Expectation-Maximization (EM) algorithm

Dr. Kuppusamy .P
Associate Professor / SCOPE

Expectation-Maximization (EM) algorithm

- **Estimation:** Estimate the expectation from some random data
- **Maximization:** Estimated value should be maximized to find the best result.
- EM algorithm is used to find the latent (not directly observed i.e., unobserved) or missing or unknown variables in the dataset.
- Observed variables are measured whereas unobserved (latent/hidden) variables are inferred from observed variables.
- E.g., Intelligence , happiness, depression, personality – these variables cannot measure directly.
- EM is used to find the **local maximum likelihood** or maximum a posteriori (MAP) parameters for latent variables in a statistical model. It is used to predict these values that is missing or incomplete.

- **When EM can be used?**
- If dataset is incomplete (i.e., groups or labels not given), but model need to predict its group or label.

Expectation-Maximization (EM) algorithm

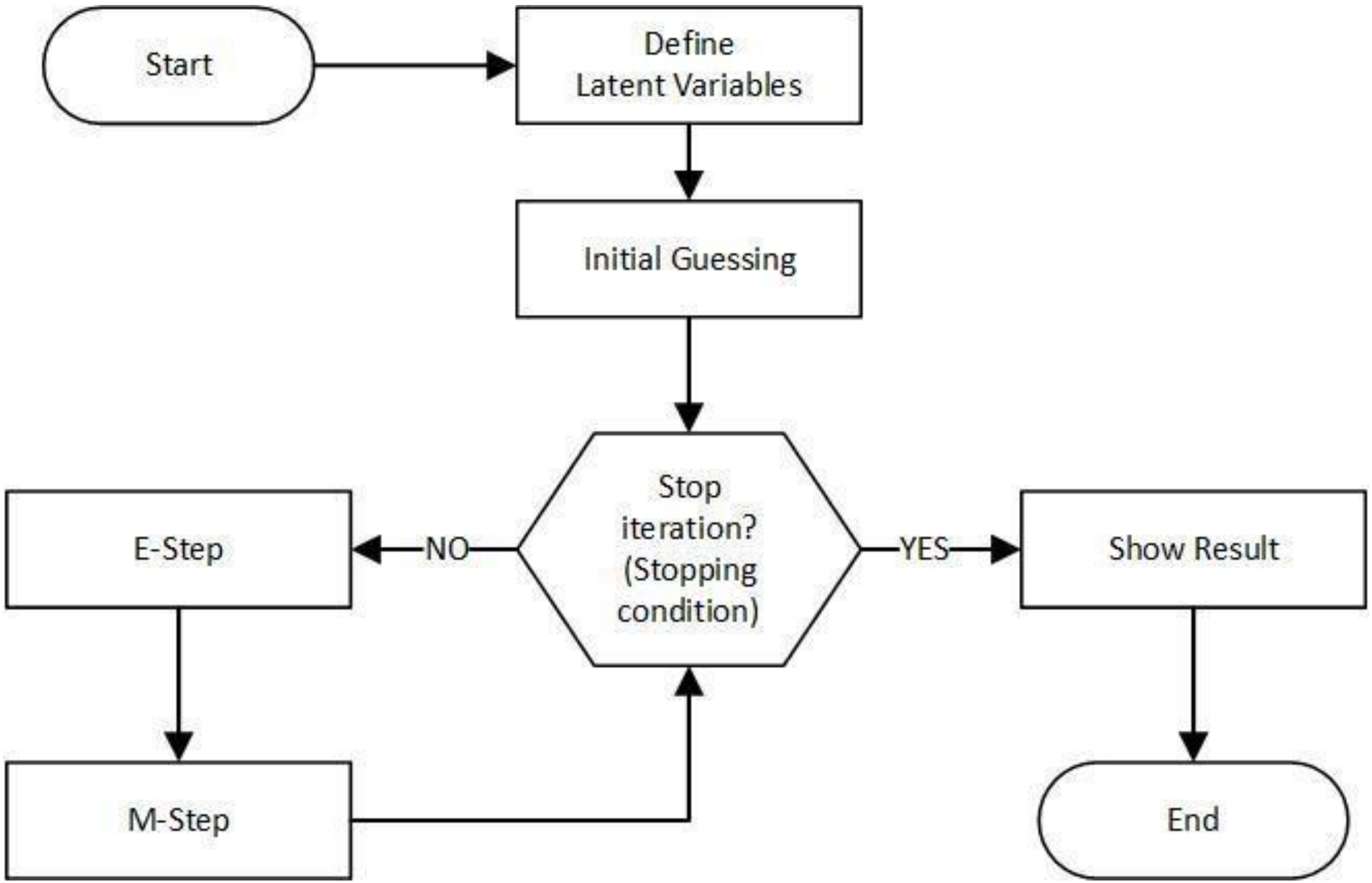
1. Initially, a set of random initial values of the parameters are considered. Assume, set of incomplete observed data (generated from normal or uniform or exponential distribution, etc) is given to the system.
2. **Expectation step (E – step):** Estimate (guess) the values of the missing or incomplete data using the previous observed data of the dataset. Estimated value is used to update the variable values.
3. **Maximization step (M – step):** Update the parameter values using Complete (Estimated) data that is generated in the (E) step. It is used to update value of hypothesis.
4. Check the values are converging or not.

   If not, Repeat step 2 and step 3 until convergence.

EM algorithm can be used:

- To fill the missing data in a sample.
- As basis of unsupervised learning of clusters.
- To Estimate the parameters of Hidden Markov Model (HMM).
- To Discover the values of latent variables.

Expectation-Maximization (EM) algorithm



Two Coins Tosses distribution Example - EM algorithm

- Let consider two Coins A and B; where both have a different head-up probability. Randomly choose a coin 5 times either coin A or B. Then, each coin selection is followed by tossing it 10 times.
- Therefore, we get the following outcomes:

Set 1: H T T T H H T H T H (5H 5T)

Set 2: H H H H T H H H H H (9H 1T)

Set 3: H T H H H H H T H H (8H 2T)

Set 4: H T H T T T H H T T (4H 6T)

Set 5: T H H H T H H H T H (7H 3T)

- The probability of coin will land with head-up for each of these coins denoted as θ.
- Estimate θ for each coin?
- Find the coin identity i,.e., what is the probability of each coin in above sets?

Two Coins Tosses distribution Example - EM algorithm

Initially, define what variables are required that are not observed in the data.

Goal: Estimate the probability of getting heads-up for each coin.

- However, it cannot be calculated directly due to not aware of the identity of the coin used in each set.
- Therefore, find which coin is used in each set i.e., coin identity is latent variable.
- As of now, we know that two coins A and B used in the 5 sets of outcomes.

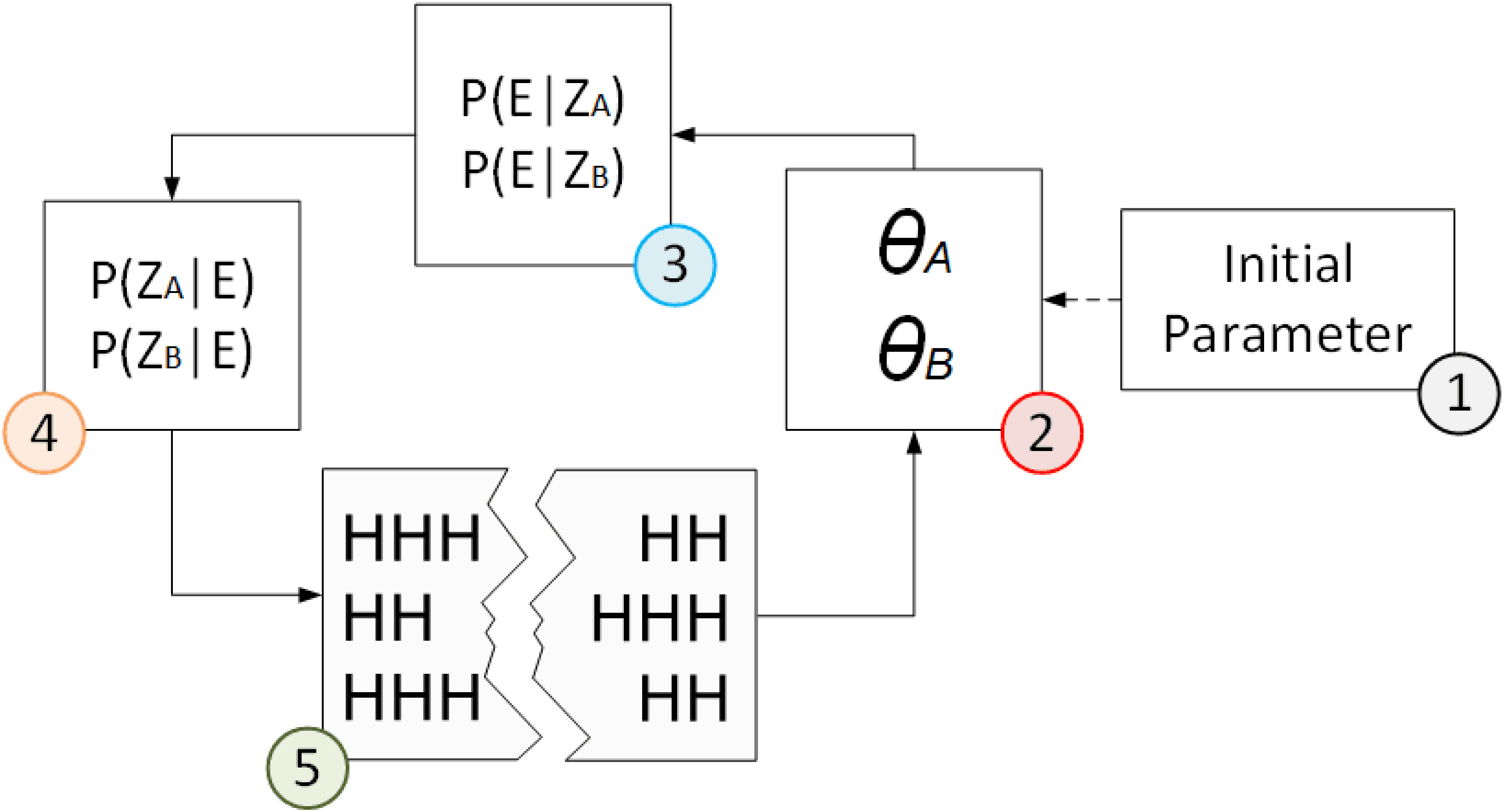
Solution:

1.Initial guess:

- Initially, model does not aware the coin identity.
- The probability of getting head-up for each of these coins denoted as θ. Therefore, guess the initial θ parameter for each coin i.e., $\theta_A\ and \theta_B$ (Both are unknown initially).
- Now, choose **randomly** between range 0 to 1 for each θ. E.g., θ_A is 0.5, 0.6, 0.7. and θ_B = 0.2, 0.4, 0.5.
- No relationship between parameters guessing both $\theta_A\ and \theta_B$. i.e., **no** need of sum of $\theta_A\ and \theta_B$ must be = 1, Because this probability represents the individual value of getting heads-up on each coin.
- Finally let consider **random initial values** for $\theta_A = 0.6\ and \theta_B = 0.5$

Two Coins Tosses distribution Example - EM algorithm

a



Two Coins Tosses distribution Example - EM algorithm

2.E-step:

- Now estimate the identity of the coin used in each set based on $\theta_A = 0.6\ and \theta_B = 0.5$.
- Calculate each coin's probability to get the outcomes of each set $\mathrm{P(E|Z_A)}$ and $\mathrm{P(E|Z_B)}$ based on the current θ parameters.
- E.g., the probability of coin A to get 8H 2T in 10 tosses in set 3.
- If $\theta_A = 0.6$ i.e., the probability of getting head is 0.6 (and tail 0.4) from coin A.
- **Compute** the probability of **coin A** will give **8H 2T** in **10 tosses (a set) i.e., $\mathbf{P(E|Z_A)}$**
- Similarly, **Compute** the probability of **coin B** giving E = 8H 2T, i.e., $\mathrm{P(E|Z_B)}$.
- The equation for probability distribution of a binomial random variable is

$$P(\mathrm{E}|Z_x) = \frac{n!}{h!\,(n-h)!}\theta_x^h\,(1-\theta_x)^{\,n-h}$$

- $P(\mathrm{E}|Z_x)$ is the probability of coin x giving E ; n - total coin tosses in a set E.
- h - total number of heads in a set of E ; θ_x - the probability of getting head-up using coin x.
- Compute for coin A and B

$P(E_{8H2T}|Z_A) = \frac{10!}{8!2!!} * 0.6^8 * 0.4^2$=0.121 ; $P(E_{8H2T}|Z_B) = \frac{10!}{8!2!!} * 0.5^8 * 0.5^2$=0.044

Two Coins Tosses distribution Example - EM algorithm

- Compare the probabilities both of coin A and coin B given E i.e., $P(Z_A|E)$ and $P(Z_B|E)$.
- Compute the ratio of probability using Bayes' theory and total probability:

$$P(Z_x|E) = \frac{\text{probability of coins x giving E}}{\text{Total probability of coins x and y in giving E}}$$

$$P(Z_x|E) = \frac{P(\mathrm{E}|Z_x) * P(Z_x)}{P(\mathrm{E}|Z_x) * P(Z_x) + P(\mathrm{E}|Z_y) * P(Z_y)}$$

$P(Z_x|E)$ - probability of coin x given E (compared to coin y).

$P(Z_x)$ - probability of choosing coin x ; $P(Z_y)$ - probability of choosing coin y.

- Among the 2 coins A and B, the probability of choosing one of them is 50:50. Then $P(Z_A) = P(Z_B) = 0.5$.

Two Coins Tosses distribution Example - EM algorithm

For Coin A selection,

$$P(Z_A|E) = \frac{P(E|Z_A)\cancel{P(Z_A)}}{P(E|Z_A)\cancel{P(Z_A)} + P(E|Z_B)\cancel{P(Z_B)}}$$

$$= \frac{P(E|Z_A)}{P(E|Z_A) + P(E|Z_B)}$$

$$P(Z_A|E_{8H2T}) = \frac{\cancel{\frac{10!}{8!\,2!}}\ 0.6^8\ 0.4^2}{\cancel{\frac{10!}{8!\,2!}}\ 0.6^8\ 0.4^2 + \cancel{\frac{10!}{8!\,2!}}\ 0.5^8\ 0.5^2}$$

$$= \frac{0.6^8\ 0.4^2}{0.6^8\ 0.4^2 + 0.5^8\ 0.5^2}$$

$$= 0.73$$

For Coin B selection

$$P(Z_B|E_{8H2T}) = \frac{\cancel{\frac{10!}{8!\,2!}}\ 0.5^8\ 0.5^2}{\cancel{\frac{10!}{8!\,2!}}\ 0.6^8\ 0.4^2 + \cancel{\frac{10!}{8!\,2!}}\ 0.5^8\ 0.5^2}$$

$$= \frac{0.5^8\ 0.5^2}{0.6^8\ 0.4^2 + 0.5^8\ 0.5^2}$$

$$= 0.27$$

Two Coins Tosses distribution Example - EM algorithm

Estimates the number of heads for each coin

- The ratio of coins A and B in giving each E is calculated from set 1 to 5.

Coin Tosses	E	Coin A Probability	Coin B Probability
HTTTHHTHTH	5H 5T	0.45	0.55
HHHHTHHHHH	9H 1T	0.8	0.2
HTHHHHHTHH	8H 2T	0.73	0.27
HTHTTTHHTT	4H 6T	0.35	0.65
THHHTHHHTH	7H 3T	0.65	0.35

- Now, estimate the **total number of H** for each coin. It is calculated based on the coin ratio above.
- To calculate "total heads and tails" for coin x, it is similar to the "complete data".
- multiply the ratio of each coin to the number of heads in each E.

Coin Tosses	E	Estimated H for Coin A	Estimated H for Coin B
HTTTHHTHTH	5H 5T	5 * 0.45 = 2.25	5 * 0.55 = 2.75
HHHHTHHHHH	9H 1T	9 * 0.80 = 7.2	9 * 0.20 = 1.8
HTHHHHHTHH	8H 2T	8 * 0.73 = 5.84	8 * 0.27 = 2.16
HTHTTTHHTT	4H 6T	4 * 0.35 = 1.4	4 * 0.65 = 2.6
THHHTHHHTH	7H 3T	7 * 0.65 = 4.55	7 * 0.35 = 2.45

No necessary to calculate the tails for each coin

Two Coins Tosses distribution Example - EM algorithm

3.M-Step

- The results of E-step can be used to improve the θ parameter. So, use the **Maximum Likelihood Estimation** (MLE) equation that is similar to the “completed data”.
- For sum (total estimated tosses)=1, we need the heads and tails for each coin. But it can be done as follows:

Multiply the coin ratio with 10 tosses:

$$\theta'_A = \frac{2.25 + 7.2 + 5.84 + 1.4 + 4.55}{10 * (0.45 + 0.8 + 0.73 + 0.35 + 0.65)} = 0.713$$

$$\theta'_B = \frac{2.75 + 1.8 + 2.16 + 2.6 + 2.45}{10 * (0.55 + 0.2 + 0.27 + 0.65 + 0.35)} = 0.581$$

- Finally, the both parameters of $\theta_A\ and \theta_B$ for the first iteration have been improved.
- For the next iteration, the E-Step will use this new parameter value, and re-improved at the next M-step.
- This iteration will always repeat the E-step and M-step, until it reaches any stop condition.

Two Coins Tosses distribution Example - EM algorithm

4.Stopping condition and the final result

- The iteration of the E-Step and M-Step, will be repeated until they meet the stopping condition.
- Commonly, the EM algorithm has two options of stopping condition:
- **Maximum iteration:** EM Algorithm will stop if a certain number of iterations has been reached.

  E.g., maximum iteration =10, then EM Algorithm will not be more than 10 iterations.
- Or,
- **Convergence threshold:** M-step gives no significant parameter improvement compared to the improvement in the previous iteration. The changes are very small below our threshold.

Two Coins Tosses distribution Example - EM algorithm

Final result of EM algorithm:

The parameter improvement in each iteration

Iteration	θ_A	θ_B	Differences
0	0.6	0.5	0.781
1	0.713	0.581	0.139
2	0.745	0.569	0.0342
3	0.768	0.55	0.0298
4	0.783	0.535	0.0212
5	0.791	0.526	0.012
6	0.795	0.522	0.0057
7	0.796	0.521	0.0014
8	0.796	0.52	0.001
9	0.796	0.52	0

To calculate the differences or improvements in each iteration, use Euclidean Distance.

E.g.,

$$d(\theta, \theta') = \sqrt{(\theta_A - \theta'_A)^2 + (\theta_B - \theta'_B)^2}$$

$$d(iter1, iter2) = \sqrt{(0.713 - 0.745)^2 + (0.581 - 0.569}$$

$$= 0.0342$$

Two Coins Tosses with Completed data

- Therefore, we get the following outcomes:

Set 1: Coin B: H T T T H H T H T H (5H 5T)

Set 2: Coin A: H H H H T H H H H H (9H 1T)

Set 3: Coin A: H T H H H H H T H H (8H 2T)

Set 4: Coin B: H T H T T T H H T T (4H 6T)

Set 5: Coin A: T H H H T H H H T H (7H 3T)

- The probability of coin will land with head-up for each of these coins $\theta_A\ and \theta_B$.
- Simple equation for MLE $\theta_x = \frac{No.of\ heads-up\ using\ coin\ X}{Total\ no.of\ tosses\ with\ coin\ X}$

$\theta_A = \frac{24}{30} = 0.8;$ $\theta_B = \frac{9}{20} = 0.45;$

Expectation-Maximization (EM) algorithm

Advantages

- It is always guaranteed that likelihood will increase with each iteration.
- The E-step and M-step are easy for implementation.
- Solutions to the M-steps often exist in the closed form.

Disadvantages

- Slow convergence.
- It makes convergence to the local optima only.
- It requires both the probabilities, forward and backward (numerical optimization requires only forward probability).

References

1. Tom M. Mitchell, Machine Learning, McGraw Hill , 2017.

2. EthemAlpaydin, Introduction to Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2017.

3. Wikipedia

4. https://www.indowhiz.com/articles/en/the-simple-concept-of-expectation-maximization-em-algorithm/