Computational Genomics **Fall Semester, 2014**

Lecture 7: December 9, 2014

Lecturer: Roded Sharan *Scribe: Livnat Jerby-Arnon, Yotam Frank*

7.1 Introduction to Expectation Maximization (EM)

The Expectation Maximization (EM) algorithm is a popular tool for simplifying difficult maximum likelihood problems [1, 2]. Typically these problems are ones for which maximization of the likelihood is difficult, but made easier by data augmentation, meaning, by enlarging the sample with latent (unobserved) data. Throughout, we will denoted the model parameters as θ, and the latent data as y [3].

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, that is, the model parameters θ and the latent variables y, and simultaneously solving the resulting equations. In statistical models with latent variables, this is usually impossible, as the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that the following is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set. The algorithm keeps alternating between the two until the resulting values both converge to fixed points. As we will show, although this process will not guarantee to find a solution with maximum likelihood, it will obtain a solution with a local maximum likelihood. Note that although in general, there may be multiple maxima, and there is no guarantee that the global maximum will be found, in specific cases, for example if the likelihood function is strictly concave, there is only one maximum which is likely to be found via EM. We first describe EM in the context of two simple two-component mixture models: a normal mixture model, and a Bernoulli mixture model.

7.2 Two-Component Gaussian Mixture Model

7.2.1 Problem Formulation

In this section we describe a simple mixture model for density estimation, and the associated EM algorithm for carrying out maximum likelihood estimation. Our data X is samples from

one of two normal distributions. Note that this model can be generalized to a mixture of more than two normal distributions. Our latent data y_i denotes for each sample x_i whether it was sampled from the first distribution or from the second distribution. We will show that given the latent data the model and likelihood function are simplified.

$$X = X_1 \cup X_2, X_1 = \{x_i | x_i \in X, y_i = 1\}, X_2 = \{x_i | x_i \in X, y_i = 2\} \tag{7.1}$$

$$x_i \in X_1 \rightarrow x_i \sim N(\mu_1, \sigma), x_i \in X_2 \rightarrow x_i \sim N(\mu_2, \sigma) \tag{7.2}$$

The probability that a sample in X was sampled from each one of the distributions is given by

$$P(y_i = 1) = p_1; P(y_i = 2) = p_2 = 1 - p_1 \tag{7.3}$$

Given that we know from which distribution the sample was sampled from we can easily compute its probability.

$$P(x_i | y_i = j) = \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-(x_i - \mu_j)^2)}{2\sigma^2}) \tag{7.4}$$

Our goal is to find the model parameters $\theta = \{\mu_1, \mu_2, p_1\}$ that maximize the likelihood:

$$L(\theta) = \prod_i P(x_i|\theta) = \prod_i \sum_j P(x_i, y_i = j|\theta) = \prod_i \sum_j P(y_i = j|\theta) P(x_i|y_i = j, \theta) \tag{7.5}$$

The first equality results from the independence of the samples in X; the second equality results from the law of total probability, and the last equality is due to Bayes' theorem. Finding the maximum of the resulting log likelihood function is therefore not trivial as its partial derivatives depend both on the model parameters and on the latent variables.

$$log(L(\theta)) = \sum_i log \left(\sum_j \frac{p_i}{\sigma\sqrt{2\pi}} exp(\frac{-(x_i - \mu_j)^2)}{2\sigma^2}) \right) \tag{7.6}$$

7.2.2 Simplified Problem: Finding a Local Maximum of the Likelihood Function

Instead of finding the θ that maximizes the likelihood, the EM algorithm finds θ that obtains a local maximum of the likelihood. It starts from an initial guess of the model parameters, and adjusts them iteratively. We will denote the model parameters of iteration t as θ^t, and show that given these parameters we can choose the parameters of the next iteration, denoted as θ such that their likelihood will be greater than the likelihood of θ^t. Meaning, we will show how to adjust θ^t such that

$$logP(X|\theta) \geq logP(X|\theta^t) \tag{7.7}$$

According to Bayes' therem

$$P(x|\theta) = \frac{P(x,y|\theta)}{P(y|x,\theta)} \tag{7.8}$$

Take a log from both sides of the equation

$$logP(x|\theta) = log\frac{P(x,y|\theta)}{P(y|x,\theta)} = logP(x,y|\theta) - logP(y|x,\theta) \tag{7.9}$$

Multiply both sides of the equation by $P(y|x,\theta^t)$

$$P(y|x,\theta^t)logP(x|\theta) = P(y|x,\theta^t)logP(x,y|\theta) - P(y|x,\theta^t)logP(y|x,\theta) \tag{7.10}$$

Sum the above over all y

$$\sum_y P(y|x,\theta^t)logP(x|\theta) = \sum_y P(y|x,\theta^t)logP(x,y|\theta) - \sum_y P(y|x,\theta^t)logP(y|x,\theta) \tag{7.11}$$

We will now simplify the right part of the equation by noting that $logP(x|\theta)$ is independent of y, and that $\sum_y P(y|x,\theta^t) = 1$

$$\sum_y P(y|x,\theta^t)logP(x|\theta) = logP(x|\theta)\sum_y P(y|x,\theta^t) = logP(x|\theta) \tag{7.12}$$

We therefore get the following

$$logP(x|\theta) = \sum_y P(y|x,\theta^t)logP(x,y|\theta) - \sum_y P(y|x,\theta^t)logP(y|x,\theta) \tag{7.13}$$

To adjust the parameters such that $logP(x|\theta) \geq logP(x|\theta^t)$ we will define

$$\Delta = logP(x|\theta) - logP(x|\theta^t) \tag{7.14}$$

and search for θ such that $\Delta \geq 0$.
Let

$$Q(\theta|\theta^t) = \sum_y P(y|x,\theta^t)logP(x,y|\theta) \tag{7.15}$$

Then

$$\Delta = Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \sum_y P(y|x,\theta^t) log \frac{P(y|x,\theta^t)}{P(y|x,\theta)} \tag{7.16}$$

Definition. Kullback-Leiber (KL) divergence: For discrete probability distributions P and Q, the KL divergence of Q from P is the expectation of the logarithmic difference between the probabilities P and Q taken using the probabilities P:

$$D_{KL}(p||q) = \sum_{i\in+} P(x_i) log \frac{Q(x_i)}{P(x_i)} \tag{7.17}$$

Where $i \in +$ denotes that $P(x_i) > 0$. The KL divergence is defined only if $Q(x_i) = 0 \rightarrow P(x_i) = 0$ for all x_i. If $0ln0$ appears in the formula, it is interpreted as zero, because $lim_{x\rightarrow 0} xlnx = 0$

Lemma. Kullback-Leiber divergence is positive: As for all $x > 0$: $x - 1 \geq log(x)$

$$D_{KL}(p||q) \leq \sum_{i\in+} P(x_i) \left(\frac{Q(x_i)}{P(x_i)} - 1 \right) = \sum_{i\in+} Q(x_i) - \sum_{i\in+} P(x_i) = \sum_{i\in+} Q(x_i) - 1 \leq 0 \tag{7.18}$$

Theorem. It is sufficient to maximize $Q(\theta|\theta^t)$ to ensure $\Delta \geq 0$

Proof.

1. Let $\theta^* = argmax_\theta Q(\theta|\theta^t)$ then $Q(\theta^*|\theta^t) - Q(\theta^t|\theta^t) \geq 0$.
2. As KL divergence is positive:

$$\sum_y P(y|x,\theta^t) log \frac{P(y|x,\theta^t)}{P(y|x,\theta)} \geq 0 \tag{7.19}$$

■

Corollary. The main component to maximize is $Q(\theta|\theta^t) = \sum_y P(y|x,\theta^t) logP(x,y|\theta)$.

Q can be represented also as the expectation of $logP(x,y|\theta)$ over the distribution of y given by the current parameters θ^t, that is, $Q(\theta|\theta^t) = E_y[logP(x,y|\theta)]$. $logP(x,y|\theta)$ is also called the complete log-likelihood, which is the likelihood of the model given the observed and latent data. Q is therefore the expectation of the complete log-likelihood.

7.2.3 The EM algorithm:

1. **E-step**: Calculate the Q function.
2. **M-step**: Maximize $Q(\theta|\theta^t)$ with respect to θ.
3. **Stopping criterion**: improvement in log likelihood $\leq \epsilon$.

7.2.4 Application to the Gaussian Mixture Model

Going back to the Gaussian mixture model we now do not need to maximize the complex likelihood function, but the following function

$$Q(\theta|\theta^t) = \sum_y P(y|x,\theta^t) log P(x,y|\theta) \tag{7.20}$$

We will now simplify Q

$$P(x,y|\theta) = \prod_i P(x_i,y_i|\theta) = \prod_i \prod_j P(x_i,y_i=j|\theta)^{y_{ij}}, y_{ij} = \begin{cases} 1, y_i = j \\ 0, y_i \neq j \end{cases}$$

$$log P(x,y|\theta) = \sum_i \sum_j y_{ij} log P(x_i,y_i=j|\theta)$$

$$Q(\theta|\theta^t) = \sum_y P(y|x,\theta^t) \sum_i \sum_j y_{ij} log P(x_i,y_i=j|\theta) \tag{7.21}$$

Rearranging the summation we get

$$Q(\theta|\theta^t) = \sum_i \sum_j \sum_y P(y|x,\theta^t) y_{ij} log P(x_i,y_i=j|\theta) \tag{7.22}$$

As $P(y|x,\theta^t)y_{ij}$ is the only part that depends on y we have

$$Q(\theta|\theta^t) = \sum_i \sum_j log P(x_i,y_i=j|\theta) \sum_y P(y|x,\theta^t) y_{ij} \tag{7.23}$$

As $\sum_y P(y|x,\theta^t)y_{ij} = P(y_{ij}=1|x_i,\theta^t)$

$$Q(\theta|\theta^t) = \sum_i \sum_j log P(x_i,y_i=j|\theta) P(y_{ij}=1|x_i,\theta^t) \tag{7.24}$$

We will define

$$w_{ij}^t = P(y_{ij}=1|x_i,\theta^t) = \frac{P(y_i=j,x_i|\theta^t)}{\sum_j P(y_i=j,x_i|\theta^t)} \tag{7.25}$$

$$Q(\theta|\theta^t) = \sum_i \sum_j w_{ij}^t \left(log \frac{1}{\sqrt{2\pi}} - log\sigma + logp_i - \frac{(x_i-\mu_j)^2}{2\sigma^2} \right) \tag{7.26}$$

Finally we obtained a convenient representation of the concave function Q such that the parameters that maximize it can be easily found based on its partial derivatives.

7.3 Coin Flipping Model

We will now describe another problem that can be solved via EM. Assume we have two coins, each with a different probability to get heads, denoted as h_j. Each time we flip a coin we chose the first coin with probability p_1 and the second coin with probability $p_2 = 1 - p_1$. Our data X consists of a sequence of heads ($x_i = 1$) and tails ($x_i = 0$). Our latent variables are y_i, such that $y_i = j$ denotes that the i^{th} position in the sequence X was obtained by flipping coin j, where $j \in \{1, 2\}$. We want to find the model parameters $\theta = \{h_1, h_2, p_1\}$ that maximize the likelihood of X.

$$L(\theta|X) = P(X|\theta) = \prod_i P(x_i|\theta) = \prod_i \sum_{j=1,2} p_j[h_j^{x_i}(1-h_j)^{1-x_i}] \tag{7.27}$$

The second equality results from the independence of the samples in X; the last equality results from the law of total probability. Once again, it is not trivial to find the maximum of the likelihood function. Luckily, as we have seen in the previous section, the EM algorithm will need to find the maximum of a simpler function - Q.

$$logP(x, y|\theta) = \sum_i logP(x_i, y_i|\theta) = \sum_i log \prod_j P(x_i, y_i = j|\theta)^{y_{ij}} = \sum_i \sum_j y_{ij} logP(x_i, y_i|\theta)$$

where

$$y_{ij} = \begin{cases} 1, y_i = j \\ 0, y_i \neq j \end{cases}$$

$$Q(\theta|\theta^t) = E_y[logP(x, y|\theta)] = E_y\left[\sum_i \sum_j y_{ij} logP(x_i, y_i|\theta)\right] = \sum_i \sum_j E_y(y_{ij}) logP(x_i, y_i|\theta)$$

$$= \sum_i \sum_j E_y(y_{ij}) log(p_j[h_j^{x_i}(1-h_j)^{1-x_i}]) = \sum_i \sum_j E_y(y_{ij})[log(p_j) + x_i log h_j + (1-x_i) log(1-h_j)]$$

$$= \sum_i \sum_j w_{ij}^t [log(p_j) + x_i log h_j + (1-x_i) log(1-h_j)] \tag{7.28}$$

Where w_{ij}^t is defined according to (7.25).

Once again we obtained a simplified representation of Q, such that the partial derivatives can be easily manipulated. For example, to find $p_1^* = argmax_{p_1} Q(\theta|\theta^t)$ we can compute the following partial derivative

$$\frac{\partial Q}{\partial p_1} = \frac{\sum_i w_{i1}^t}{p_1} - \frac{\sum_i w_{i2}^t}{1-p_1} = 0 \tag{7.29}$$

$$p_1^* = \frac{\sum_i w_{i1}^t}{\sum_i w_{i1}^t + \sum_i w_{i2}^t} \tag{7.30}$$

7.4 The Baum-Welch Algorithm

The Baum-Welch Algorithm is an adptation of the generalized EM algorithm used to estimate the parameters of an HMM (so we can use it to run algorithms like Viterbi). The setting of the algorithm is as follows. The hidden data is the vector $y = \pi$ of hidden states travesed by the model. The log-likelihood is:

$$logP(x|\theta) = log \sum_{\pi} P(x, \pi|\theta) \tag{7.31}$$

And the Q function is:

$$Q(\theta|\theta^t) = \sum_{\pi} P(\pi|x, \theta^t) \cdot logP(x, \pi|\theta) \tag{7.32}$$

Let $e_k(b)$ be the emission probability of character b by state $1 \leq k \leq M$. Let $E_k(b, \pi)$ be the number of times we saw b emitted by k in π. Let a_{kl} be the transition probability from state k to state l. Let A_{kl} be the number of times we saw a transition from state k to state l in π. We now have:

$$P(x, \pi|\theta) = \prod_{k=1}^{M} \prod_{b} [e_k(b)]^{E_k(b,\pi)} \cdot \prod_{k=1}^{M} \prod_{l=1}^{M} a_{kl}^{A_{kl}(\pi)} \tag{7.33}$$

By subtituting $P(x, \pi|\theta)$ we obtain:

$$Q(\theta|\theta^t) = \sum_{\pi} P(\pi|x, \theta^t) \left[\sum_{k=1}^{M} \sum_{b} E_k(b, \pi) \cdot log(e_k(b)) + \sum_{k=1}^{M} \sum_{l=1}^{M} A_{kl}(\pi) \cdot log(a_{kl}) \right] \tag{7.34}$$

And by changing the order of summation:

$$Q(\theta|\theta^t) = \sum_{k=1}^{M} \sum_{b} \sum_{\pi} P(, \pi|x, \theta^t) \cdot E_k(b, \pi) \cdot log(e_k(b)) + \\ + \sum_{k=1}^{M} \sum_{l=1}^{M} \sum_{\pi} P(, \pi|x, \theta^t) A_{kl}(\pi) \cdot log(a_{kl}) \tag{7.35}$$

Let $E_k(b)$ be the expected number of times character b is emitted by state k. Let A_{kl} be the expected number of times we had trasition from state k to state l. By definition:

$$\sum_{\pi} P(\pi|x,\theta^t) \cdot E_k(b,\pi) = E_k(b) \tag{7.36}$$

$$\sum_{\pi} P(\pi|x,\theta^t) \cdot A_{kl}(\pi) = A_{kl} \tag{7.37}$$

This is simply because:

$$\sum probability \times value = expectation \tag{7.38}$$

We wish to find a set of parameters θ^{t+1} that maximizes:

$$\sum_{k=1}^{M}\sum_{b} E_k(b) \cdot log(e_k(b)) + \sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl} \cdot log(a_{kl}) \tag{7.39}$$

$E_k(b)$, A_{kl} can be computed using forward / backward:

$$P(\pi_i = k, \pi_{i+1} = l|x,\theta^t) = [1/P(x)] \cdot f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1) \tag{7.40}$$

$$A_{kl} = [1/P(x)] \cdot \sum_{i} f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1) \tag{7.41}$$

And similarly,

$$E_k(b) = [1/P(x)] \cdot \sum_{i|x_i=b} f_k(i) \cdot b_k(i) \tag{7.42}$$

For maximization, we select:

$$a_{ij}^{chosen} = \frac{A_{ij}}{\sum_k A_{ik}}, e_k(b)^{chosen} = \frac{E_k(b)}{\sum_{b'} E_k(b')} \tag{7.43}$$

a_{ij} is merely the weight of transitions from state i to state j relative to the total amount of (expected) transitions from state i to any other state. e_{kb} is the weight of the expected amount of emissions of b by state k divided by the sum of emissions of other characters attributed to state k.

Theorem. The Baum-Welch algorithm increases the likelihood at each step.
Proof. We will show the difference between our chosen estimates and any other estimate is non-negative:

$$\sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl} \cdot log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right) = \sum_{k}\left(\sum_{k'} A_{kk'}\right)\sum_{l}\frac{A_{kl}}{\sum_{k'} A_{kk'}} \cdot log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right) \geq 0 \quad (7.44)$$

And this is true by the Kullback-Liebler divergence property. A similar arument holds for $e_k(b)^{chosen}$.

■

7.4.1 Summary - Parameter Estimation in HMM

Input: $X^1, ..., X^n$ independent training sets.
Baum-Welch alg. (1972):

- **E-step:** By the equation

$$P(\pi_i = k, \pi_{i+1} = l|X, \Theta) = [1/P(X)] \cdot f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1) \quad (7.45)$$

  We compute:

$$A_{kl} = \sum_j [1/P(X^j)] \cdot \sum_i f_k^j(i) \cdot a_{kl} \cdot e_l(x_{i+1}^j) \cdot b_l^j(i+1) \quad (7.46)$$

  Where A_{kl} is the expected no. of $k \to l$ state transitions.

  And:

$$E_k(b) = \sum_j [1/P(X^j)] \cdot \sum_{i|x_i^j=b} f_k^j(i) \cdot b_k^j(i) \quad (7.47)$$

  Where $E_k(b)$ is the expected no. of symbol b appearances attributed to state k.

- **M-step:** re-compute new parameters from A_{kl} , $E_k(b)$ by maximizing the likelihood.

Repeat E-step and M-step until *improvement* $\leq \epsilon$ for some predefined ϵ.

7.5 Bailey and Elkan EM

We introduce a new EM algorithm whose purpose is to find transcription factor motifs. "A sequence **motif** is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance" (from Wikipedia). As usual, we first introduce the setting of the algorithm:

- Model sequences as created from a mixture of motif model (profile of length l) and a background model.
- Sequences are broken into thier n overlapping l-mers $(X_1, ..., X_n)$
- motif model (θ_1) : prob. $f_{i,b}$ for base b at position i of the motif.
- Background model (θ_2): prob. $f_{0,b}$ for base b in a promoter sequence.
- Mixture: model j is used with prob. λ_j. Where $\lambda_1 + \lambda_2 = 1$.
- Motif indicators (hidden): we define indictor variables $Z_{ij} = 1$ iff X_i follows model j.

The log-likelihhood function is:

$$logL(\theta, \lambda|X, Z) = logP(X|\theta, \lambda) = log \sum_z P(X, Z|\theta, \lambda) \tag{7.48}$$

$$P(X, Z|\theta, \lambda) = \prod_i \prod_j [P(X_i|\theta_j)\lambda_j]^{Z_{ij}} \tag{7.49}$$

By subtituting $P(X, Z|\theta, \lambda)$ in the first equation, the complete log-likelihhood is:

$$\begin{aligned} logL(\theta, \lambda|X, Z) &= \sum_{i=1}^{n} \sum_{j=1}^{2} Z_{ij} \cdot log(P(X_i|\theta_j)\lambda_j) = \\ = \sum_{i=1}^{n} log(P(X_i|\theta_1)\lambda_1) + \sum_{i=1}^{n} log(P(X_i|\theta_2)\lambda_2) &= \sum_{i=1}^{n} log\left[\prod_{j=1}^{l} f_{j,X_{ij}}\lambda_1\right] + \sum_{i=1}^{n} log\left[\prod_{j=1}^{l} f_{0,X_{ij}}\lambda_2\right] = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{l} log\left(f_{j,X_{ij}}\lambda_1\right) + \sum_{i=1}^{n}\sum_{j=1}^{l} log\left(f_{0,X_{ij}}\lambda_2\right) \end{aligned} \tag{7.50}$$

E-step (iteration t)

The expected log-likelihood:

$$\begin{aligned} Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) = E[logL(\theta, \lambda|X, Z)] &= \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(t)} log(P(X_i|\theta_j)) + \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(t)} log\lambda_j = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{l} Z_{ij}^{(t)} log(f_{j,X_{ij}}) + \sum_{i=1}^{n}\sum_{j=1}^{l} Z_{ij}^{(t)} log(f_{0,X_{ij}}) + \sum_{i=1}^{n}\sum_{j=1}^{2} Z_{ij}^{(t)} log\lambda_j \end{aligned} \tag{7.51}$$

Where:

$$Z_{ij}^{(t)} = E[Z_{ij}] = P(Z_{ij} = 1|X, \lambda, \theta) = \frac{P(X_i|\theta_j^{(t)})\lambda_j^{(t)}}{\sum_{k=1}^{2} P(X_i|\theta_k^{(t)})\lambda_k^{(t)}} \tag{7.52}$$

Finally, by using the constraint $\lambda_1 + \lambda_2 = 1$, we have for $k = 1, 2$:

$$Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{l} Z_{ij}^{(t)} log(f_{j,X_{ij}}) + \sum_{i=1}^{n}\sum_{j=1}^{l} Z_{ij}^{(t)} log(f_{0,X_{ij}}) + \\ + log\lambda_k \sum_{i=1}^{n} Z_{ik}^{(t)} + log(1-\lambda_k)\sum_{i=1}^{n} Z_{ik}^{(t)} \tag{7.53}$$

M-step (iteration $t + 1$)

We wish to find:

$$(\theta^{(t+1)}, \lambda^{(t+1)}) = \underset{\theta,\lambda}{\text{argmax}} Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) \tag{7.54}$$

First we need to find extrema points of Q, so we take the partial derivative $\frac{\partial Q}{\partial \lambda_1}$ (case: λ_2 is symmetric) :

$$\frac{\partial Q}{\partial \lambda_1} = \frac{1}{\lambda_1}\sum_{i=1}^{n} Z_{i1}^{(t)} + \frac{1}{\lambda_1 - 1}\sum_{i=1}^{n} Z_{i2}^{(t)} = 0 \tag{7.55}$$

We define:

$$A_1 := \sum_{i=1}^{n} Z_{i1}^{(t)} \quad ; \quad A_2 := \sum_{i=1}^{n} Z_{i2}^{(t)} \tag{7.56}$$

We have:

$$\frac{(\lambda_1 - 1)A_1 + \lambda_1 A_2}{\lambda_1(\lambda_1 - 1)} = 0 \tag{7.57}$$

And therefore:

$$\lambda_1 = \frac{A_1}{A_1 + A_2} \tag{7.58}$$

Which implies for $(j = 1, 2)$:

$$\lambda_j = \frac{\sum_{i=1}^{n} Z_{ij}^{(t)}}{\sum_{i=1}^{n}\sum_{k=1}^{2} Z_{ik}^{(t)}} = \frac{\sum_{i=1}^{n} Z_{ij}^{(t)}}{\sum_{i=1}^{n} 1} = \frac{\sum_{i=1}^{n} Z_{ij}^{(t)}}{n} \tag{7.59}$$

By taking the second derivative:

$$\frac{\partial\partial Q}{\partial\lambda_1\partial\lambda_1} = -\frac{A_1}{\lambda_1^2} - \frac{A_2}{(\lambda_1 - 1)^2} < 0 \tag{7.60}$$

We see λ_j is a maximum point which gives us the update rule for λ_j:

$$\lambda_j^{(t+1)} = \sum_{i=1}^{n} \frac{Z_{ij}^{(t)}}{n} \tag{7.61}$$

In a similar manner we get the update rule for $f_{j,k}$:

$$f_{j,k}^{(t+1)} = \frac{c_{jk}}{\sum_{k'=1}^{L} c_{jk'}} \tag{7.62}$$

Where c_{jk} is the expected number of times letter k is produced in column j:

$$c_{0k} = \sum_{i=1}^{n}\sum_{j=1}^{l} Z_{i2}^{(t)} \cdot I(k, X_{ij}) \quad ; \quad c_{jk} = \sum_{i=1}^{n} Z_{i1}^{(t)} \cdot I(k, X_{ij}) \tag{7.63}$$

And $I(k, X_{ij})$ is an indicator variable:

$$I(k, X_{ij}) = \begin{cases} 1 & X_{ij} = k, k \in \Sigma \\ 0 & \text{otherwise} \end{cases} \tag{7.64}$$

7.5.1 Multiple EM for Motif Elicitation

Multiple EM for Motif Elicitation (MEME) is a tool for discovering motifs in a group of related DNA or protein sequences `http://meme.nbcr.net/meme/`[4]. With MEME one can find similar biological functions and structures in different sequences. In order to use MEME one has to carefully choose:

- The best range of widths for the motif.
- The number of occurrences in each sequence.
- The composition of each motif. Does it involove any gaps?

Bibliography

[1] Do, C. B. and S. Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, pages 897–899, 2008.

[2] Frank Dellaert. The expectation maximization algorithm. Technical report, Georgia Institute of Technology, 2002.

[3] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 1977.

[4] Bailey, T. L. and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *International Conference on Intelligent Systems for Molecular Biology; ISMB.*, pages 28–36, 1994.