EXPECTATION MAXIMIZATION: A TUTORIAL

1. Introduction

I was motivated to write this tutorial because I was not able to find an easily-followed derivation of the EM equations showing clearly how the E-step and M-step are derived using, as far as possible, only an elementary application of conditional probabilities and Bayes' theorem. This derivation assumes all random variables are discrete, but it can be easily adapted to continuous distributions.

2. Maximizing the Log Likelihood

We start with a distribution over a random variable X, parametrized by some parameters θ. Let's denote this distribution by the mass function

$$P(X = x|\theta) \coloneqq p(x|\theta),$$

where we explicitly show the parameter dependence by the conditioning. In addition, we will use the notation

$$p(x = x_0|\theta)$$

to denote the value of the probability at $x = x_0$.

We know that there exist some hidden variables upon which this probability depends, denoted by random vector Z. Our aim is to find

$$\theta^* \coloneqq \arg\max_\theta \log p(x|\theta), \tag{1}$$

which are the parameters that maximize the log likelihood of $p(x|\theta)$. We can capture that dependence on the hidden variables by marginalizing over the joint distribution $p(x, z|\theta)$ to write

$$\log p(x|\theta) = \log \sum_z p(x, z|\theta) = \log \sum_z p(x|z, \theta)p(z|\theta)$$

Maximizing the RHS directly is challenging, since it includes a weighted sum involving a number of distributions of x conditioned on the hidden variables. Since each of these may have their own local maxima, the marginal distribution will usually end up being multi-modal, making it challenging to locate the global maximizer using conventional numerical methods, which may be ill-conditioned.

3. Approximating the Log Likelihood

Instead, we can try approximating the conditional distribution of z given x with an estimate for our parameters. To see how this can be achieved, note that by the definition of conditional probability,

$$p(x, z|\theta) = p(z|x, \theta)p(x|\theta)$$

Rearranging gives us

$$p(x|\theta) = \frac{p(x, z|\theta)}{p(z|x, \theta)}$$

Then, we can multiply both sides by $p(z|x,\theta)$ to give

$$p(x|\theta)p(z|x,\theta) = \frac{p(x,z|\theta)}{p(z|x,\theta)}p(z|x,\theta)$$

We know that $p(z|x,\theta)/p(z|x,\theta) = 1$, so we can replace this fraction by any term that evaluates to 1. Since we are trying to estimate the parameter vector, let's use our current best estimate for it at update step k, which we denote $\theta = \hat{\theta}^{(k)}$. Then, we can replace $p(z|x,\theta)$ by $p(z|x,\hat{\theta}^{(k)})$ in the fraction to give

$$p(x|\theta)p(z|x,\theta) = \frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(k)})}p(z|x,\hat{\theta}^{(k)})$$

Now, we sum both sides over z:

$$\sum_z p(x|\theta)p(z|x,\theta) = \sum_z \frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(k)})}p(z|x,\hat{\theta}^{(k)})$$
$$p(x|\theta)\sum_z p(z|x,\theta) = \sum_z \frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(k)})}p(z|x,\hat{\theta}^{(k)})$$

Because $p(z|x,\theta)$ is a probability distribution, it sums to unity, and vanishes on the LHS. Noticing that the sum on the RHS defines the conditional expectation over z, it follows that

$$p(x|\theta) = \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(k)})}\right]$$

Now, let's go back to the log likelihood, which is what we were interested in originally. Taking logarithms of both sides gives

$$\log p(x|\theta) = \log \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(k)})}\right] \tag{2}$$

We now have to make use of a result from convex analysis called *Jensen's Inequality*. This result states that we can switch the order of a convex function and an expectation. For a given convex function $f(x)$, Jensen's inequality states that

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

If $f(\cdot)$ is instead concave, like the log function, then $-f(\cdot)$ is convex, so the inequality becomes

$$f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$$

Applying this result to (2),

$$\log p(x|\theta) \geq \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log\left(\frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(k)})}\right)\right] =: \ell(\theta,\hat{\theta}^{(k)}), \tag{3}$$

where we will use the function $\ell(\cdot,\cdot)$ as a shorthand to represent the conditional expectation of the logarithm on the RHS. Notice that the inequality becomes an equality when $\theta = \hat{\theta}^{(k)}$. That is, $\ell(\hat{\theta}^{(k)},\hat{\theta}^{(k)}) = \log p(x|\hat{\theta}^{(k)})$ (prove this!)

We are now interested in how, by repeatedly maximizing a lower bound on the log likelihood, represented by $\ell(\cdot,\cdot)$, we can get close to the parameter θ that maximizes

the original log likelihood $\log p(x|\theta)$. This turns out to be relatively simple. Given $\hat{\theta}^{(k)}$, if we choose the updated parameter at step $k+1$ as

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta} \ell(\theta, \hat{\theta}^{(k)}), \tag{4}$$

then notice that

$$\log p(x|\hat{\theta}^{(k)}) = \ell(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}) \leq \max_{\theta} \ell(\theta, \hat{\theta}^{(k)}) = \ell(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}) \leq \log p(x|\hat{\theta}^{(k+1)}),$$

Each time we update θ using (4), we end up with a parameter that increases our log likelihood, provided we are not at a local maximum already. This forms the basis for the EM algorithm.

4. The EM algorithm

It turns out we do not need solve the full optimization problem (4). We can use the properties of the logarithm function and expected values to break (3) into two components, an *energy* term and an *entropy* term:

$$\begin{aligned}\log p(x|\theta) &\geq \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log p(x,z|\theta) - \log p(z|x,\hat{\theta}^{(k)})\right] \\ &= \underbrace{\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log p(x,z|\theta)\right]}_{\text{energy}} - \underbrace{\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log p(z|x,\hat{\theta}^{(k)})\right]}_{\text{entropy}}\end{aligned}$$

The energy term is the only one that varies with θ (the entropy term stays constant as θ is varied), so to find the maximizing θ in (4), we only need to consider the energy term. Once we find this θ, we can update our parameter estimate to this new value. Assuming we start with an initial guess for the parameter vector $\hat{\theta}^{(0)}$, we can iteratively update this and increase the log likelihood monotonically by applying the following two steps:

E-step: Given the current parameter estimate $\hat{\theta}^{(k)}$, calculate the conditional distribution $p(z|x,\hat{\theta}^{(k)})$. We can use Bayes' theorem to write

$$\begin{aligned}p(z|x,\hat{\theta}^{(k)}) &= \frac{p(x|z,\hat{\theta}^{(k)})p(z|\hat{\theta}^{(k)})}{p(x|\hat{\theta}^{(k)})} \\ &= \frac{p(x|z,\hat{\theta}^{(k)})p(z|\hat{\theta}^{(k)})}{\sum_{z'} p(x|z',\hat{\theta}^{(k)})p(z'|\hat{\theta}^{(k)})}\end{aligned}$$

We will generally be able to compute all of the terms in this expression easily. For example, if z is a categorical variable, the distribution of x will be conditional on the category. The denominator normalizes the expression so that it is a probability distribution (i.e. sums to unity). Using this distribution, compute the expectation of the log likelihood

$$\ell(\theta, \hat{\theta}^{(k)}) = \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log p(x,z|\theta)\right]$$

M-step: Maximize the expectation computed in the E-step with respect to θ and use this to update the parameter estimate according to

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta} \ell(\theta, \hat{\theta}^{(k)})$$

To find the parameter updates in practice, we can find the gradient of $\ell(\cdot,\cdot)$ with respect to θ and set it to zero to find the critical point(s).

We repeat the E-step and M-step until convergence is reached. Typically, we will check that $\|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\|$ is sufficiently small.

5. Example: Coin Tossing

We will now consider a simple example of EM applied to a coin-tossing experiment, as described in [1]. We start with two biased coins. At each iteration of our experiment, we choose one of the coins with some unknown probability of selection, and then toss it M times, recording the number of heads that we observe. We repeat this experiment N times, and want to estimate the biases on each coin, as well as the probability of choosing a certain coin. In this case, the hidden variable indicates which coin was picked for each experiment.

5.1. Mathematical Description of Problem.

Let α_1 be the bias on one coin, let α_2 be the bias on the other coin, and let π_1 and π_2 be the probabilities of choosing each coin respectively. Note that π_1 and π_2 are not independent and must sum to 1. The parameter vector we wish to estimate is $\theta = [\alpha_1, \alpha_2, \pi_1, \pi_2]$. We can now define our distributions. Firstly, the distribution of choosing coin z_n on the nth set of tosses can be written

$$p(z_n|\theta) = \pi_1^{\mathbb{I}[z_n=1]}\pi_2^{\mathbb{I}[z_n=2]} = \prod_{i=1}^{2} \pi_i^{\mathbb{I}[z_n=i]}, \tag{5}$$

where we have used the *indicator function* $\mathbb{I}[\cdot]$ that returns 1 when the expression inside the brackets is true, and 0 otherwise. It allows us to combine the two probabilities into a single expression. Work through it and verify for yourself that when $z = 1$ the probability is π_1 and when $z = 2$, the probability is π_2. Assuming each coin toss is independent, we can define the probability over all tosses by

$$p(\boldsymbol{z}|\theta) = \prod_{n=1}^{N} p(z_n|\theta),$$

where $\boldsymbol{z} \coloneqq [z_1, z_2, \ldots, z_N]$. We can also now define the probability of getting x_n heads on the nth set of M tosses, which for a fixed z_n, follows a binomial distribution. Taking into account both coins, we can write this as

$$p(x_n|z_n, \theta) = \binom{M}{x_n} \prod_{i=1}^{2} \left(\alpha_i^{x_n}(1-\alpha_i)^{(M-x_n)}\right)^{\mathbb{I}[z_n=i]} \tag{6}$$

Then, the conditional probability over all tosses is given by

$$p(\boldsymbol{x}|\boldsymbol{z}, \theta) = \prod_{n=1}^{N} p(x_n|z_n, \theta),$$

where $\boldsymbol{x} \coloneqq [x_1, x_2, \ldots, x_N]$. This follows from assuming independence of each set of M tosses.

5.2. **E-Step.** We compute the distribution

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{x},\hat{\theta}^{(k)}) &= \frac{p(\boldsymbol{x}|\boldsymbol{z},\hat{\theta}^{(k)})p(\boldsymbol{z}|\hat{\theta}^{(k)})}{\sum_{\boldsymbol{z'}} p(\boldsymbol{x}|\boldsymbol{z'},\hat{\theta}^{(k)})} \\ &= \frac{\prod_{n=1}^{N} p(x_n|z_n,\hat{\theta}^{(k)})p(z_n|\hat{\theta}^{(k)})}{\sum_{\boldsymbol{i}\in\{1,2\}^N}\prod_{n=1}^{N} p(x_n|z_n=i_n,\hat{\theta}^{(k)})p(z_n|\hat{\theta}^{(k)})} \\ &= \frac{\prod_{n=1}^{N} p(x_n|z_n,\hat{\theta}^{(k)})p(z_n|\hat{\theta}^{(k)})}{\prod_{n=1}^{N}\sum_{i_n=1}^{2} p(x_n|z_n=i_n,\hat{\theta}^{(k)})p(z_n|\hat{\theta}^{(k)})}, \end{aligned}$$

Note that denominator is summed over all possible combinations of coin chosen for the N sets of ten coin tosses. We use the vector $\boldsymbol{i} = [i_1, i_2, \ldots, i_N]$ as a convenience to represent N nested summations, where each element of $\boldsymbol{i}$ is an index of summation that varies from 1 to 2. These nested summations are then simplified by factoring the denominator, given that the summations are separable in each z_n. Once we have this distribution, computing the expectation $\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\cdot]$ gives

$$\begin{aligned} \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] &= \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)p(\boldsymbol{z}|\theta)] \\ &= \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)] + \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\log p(\boldsymbol{z}|\theta)] \end{aligned} \quad (7)$$

5.3. **M-Step.** Notice that the first term of (7) depends only on parameters α_1 and α_2 and the second term depends only on parameters π_1 and π_2. We can therefore maximize each term separately with respect to its parameters.

Consider the first term.

$$\begin{aligned} \log p(\boldsymbol{x}|\boldsymbol{z},\theta) &= \sum_{n=1}^{N} \log p(x_n|z_n,\theta) \\ &= \sum_{n=1}^{N} \log \binom{M}{x_n} + \sum_{n=1}^{N}\sum_{i=1}^{2} \mathbb{I}[z_n=i](x_n \log\alpha_i + (M-x_n)\log(1-\alpha_i)), \end{aligned}$$

where we have substituted (6) and used properties of the logarithm. We now need to take the expectation with respect to the conditional distribution on the hidden variables using the parameter estimates:

$$\begin{aligned} \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)] &= \sum_{n=1}^{N} \log \binom{M}{x_n} \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{2} \mathbb{E}_{Z|X,\hat{\theta}^{(k)}}[\mathbb{I}[z_n=i]]\left(x_n\log\alpha_i + (M-x_n)\log(1-\alpha_i)\right), \end{aligned} \quad (8)$$

where we have used the fact that the expectation can be taken inside the summations, as well as the fact that the only term in the expression that depends on the hidden variables is the indicator function $\mathbb{I}[z_n = i]$. To find the maximum, we need

to take the partial derivatives of (8) with respect to α_1 and α_2 and set these to zero. We will cover both these cases by taking the derivative with respect to α_j, for $j = 1, 2$. Then, we have

$$\frac{\partial}{\partial \alpha_j} \mathbb{E}_{Z|X,\hat{\theta}^{(k)}} \left[\log p(\boldsymbol{x}|\boldsymbol{z}, \theta)\right] = \sum_{n=1}^{N} \mathbb{E}_{Z|X,\hat{\theta}^{(k)}} \left[\mathbb{I}[z_n = j]\right] \left(\frac{x_n}{\alpha_j} - \frac{M - x_n}{1 - \alpha_j}\right) \tag{9}$$

Given that we are differentiating only with respect to α_j, any constants or terms involving α_i, $i \neq j$ vanish, so the binomial coefficient and the summation over $i = 1, 2$ disappear.

We can compute $\mathbb{E}_{Z|X,\hat{\theta}^{(k)}} \left[\mathbb{I}[z_n = j]\right]$ by applying the law of the unconscious statistician, which allows us to write the expectation of a function of a random variable as a probability-weighted sum of function values of the random variable. Writing this summation,

$$\mathbb{E}_{Z|X,\hat{\theta}^{(k)}} \left[\mathbb{I}[z_n = j]\right] = \sum_{\boldsymbol{i} \in \{1,2\}^N} \mathbb{I}[z_n = j] p(\boldsymbol{z} = \boldsymbol{i}|\boldsymbol{x}, \hat{\theta}^{(k)}),$$

where $\boldsymbol{i} \coloneqq [i_1, i_2, \dots\ i_N]$ are the index variables of the N summations as before. From this expression, we observe that the indicator function is simply "picking out" the events where $z_n = j$ (it has value zero everywhere else). Then, we can rewrite the expression using the independence assumption, factorize and choose only those terms of the summation where $z_n = j$:

$$\begin{aligned}
\mathbb{E}_{Z|X,\hat{\theta}^{(k)}} \left[\mathbb{I}[z_n = j]\right] &= \sum_{\boldsymbol{i} \in \{1,2\}^N} \prod_{q=1}^{N} \mathbb{I}[z_n = j] p(z_q = i_q|x_q, \hat{\theta}^{(k)}) \\
&= \prod_{q=1}^{N} \sum_{i_q=1}^{2} \mathbb{I}[z_n = j] p(z_q = i_q|x_q, \hat{\theta}^{(k)}) \\
&= p(z_n = j|x_n, \hat{\theta}^{(k)}) \prod_{\substack{q=1 \\ q \neq n}}^{N} \sum_{i_q=1}^{2} p(z_q = i_q|x_q, \hat{\theta}^{(k)}) \qquad (10) \\
&= p(z_n = j|x_n, \hat{\theta}^{(k)}), \qquad (11)
\end{aligned}$$

where the summations on the RHS of (10) are all unity (why?) This shows that the expectation of the indicator function is simply given by the probability of choosing coin j on the nth toss, given the number of heads x_n and estimated parameter vector $\hat{\theta}^{(k)}$. However, it is easier for us to use Bayes' theorem to rewrite this as

$$p(z_n = j|x_n, \hat{\theta}^{(k)}) = \frac{p(x_n|z_n = j, \hat{\theta}^{(k)}) p(z_n = j|\hat{\theta}^{(k)})}{\sum_{i=1}^{2} p(x_n|z_n = i, \hat{\theta}^{(k)}) p(z_n = i|\hat{\theta}^{(k)})} \tag{12}$$

We can now substitute (5) and (6) into (12) and replace the actual parameter vector with its estimate

$$\hat{\theta}^{(k)} \coloneqq [\hat{\pi}_1^{(k)}, \hat{\pi}_2^{(k)}, \hat{\alpha}_1^{(k)}, \hat{\alpha}_2^{(k)}]. \tag{13}$$

Making the substitution gives

$$p(z_n = j|x_n, \hat{\theta}^{(k)}) = \frac{\hat{\pi}_j^{(k)} \hat{\alpha}_j^{(k)x_n} (1 - \hat{\alpha}_j^{(k)})^{(M-x_n)}}{\sum_{i=1}^{2} \hat{\pi}_i^{(k)} \hat{\alpha}_i^{(k)x_n} (1 - \hat{\alpha}_i^{(k)})^{(M-x_n)}}. \tag{14}$$

Note that again, the binomial coefficients in the numerator and denominator cancel out. Now, substituting (11) into (9) and setting this to zero to find the updated parameter value,

$$\frac{\partial}{\partial \alpha_j} \mathbb{E}_{Z|X,\hat{\theta}^{(k)}} \left[\log p(\boldsymbol{x}|\boldsymbol{z}, \theta)\right] = \sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)}) \left(\frac{x_n}{\alpha_j} - \frac{M - x_n}{1 - \alpha_j}\right) = 0$$

Rearranging and solving for α_j gives

$$\alpha_j = \frac{\sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)}) x_n}{\sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)}) x_n + p(z_n = j|x_n, \hat{\theta}^{(k)})(M - x_n)},$$

where $p(z_n = j|x_n, \hat{\theta}^{(k)})$ is computed as in (14). This gives us the maximum likelihood estimate for the coin biases, which we can use to update the parameter values. If instead of being latent, we consider the probabilities of choosing the coins to be known parameters with $\pi_1 = \pi_2 = 0.5$ and set $M = 10$ and $N = 5$ then this expression matches the calculations in [1], where $10 - x_n$ is the number of tails in the nth set of 10 tosses. This expression can be further simplified to give

$$\alpha_j = \frac{\sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)}) x_n}{M \sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)})}$$

We now turn to estimating the probability of choosing a particular coin. This is slightly more complicated than the coin biases, since the probabilities of choosing the coins are not independent, but must sum to 1. We write this constraint in the form

$$g(\pi_1, \pi_2) = \pi_1 + \pi_2 - 1 = 0 \tag{15}$$

The partial derivatives with respect to π_j of the second term of (7) are given by

$$\begin{aligned}\frac{\partial}{\partial \pi_j}\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log p(\boldsymbol{z}|\theta)\right] &= \frac{\partial}{\partial \pi_j}\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log \prod_{n=1}^{N}\prod_{i=1}^{2}\pi_i^{\mathbb{I}[z_n=i]}\right] \\ &= \frac{\partial}{\partial \pi_j}\sum_{n=1}^{N}\sum_{i=1}^{2}\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\mathbb{I}[z_n=i]\right]\log \pi_i \\ &= \frac{1}{\pi_j}\sum_{n=1}^{N}\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\mathbb{I}[z_n=j]\right] \\ &= \frac{1}{\pi_j}\sum_{n=1}^{N} p(z_n=j|x_n,\hat{\theta}^{(k)}), \end{aligned} \tag{16}$$

where we have substituted (11) for the expected value of the indicator function as calculated before. Given the equality constraint, we cannot simply set the derivatives (16) to zero. Using the method of Lagrange multipliers, we instead need to ensure that

$$\frac{\partial}{\partial \pi_j}\left(\mathbb{E}_{Z|X,\hat{\theta}^{(k)}}\left[\log p(\boldsymbol{z}|\theta)\right] - \lambda g(\pi_1,\pi_2)\right) = 0,$$

for Lagrange multiplier λ and $j = 1, 2$. Then, it follows that

$$\frac{1}{\pi_j}\sum_{n=1}^{N} p(z_n=j|x_n,\hat{\theta}^{(k)}) = \lambda$$

This means that we can equate the LHS for $j = 1, 2$ and use (15) to give

$$\frac{1}{\pi_j}\sum_{n=1}^{N} p(z_n=j|x_n,\hat{\theta}^{(k)}) = \frac{1}{1-\pi_j}\sum_{n=1}^{N}(1-p(z_n=j|x_n,\hat{\theta}^{(k)})),$$

where we have used the fact that

$$\sum_{i=1}^{2} p(z_n=i|x_n,\hat{\theta}^{(k)}) = 1$$

Rearranging this,

$$\pi_j\sum_{n=1}^{N}(1-p(z_n=j|x_n,\hat{\theta}^{(k)})) = (1-\pi_j)\sum_{n=1}^{N} p(z_n=j|x_n,\hat{\theta}^{(k)})$$

This leaves us with the simple maximum-likelihood estimates

$$\pi_j = \frac{1}{N}\sum_{n=1}^{N} p(z_n=j|x_n,\hat{\theta}^{(k)})$$

for $j = 1, 2$.

5.4. **EM Update Rules.** Using the results from the previous section, we can update our parameters using the rules

$$p(z_n = j|x_n, \hat{\theta}^{(k)}) = \frac{\hat{\pi}_j^{(k)} \hat{\alpha}_j^{(k)x_n} (1 - \hat{\alpha}_j^{(k)})^{(M-x_n)}}{\sum_{i=1}^{2} \hat{\pi}_i^{(k)} \hat{\alpha}_i^{(k)x_n} (1 - \hat{\alpha}_i^{(k)})^{(M-x_n)}}$$

$$\hat{\alpha}_j^{(k+1)} = \frac{\sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)}) x_n}{M \sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)})}$$

$$\hat{\pi}_j^{(k+1)} = \frac{1}{N} \sum_{n=1}^{N} p(z_n = j|x_n, \hat{\theta}^{(k)}),$$

for $j = 1, 2$. This will guarantee our parameters converging to a local maximum, but not necessarily the global optimum. We can improve performance by running EM from several different initial parameter estimates $\hat{\theta}^{(0)}$. The analysis presented here can easily extend to more than two coins by defining extra probabilities on the coins being chosen and changing the summation indices appropriately.

References

[1] Do, C. B., and Batzoglou, S. (2008). What is the expectation maximization algorithm? Nature Biotechnology, 26(8), 897–899. doi: 10.1038/nbt1406