Expectation Maximization

Daniel Khashabi

Summer 2013
Last Update: December 1, 2015

1 Introduction

Consider the problem of parameter learning by maximizing the likelihood of the observations for random variables $(X, Y) \sim \{(x_i, y_i)\}_{i=1}^{N}$. Assume we model the joint distribution between X and Y using $p(X, Y; \theta)$ which is resulted designer's domain knowledge, and is characterized by the parameter θ.

$$\mathcal{L}(\theta) = \log \prod_{i=1}^{N} p(X = x_i, Y = y_i; \theta) = \sum_{i=1}^{N} \log p(X = x_i, Y = y_i; \theta)$$

To find the ML estimated parameters, one can maximize $\mathcal{L}(\theta)$ with respect to the model parameters θ. But what if we don't observe anything from Y? We call this scenario the "missing data" case. Assume the following definition of the likelihood,

$$\mathcal{L}(\theta) = \log \prod_{i=1}^{N} p(X = x_i; \theta) = \log \prod_{i=1}^{N} \sum_{Y} p(X = x_i, Y = y_i; \theta) = \sum_{i=1}^{N} \log \sum_{Y} p(X = x_i, Y = y_i; \theta)$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \sum_{Y} p(X = x_i, Y = y_i; \theta) \tag{1}$$

Performing parameter learning in the case of missing data (latent variables) by maximizing 1 is not trivial. But EM introduced a formalized way to approximate the maximization, by maximizing the lower-bound on this function.

2 Expectation-Maximization

We first introduce the algorithm, and then analyze its properties. Define the function

$$\mathcal{Q}(\theta; \theta^{(t)}) = \mathbb{E}_{Y|X,\theta^{(t)}} \left[\log p\left(X, Y | \theta\right) \right] \tag{2}$$

The EM algorithm is the following,

- **Initialization:** Initialize the parameters of the mode $\theta^{(0)}$.
- Repeat until convergence:

  1. **Expectation:** Find the expected objective $\mathcal{Q}(\theta; \theta^{(t)})$
  2. **Maximization:** Maximize the EM objective $\mathcal{Q}(\theta; \theta^{(t)})$ with respect to θ:

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{Q}(\theta; \theta^{(t)})$$

We prove this iterations will converge to the maximization of Equation 1 in several steps.

Lemma 1. *The EM objective in Equation 2 is maximizing a lower bound on Equation 1.*

Proof.

$$p(Y|X;\theta) = p(X,Y;\theta)/p(X;\theta)$$

$$\log p(X;\theta) = \log p(X,Y;\theta) - \log p(Y|X;\theta)$$

$$\begin{aligned}\log p(X;\theta) &= \sum_{Y} p(Y|X;\theta^{(t)}) \log p(X,Y;\theta) - \sum_{Y} p(Y|X;\theta^{(t)}) \log p(Y|X;\theta) \\ &= \mathcal{Q}(\theta;\theta^{(t)}) + \mathcal{H}(\theta;\theta^{(t)})\end{aligned}$$

Similarly:

$$\log p(X;\theta^{(t)}) = \mathcal{Q}(\theta^{(t)};\theta^{(t)}) + \mathcal{H}(\theta^{(t)};\theta^{(t)})$$

Using Gibb's inequality we know that

$$\mathcal{H}(\theta;\theta^{(t)}) \geq \mathcal{H}(\theta^{(t)};\theta^{(t)})$$

Which results in

$$\log p(X;\theta) - \log p(X;\theta^{(t)}) \geq \mathcal{Q}(\theta;\theta^{(t)}) - \mathcal{Q}(\theta^{(t)};\theta^{(t)})$$

Which shows improving $\mathcal{Q}$ is lower-bound on improvements in $\log p(X;\theta)$. □

Example 1 (Gaussian Mixture Model)**.** *Suppose we have observations $\{X_i\}_{i=1}^{N}$ and we want to cluster them into K clusters. Define a model as following:*

$Z_n \in \{1, ..., K\}$	*cluster assignment latent variables*
$p(X|Z=k) = \mathcal{N}(\mu_k, 1)$	*probability of sampling a point from cluster k*
$\pi_k = p(Z=k)$	*prior distribution over each cluster*

Therefore the joint model can be written as a mixture of cluster components times their prior:

$$p(X) = \sum_{k=1}^{K} \pi_k p(X|Z=k)$$

The parameters need to be estimated in this model are $\theta = \{\pi_k\}_{k=1}^K \cup \{\mu_k\}_{k=1}^K$. *Finding the latent variables* $\{Z_n\}_{n=1}^N$ *amounts to finding the cluster assignment of the points.*

To derive the EM, we need to estimate the $\mathcal{Q}$ *function. For that we first form the joint distribution of* $p(X, Z; \theta)$.

$$p(X, Z; \theta) = \prod_{k=1}^K [p(X|Z = k)\pi_k]^{1\{Z=k\}}$$
$$\Rightarrow \log p(X, Z; \theta) = \sum_{k=1}^K 1\{Z = k\} \log\left(p(X|Z = k)\pi_k\right)$$

We plug in the observations into the joint distribution:

$$p(\{(X_n = x_n, Z_n)\}_{n=1}^N; \theta) = \prod_{n=1}^N p(X_n = x_n, Z_n; \theta)$$

$$\Rightarrow \log p(\{(X_n = x_n, Z_n)\}_{n=1}^N; \theta) = \sum_{n=1}^N \sum_{k=1}^K 1\{Z_n = l\} \log\left(p(X_n = x_n|Z_n = k; \theta)\pi_k\right)$$

$$\begin{aligned}\Rightarrow \mathcal{Q}(\theta; \theta^{(t)}) &= \mathbb{E}_{Z|X,\theta^{(t)}} \log p(\{(X_n = x_n, Z_n)\}_{n=1}^N; \theta) \\ &= \mathbb{E}_{Z|X,\theta^{(t)}} \sum_{n=1}^N \sum_{k=1}^K 1\{Z_n = k\} \log\left(p(X_n = x_n|Z_n = k; \theta)\pi_k\right) \\ &= \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}_{Z|X,\theta^{(t)}} 1\{Z_n = k\} \log\left(p(X_n = x_n|Z_n = k; \theta)\pi_l\right)\end{aligned}$$

Define $h_{nk}^{(t)} = \mathbb{E}_{Z|X,\theta^{(t)}} 1\{Z_n = k\} = p(Z_n = k|X_n = x_n; \theta^{(t)})$. *This probability tell us, what is the probability that the point* x_n *is assigned to the cluster* k *(and not other clusters).*

$$\mathcal{Q}(\theta; \theta^{(t)}) = \sum_{n=1}^N \sum_{k=1}^K h_{nk} \log\left(p(X_n = x_n|Z_n = k; \theta)\pi_l\right) \tag{3}$$

We can calculate $h_{nk}^{(t)}$ *based on the known distributions in the problem as following using the Bayes rule:*

$$\begin{aligned}h_{nk}^{(t)} = \mathbb{E}_{Z|X,\theta^{(t)}} 1\{Z_n = k\} &= p(Z_n = k|X_n = x_n; \theta^{(t)}) \\ &= \frac{p(X_n = x_n|Z_n = k; \theta^{(t)})p(Z_n = k; \theta^{(t)})}{\sum_{l=1}^K p(X_n = x_n|Z_n = k; \theta^{(t)})p(Z_n = k; \theta^{(t)})} \\ &= \frac{p(X_n = x_n|Z_n = k; \theta^{(t)})\pi_k^{(t)}}{\sum_{l=1}^K p(X_n = x_n|Z_n = l; \theta^{(t)})\pi_l^{(t)}}\end{aligned}$$

Now we have a way to calculate the $\mathcal{Q}(\theta; \theta^{(t)})$ *function (E-step). The next step is to calculate the M-step, i.e maximization of this function with respect to its parameters* θ:

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_{\theta} \mathcal{Q}(\theta|\theta^{(t)}) \\ &= \arg\max_{\theta} \sum_{n=1}^{N}\sum_{k=1}^{K} h_{nk}^{(t)} \log\left(p(X_n = x_n|Z_n = k;\theta)\pi_k\right)\end{aligned}$$

We need to solve this optimization for each element of $\mathcal{Q}$, i.e. for any π_k Note that there is an implicit constraint on this optimization on π_ks, which is $\sum_{k=1}^{K} \pi_k = 1$. To solve this constrained optimization we form the Lagrangian:

$$\begin{aligned}L(\theta,\lambda) &= \sum_{n=1}^{N}\sum_{k=1}^{K} h_{nk}^{(t)} \log\left(p(X_n = x_n|Z_n = k;\theta)\pi_k\right) - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} -\frac{h_{nk}^{(t)}\|x_n-\mu_k\|^2}{2} + h_{nk}^{(t)}\log\frac{1}{\sqrt{2\pi}} + h_{nk}^{(t)}\log\pi_k - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)\end{aligned}$$

$$\frac{\partial L}{\partial \pi_l} = 0 \Rightarrow \sum_{n=1}^{N} h_{nl}^{(t)}/\pi_l = 1 \Rightarrow \pi_l^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N} h_{nl}^{(t)} \tag{4}$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \sum_{k=1}^{K} \pi_k = 1 \tag{5}$$

$$\frac{\partial L}{\partial \mu_l} = 0 \Rightarrow \sum_{n=1}^{N} -h_{nl}^{(t)}(x_n - \mu_l) = 0 \Rightarrow \mu_l^{(t+1)} = \frac{\sum_{n=1}^{N} h_{nl}^{(t)} x_n}{\sum_{n=1}^{N} h_{nl}^{(t)}} \tag{6}$$

Now we have derived all equations necessary for the EM algorithm. In summary, the E-step (estimation of $\mathcal{Q}$) is done using Equation 3 and the M-step (the optimization step) is done using Equation 6 and Equation 4.

Example 2 (Latent categories in documents)**.** *Denote each document with D, which contains a bunch of words, which we denote with W. The assumption is that there is an underlying topical categories which we denote with C.*

- *There are M documents:*
$$D \in \{d_1, \dots, d_M\}$$
- *There are V possible words:*
$$W \in \{w_1, \dots, w_V\}$$
- *There are K possible topical categories*
$$C \in \{c_1, \dots, c_K\}$$

Here is the generative process which defines the model:

- *Prior on each document d: $p(D = d)$*

- *Probability of a category c being inside a document d: $p(C = c|D = d)$*
- *Probability of word w being generated from category c: $p(W = w|C = c)$*

In the input we observe the M documents with their words. In other words for each document d_m, we know the count of word w_v in it, which we denote with $n(w_v, d_m)$.

Given this observation we want to estimate the all of the following probabilities:

$$\theta = \{p(D), p(C|D), p(W|C), \quad \forall D \in \{d_1, \ldots, d_M\}, W \in \{w_1, \ldots, w_V\}, C \in \{c_1, \ldots, c_K\}\}$$

Let's denote the estimations at time t with

$$\theta^{(t)} = \{p_t(D), p_t(C|D), p_t(W|C), \quad \forall D \in \{d_1, \ldots, d_M\}, W \in \{w_1, \ldots, w_V\}, C \in \{c_1, \ldots, c_K\}\}$$

We first write the joint distribution of all variables, given observation from documents. Define $\tilde{W}_j$ and $\tilde{C}_j$ to representing specific word and category at position j of a documents, where $1 \leq j \leq J$ with J being the length of each document.

$$\prod_{m=1}^{M} \prod_{j=1}^{J} p(D = d_m, \tilde{W}_j = \tilde{w}_j, \tilde{C}_j)$$

$$\begin{aligned} p(joint|\theta) = p\begin{pmatrix} \textit{All documents and words and} \\ \textit{their latent category variables} \end{pmatrix} &= \prod_{m=1}^{M} p(\textit{ Words in document } d_m \textit{ and their latent categories }) = \\ &= \prod_{m=1}^{M} \prod_{v=1}^{V} \prod_{k=1}^{K} p(D = d_m, W = w_v, C = c_k)^{n(d_m, w_v) 1\{C = c_k\}} \end{aligned}$$

$$\log p(joint|\theta) = \sum_{m=1}^{M} \sum_{v=1}^{V} n(d_m, w_v) \sum_{k=1}^{K} 1\{C = c_k\} p(D = d_m, W = w_v, C = c_k)$$

$$\begin{aligned} \mathcal{Q}(\theta|\theta^{(t)}) &= \mathbb{E}_{C|W,D,\theta^{(t)}} p(joint|\theta) \\ &= \sum_{m=1}^{M} \sum_{v=1}^{V} n(d_m, w_v) \sum_{k=1}^{K} p_t(C = c_k|D = d_m, W = w_v) p(D = d_m, W = w_v, C = c_k) \\ &= \sum_{m=1}^{M} \sum_{v=1}^{V} n(d_m, w_v) \sum_{k=1}^{K} p_t(C = c_k|D = d_m, W = w_v) p(D = d_m) p(W = w_v|C = c_k) p(C = c_k|D = d_m) \end{aligned}$$

$$\begin{aligned} p_t(C = c_k|D = d_m, W = w_v) &= \frac{p_t(D = d_k) p_t(W = w_v|C = c_k) p_t(C = c_k|D = d_m)}{\sum_{k=1}^{K} p_t(D = d_k) p_t(W = w_v|C = c_k) p_t(C = c_k|D = d_m)} \\ &= \frac{p_t(W = w_v|C = c_k) p_t(C = c_k|D = d_m)}{\sum_{k=1}^{K} p_t(W = w_v|C = c_k) p_t(C = c_k|D = d_m)} \end{aligned}$$

Now we need to find the maximizers of $\mathcal{Q}(\theta|\theta^{(t)})$. We form the Lagrangian:

$$L = \mathcal{Q}(\theta|\theta^{(t)})+\alpha \Bigg(1-\sum_{m=1}^{M} p(D=d_m)\Bigg)+\sum_{k=1}^{K}\beta_k \Bigg(1-\sum_{v=1}^{V} p(W=w_v|C=c_k)\Bigg)+\sum_{m=1}^{M}\gamma_m \Bigg(1-\sum_{k=1}^{K} p(C=c_k|D=d_m$$

$$\frac{\partial L}{\partial p(D=d_k)} = 0 \Rightarrow p_{t+1}(D=d_m) = \frac{\sum_{v=1}^{V} n(d_m,w_v)}{\sum_{m=1}^{M}\sum_{v=1}^{V} n(d_m,w_v)}$$

And similarly for other two parameters we should have:

$$p_{t+1}(C=c_k|D=d_m) = \frac{\sum_{v=1}^{V} n(d_m,w_v)p_t(C=c_k|D=d_m,W=w_v)}{\sum_{v=1}^{V} n(d_m,w_v)}$$

$$p_{t+1}(W=w_v|C=c_k) = \frac{\sum_{m=1}^{M} n(d_m,w_v)p_t(C=c_k|D=d_m,W=w_v)}{\sum_{m=1}^{M}\sum_{v=1}^{V} n(d_m,w_v)p_t(C=c_k|D=d_m,W=w_v)}$$

Example 3. *Consider the following simple mixture model:*

$$\begin{cases} p_1(x;\theta) = e^{-g_1(x;\theta)} \\ p_2(x;\theta) = e^{-g_2(x;\theta)} \\ p(x;\theta) = \mu_1 p_1 + (1-\mu_1)p_2 \end{cases}$$

Given the set of observations, $x_1,\ldots,x_n$, we want to estimate parameters of this mixture model $\Theta = \{\theta_1,\theta_2,\mu_1\}$.
The conventional maximum likelihood aims at solving the following problem:

$$\Theta := \arg\max_{\Theta}\left\{\sum_i \log\left[\mu_1 e^{-g_1(x_i;\theta)} + (1-\mu_1)e^{-g_2(x_i;\theta)}\right]\right\}$$

this optimization is slightly hard, as we have logarithm of some summation. We will add some additional variables to the model to make the optimization steps more explicit. We assume that each data is coming from one specific component. For that, we define additional variable to specify the component from which the sample is coming from:

$$\delta_i = \begin{cases} 1 & \textit{if the sample is coming from the first component} \\ 0 & \textit{if the sample is coming from the second component} \end{cases}$$

The complete data likelihood is:

$$\mathcal{L}(\Theta) = \sum_i \log p(x_i,\delta_i|\Theta) = \sum_i \log p(x_i|\delta_i,\Theta) + \log p(\delta_i|\Theta)$$

Sometimes people call the above likelihood $F(\Theta,\delta)$, since we don't know the δ_i values and they need to be estimated. Therefore we can't compute the above likelihood. But we can assume a parametric form for distribution of δ_i, given an estimate to parameters Θ^n (at the n-th step), which we denote with $p(\delta|\Theta^n,X)$. To remove the random variable δ from the full-model likelihood we marginalize over δ:

$$Q(\Theta;\Theta^n) = \mathbb{E}_{p(\delta|\Theta^n,X)}\left[F(\Theta,\delta)\right]$$

Now the EM updates are the followings:

$$\begin{cases} \textit{E: Estimate the distribution of } p(\delta|\Theta^n, X)\textit{, and find } Q(\Theta;\Theta^n). \\ \textit{M: Find } \Theta^{n+1} := \arg\max_{\Theta} Q(\Theta;\Theta^n) \end{cases}$$

Let's simplify this. We know:

$$\begin{aligned} \mathcal{L}(\Theta) &= \sum_i \left[\log p(x_i|\delta_i, \Theta) + \log p(\delta_i|\Theta)\right] && (7) \\ &= \sum_i \left[-\delta_i g_1(x_i;\theta_1) - (1-\delta_i) g_2(x_i;\theta_2) + \delta_i \ln \mu_1 + (1-\delta_i)\ln(1-\mu_1)\right] && (8) \end{aligned}$$

Now we need to estimate $p(\delta|\Theta^n, X)$. Based on its definition we have:

$$\begin{aligned} p(\delta_i = 1|\Theta^n, x_i) &= \frac{p(\delta_i = 1, x_i|\Theta^n)}{p(x_i, \delta_i = 1|\Theta^n) + p(x_i, \delta_i = 0|\Theta^n)} \\ &= \frac{p(x_i|\delta_i = 1, \Theta^n) p(\delta_i = 1|\Theta^n)}{p(x_i|\delta_i = 1, \Theta^n) p(\delta_i = 1|\Theta^n) + p(x_i|\delta_i = 0, \Theta^n) p(\delta_i = 0|\Theta^n)} \\ &= \frac{\mu_1^n e^{-g_1(x_i;\theta_1^n)}}{\mu_1^n e^{-g_1(x_i;\theta_1^n)} + (1-\mu_1^n) e^{-g_2(x_i;\theta_2^n)}} \end{aligned}$$

If you are not convinced that this is a good estimation for $p(\delta_i|\Theta^n, x_i)$, we can derive it in different way. Consider $F(\Theta, \delta)$. Whatever distribution we choose for δ it needs to maximize this function. Thus we take differentiation with respect to δ (Note that this differentiation is with respect to a function, which is called functional derivative).

$$\nabla_{\delta_i} F(\Theta^n, \delta) = 0$$

$$\begin{aligned} \Rightarrow \nabla_{\delta_i} F(\Theta^n, \delta) = &- \left[\log \delta_i - \log(1-\delta_i)\right] \\ &+ \left[-g_1(x_i;\theta_1^n) + \log \mu_1^n\right] \\ &+ \left[-g_2(x_i;\theta_2^n) + \log(1-\mu_1^n)\right] = 0 \end{aligned}$$

Which will result in the same distribution for $p(\delta_i|\Theta^n, x_i)$. Given this closed form estimation for $p(\delta_i|\Theta^n, x_i)$ it is easy to find $Q(\Theta;\Theta^n)$, by plugging it into Equation 7, and maximizing it with respect Θ.

Remark 1. *One interpretation of EM is coordinate-descent optimization, over latent variable coordinate, and the observation coordinates. In the previous example we solved the E-step with another optimization over the latent variables.*

Example 4 (A simple Bayesian network(from [**?**]))**.** *Assume that a set of 3-dimensional points (x, y, z) is generated according to the following probabilistic generative model over Boolean variables $X, Y, Z \in \{0, 1\}$:*

$$Y \leftarrow X \rightarrow Z$$

1. *What parameters from the table bellow will you need to estimate in order to completely define the model?*

(1) P(X=1)	*(2)* P(Y=1)	*(3)* P(Z=1)	
(4) P(X|Y=b)	*(5)* P(X|Z=b)	*(6)* P(Y|X=b)	*(7)* P(Y|Z=b)
(8) P(Z|X=b)	*(9)* P(Z|Y=b)	*(10)* P(X|Y=b,Z=c)	*(11) 3*

Answer: *Based on the above generative model we could write the joint distribution as following:*

$$p(X,Y,Z) = p(X).p(Y|X).p(Z|X).$$

So we need to have (1), (6), (8). For this problem in order to find the whole joint distribution we need to know five parameters. For simplicity we denote the parameters using the following:

$$\begin{aligned}
&p(X=1) = \alpha \\
&p(Y=1|X=1) = a_1 \\
&p(Y=1|X=0) = a_2 \\
&p(Z=1|X=1) = b_1 \\
&p(Z=1|X=0) = b_2
\end{aligned}$$

Then we have:

$$\begin{aligned}
&p(X=x) = \alpha^x(1-\alpha)^{1-x} \\
&p(Y=y|X=1) = a_1^y(1-a_1)^{1-y} \\
&p(Y=y|X=0) = a_2^y(1-a_2)^{1-y} \\
&p(Z=z|X=1) = b_1^z(1-b_1)^{1-z} \\
&p(Z=z|X=0) = b_2^z(1-b_2)^{1-z}
\end{aligned}$$

2. *You are given a sample of m data points sampled independently at random. However, when the observations are given to you, the value of X is always omitted. Hence, you get to see $\{(y^1,z^1),\ldots,(y^m,z^m)\}$. In order to estimate the parameters you identified in part (a), in the course of this question you will derive update rules for them via the EM algorithm for the given model.*

Express $\Pr(y^j,z^j)$ *for an observed sample* (y^j,z^j) *in terms of the unknown parameters.*
Answer: *We can use the joint distribution and integrate out the unseen variables:*

$$\begin{aligned}
Pr(y^j,z^j) &= \sum_i Pr(X=x^i,y^j,z^j) \\
&= \sum_i p(X=x^i).p(Y=y^j|X=x^i).p(Z=z^j|X=x^i)
\end{aligned}$$

We can replace each term with its own parameter denoted in the previous part. -Let $p_i^j = Pr(X{=}i|y^j,z^j)$ *be the probability that hidden variable X has the value $i \in \{0,1\}$ for an*

observation $(y^j, z^j), j \in \{1, \ldots, m\}$. Express p_i^j in terms of the unknown parameters.
Answer: *Using the Bayes law we could express the probability like this:*

$$\begin{aligned} p_i^j = Pr(X{=}i|y^j, z^j) &= Pr(y^j, z^j|X{=}i).Pr(y^j, z^j)/Pr(X{=}i) \\ &= Pr(y^j|X{=}i).Pr(z^j|X{=}i).Pr(y^j, z^j)/Pr(X{=}i) \end{aligned}$$

Each of the terms are previously calculated.

3. *Let (x^j, y^j, z^j) represent the completed j^{th} example, $j \in \{1, \ldots, m\}$. Derive an expression for the expected log likelihood (LL) of the completed data set $\{(x^j, y^j, z^j)\}_{j=1}^m$, given the parameters in (a).*
Answer:
$LL = \sum_j \log p(x^j, y^j, z^j) = \sum_j \log p(X = x^j) + \sum_j \log p(Y = y^j|X = x^j) + \sum_j \log p(Z = z^j|X = x^j)$ - Maximize LL, and determine update rules for any two unknown parameters of your choice (from those you identified in part (a)).
Answer: *Because we don't haven't seen the latent variable values x we cannot explicitly plug in their values into the log-likelihood. But instead we need to takes its expectation with respect to the variable. This corresponds to the E-step in EM algorithm:*

$$Q(y, z) = \sum_x p(x|y, z) \log p(x, y, z)$$

In the next step we shall maximize the expected likelihood with respect to the parameters:

$$\theta = \arg\max_\theta Q(y, z)$$

We can calculate the expectation of the whole-data likelihood as follows:

$$Q = \mathbb{E}\left[\sum_j \log p(x^j, y^j, z^j)\right] = \sum_j \mathbb{E}\left[\log p(x^j, y^j, z^j)\right] = \sum_j [p_1^j A + p_0^j B]$$

Where,

$$A = y^j \log b_1 + (1 - y^j)\log(1 - b_1) + z^j \log a_1 + (1 - z^j)\log(1 - a_1) + \log\alpha$$

$$B = y^j \log b_2 + (1 - y^j)\log(1 - b_2) + z^j \log a_2 + (1 - z^j)\log(1 - a_2) + \log(1 - \alpha)$$

To maximize the above expression, we must take derivitive with respect to the parameters:

$$\frac{\partial Q}{\partial \alpha} = \sum_j p_1^j \frac{1}{\alpha} - \sum_j p_0^j \frac{1}{1 - \alpha} = 0 \Rightarrow \alpha = \frac{\sum_j p_1^j}{\sum_j p_1^j + \sum_j p_0^j}$$

$$\frac{\partial Q}{\partial b_1} = \sum_j p_1^j y^j \frac{1}{b_1} - \sum_j p_1^j (1 - y^j) \frac{1}{1 - b_1} = 0 \Rightarrow b_1 = \frac{\sum_j p_1^j y^j}{\sum_j p_1^j}$$