}+y_4}{\theta}=0 \\ &\Rightarrow\theta^{(k+1)}=\left(y_1\frac{\theta^{(k)}}{2+\theta^{(k)}}+y_2+y_3+y_4\right)^{-1}\left(y_1\frac{\theta^{(k)}}{2+\theta^{(k)}}+y_4\right). \end{aligned}$$

Proof of Theorem for Expectation-Maximization Algorithm

- We want to show that $L\left(\theta^{(k+1)}\middle|\mathbf{y}\right) \geq L\left(\theta^{(k)}\middle|\mathbf{y}\right)$ for $\theta^{(k+1)} = \arg\max_{\theta\in\Theta} Q\left(\theta, \theta^{(k)}\right)$.

Proof of Theorem for Expectation-Maximization Algorithm

- We want to show that $L\left(\theta^{(k+1)}\middle|\,\mathbf{y}\right) \geq L\left(\theta^{(k)}\middle|\,\mathbf{y}\right)$ for $\theta^{(k+1)} = \arg\max_{\theta\in\Theta} Q\left(\theta,\theta^{(k)}\right)$.
- *Proof*: We have

$$f\left(\mathbf{x}|\,\theta,\mathbf{y}\right) = \frac{f\left(\mathbf{x},\mathbf{y}|\,\theta\right)}{g\left(\mathbf{y}|\,\theta\right)} \Leftrightarrow g\left(\mathbf{y}|\,\theta\right) = L\left(\theta|\,\mathbf{y}\right) = \frac{f\left(\mathbf{x},\mathbf{y}|\,\theta\right)}{f\left(\mathbf{x}|\,\theta,\mathbf{y}\right)}$$

thus

$$\log L\left(\theta|\,\mathbf{y}\right) = \log f\left(\mathbf{x},\mathbf{y}|\,\theta\right) - \log f\left(\mathbf{x}|\,\theta,\mathbf{y}\right)$$

and for any value $\theta^{(k)}$

$$\log L\left(\theta|\,\mathbf{y}\right) = \underbrace{\int \log f\left(\mathbf{x},\mathbf{y}|\,\theta\right).f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)d\mathbf{x}}_{=Q\left(\theta,\theta^{(k)}\right)} - \int \log f\left(\mathbf{x}|\,\theta,\mathbf{y}\right).f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)d\mathbf{x}.$$

- Now the EM ensures by construction that $Q\left(\theta^{(k+1)}, \theta^{(k)}\right) \geq Q\left(\theta^{(k)}, \theta^{(k)}\right)$. So if we can prove that

$$\begin{aligned} & \int \log f\left(\mathbf{x}|\, \theta^{(k+1)}, \mathbf{y}\right) . f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) d\mathbf{x} \\ \leq & \int \log f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) . f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) d\mathbf{x} \end{aligned}$$

  then the Theorem is proved.

- Now the EM ensures by construction that $Q\left(\theta^{(k+1)},\theta^{(k)}\right) \geq Q\left(\theta^{(k)},\theta^{(k)}\right)$. So if we can prove that

$$\begin{aligned} & \int \log f\left(\mathbf{x}|\,\theta^{(k+1)},\mathbf{y}\right).f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)d\mathbf{x} \\ \leq & \int \log f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right).f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)d\mathbf{x} \end{aligned}$$

  then the Theorem is proved.
- We have thanks to Jensen's inequality

$$\begin{aligned} & \int \log \frac{f\left(\mathbf{x}|\,\theta^{(k+1)},\mathbf{y}\right)}{f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)}.f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)d\mathbf{x} \\ \leq & \log \int \frac{f\left(\mathbf{x}|\,\theta^{(k+1)},\mathbf{y}\right)}{f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)}.f\left(\mathbf{x}|\,\theta^{(k)},\mathbf{y}\right)d\mathbf{x} = 0 \end{aligned}$$

- Now the EM ensures by construction that $Q\left(\theta^{(k+1)}, \theta^{(k)}\right) \geq Q\left(\theta^{(k)}, \theta^{(k)}\right)$. So if we can prove that

$$\begin{aligned} & \int \log f\left(\mathbf{x}|\, \theta^{(k+1)}, \mathbf{y}\right) . f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) d\mathbf{x} \\ \leq & \int \log f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) . f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) d\mathbf{x} \end{aligned}$$

  then the Theorem is proved.
- We have thanks to Jensen's inequality

$$\begin{aligned} & \int \log \frac{f\left(\mathbf{x}|\, \theta^{(k+1)}, \mathbf{y}\right)}{f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right)} . f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) d\mathbf{x} \\ \leq & \log \int \frac{f\left(\mathbf{x}|\, \theta^{(k+1)}, \mathbf{y}\right)}{f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right)} . f\left(\mathbf{x}|\, \theta^{(k)}, \mathbf{y}\right) d\mathbf{x} = 0 \end{aligned}$$

- There are numerous variations of the EM in the literature. We will discuss some later.