Expectation Maximization Algorithm: A Review

Md. Masum Bhuiyan, Samsun Nahar Khandakar, Nadia Afrin Ritu and Md. Imdadul Islam
Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh
Correspondence E-mail: samsunnahar@juniv.edu

Abstract— **The Expectation-Maximization (EM) algorithm is used to estimate the maximum likelihood parameters of a statistical model iteratively. The EM algorithm is applicable when the given data set is incomplete, i.e., the data has two parts: the observed variable (known) and the latent variable. Latent (hidden) variables are not directly observed but inferred from observed variables using some mathematical model. Before using the EM algorithm, we need to understand observed variables, latent variables, likelihood, and Jensen's inequalities. This review presents the basic theory of the mathematical models used to implement the EM algorithm, the algorithm's operational steps with explanations, and numerical examples to provide a clear understanding.**

***Index Terms*—Likelihood function, Jensen's inequalities, latent variables, biasing probability, and E-step.**

I. INTRODUCTION

The Expectation–Maximization (EM) algorithm is a widely used iterative method for estimating the parameters of statistical models when the observed data are incomplete or when the model contains latent (hidden) variables. In many real-world problems, not all variables of interest are directly observable; instead, some variables remain hidden and must be inferred from the available observations. Under such conditions, direct maximization of the likelihood function becomes difficult, and the EM algorithm provides an effective framework by iteratively alternating between two steps: the Expectation step (E-step), which estimates the hidden variables or their expected sufficient statistics using the current parameter values, and the Maximization step (M-step), which updates the model parameters by maximizing the expected log-likelihood. Because of this ability to handle incomplete-data estimation problems, EM has become a fundamental tool in statistics, machine learning, signal processing, wireless communications, reliability analysis, image processing, and Bayesian inference.

The formal foundation of the EM algorithm was established by Dempster, Laird, and Rubin [16], who introduced it as a general maximum-likelihood estimation framework for incomplete data and proved its monotonic likelihood-improvement property. Since then, EM has been used in a broad range of applications, including process identification [1], large-scale parameter estimation [2, 20], target localization in wireless sensor networks [3], audio source separation [4], Gaussian mixture modeling for large datasets [5], reliability analysis with censored data [6, 21], hyperspectral image classification [7], radar target clustering [8], wireless channel estimation [9], Wi-Fi power-saving optimization [10], logistic regression with aggregated responses [23], and genome-wide association studies [24]. These works demonstrate both the theoretical importance and practical flexibility of EM across diverse data-driven problems.

Although the EM algorithm is widely used, many existing studies emphasize either specialized applications or advanced theoretical developments, which can make it difficult for new readers to build an intuitive understanding of how the algorithm is derived and implemented in practice. In particular, a clear connection between Jensen's inequality, likelihood and log-likelihood formulation, lower-bound construction, and the operational meaning of the E-step and M-step is often missing in a single self-contained exposition. This gap motivates the present review.

The main contribution of this paper is therefore to provide a tutorial-oriented and self-contained review of the EM algorithm that bridges mathematical foundation and practical implementation. Specifically, the paper:

- Reviews the fundamental concepts required to understand EM, including observed variables, latent variables, likelihood, log-likelihood, and Jensen's inequality;
- Derives the lower bound of the log-likelihood and explains its role in the EM procedure;
- Presents the E-step and M-step in a step-by-step manner with intuitive explanations rather than only abstract pseudocode;
- Illustrates the operation of the algorithm using explicit numerical examples, including multinomial and coin-tossing examples; and
- Consolidates representative EM applications and recent developments from the literature to provide readers with both conceptual understanding and practical context.

By integrating theoretical background, derivation, algorithmic interpretation, and worked examples in a single discussion, this paper aims to serve as an accessible reference for students, researchers, and practitioners interested in understanding the EM algorithm from first principles and applying it to incomplete-data estimation problems.

II. RELATED WORK

A. Related studies

The Expectation–Maximization (EM) algorithm has been extensively studied in both theoretical and application-oriented contexts. A review of EM in data-driven process identification was presented **by** Sammaknejad et al. **[**1],

where the authors discussed the usefulness of EM in parameter estimation under incomplete information. Since the computational efficiency of standard EM deteriorates for large datasets, Verma et al. [2] proposed several acceleration strategies, including grid-based sampling, adaptive estimation based on previous errors, and parallel implementation. In the context of wireless sensor networks, Qian et al. [3] applied a variational EM framework for multi-target localization using superimposed received signal strength, where the unknown target positions were modeled as latent variables and iteratively updated.

In signal-processing applications, Sawada et al**.** [4] investigated the use of EM and multiplicative-update algorithms for optimizing Full-Rank Spatial Covariance Analysis in audio source separation and showed that EM provides better separation performance at convergence. To address the large-scale data limitation of conventional EM, Przyborowski and Ślęzak [5] proposed an approximate EM approach for Gaussian Mixture Models by partitioning large datasets and merging local parameter estimates. A stochastic variant of EM was introduced by Wu et al. [6] for reliability analysis of progressively censored mixture Weibull data, where the method was combined with stochastic estimation to handle censored observations. In image analysis, Zhuang et al. [7] developed an improved EM-based Gaussian Mixture Model for hyperspectral image classification and demonstrated promising performance even with a limited number of training samples.

The application of EM has also been extended to radar and wireless communication systems. Yan et al. [8] proposed an EM-based algorithm for unsupervised clustering of measurements from a radar sensor network, where the latent-variable structure was used to associate received measurements with multiple moving targets. In wireless communication, Wang et al. [9] applied EM to uplink channel estimation under fading conditions in high-speed railway systems and reported improved estimation accuracy. In Wi-Fi networking, Ron and Lee [10] used EM to estimate traffic-distribution parameters for enhancing the Notice of Absence power-saving mode in Wi-Fi Direct, showing improved trade-offs between energy consumption and transmission delay.

From a theoretical perspective, the derivation of EM is closely related to Jensen's inequality and likelihood maximization. Abramovich et al. [11] presented sharpened forms of Jensen's inequality using majorization theory, while Nisar et al. [12] proposed improved bounds for integral Jensen's inequality through higher-order differentiable convex functions. Although these studies are not directly focused on EM, they are relevant to the mathematical foundation of the lower-bound formulation used in the algorithm. Related likelihood-based estimation studies also appear in the broader statistical literature. For example, da Silva et al. [13] investigated bias reduction in modified maximum-likelihood estimation for a three-parameter Weibull distribution, and Van Lierop et al**.** [14] reviewed log-likelihood ratio cost functions in forensic science. Similarly, Gemeay and Abd El-Bar [15] studied bias reduction in maximum-likelihood estimation for the unit odd Lindley half-logistic model. These works are relevant because EM is fundamentally rooted in maximum-likelihood estimation for incomplete-data problems.

The formal foundation of the EM algorithm was established by Dempster, Laird, and Rubin [16], who introduced EM as a general framework for maximum-likelihood estimation from incomplete data and proved its monotonic convergence property. Building on this foundation, Hua-Yan et al. [17] combined EM with K-means clustering to accelerate missing-data imputation. A broader textbook-style treatment of EM and related statistical models was provided by Watanabe and Yamaguchi [18], which has been useful for understanding the algorithm from a pedagogical perspective. Neal and Hinton [19] later gave a variational interpretation of EM by showing that the E-step and M-step can be viewed as coordinate-ascent updates on a lower-bound objective function, an interpretation that strongly influenced later generalized EM formulations.

Further efforts to improve the efficiency and applicability of EM can be found in large-scale data and reliability applications. Thiesson et al. [20] investigated methods for accelerating EM in large databases, while Silva et al. [21] applied EM to reliability estimation with censored industrial data. More recently, Caprio and Johansen [22] established fast convergence guarantees for EM under logarithmic Sobolev conditions, providing new theoretical insight into its convergence behavior. The adaptability of EM to different statistical models has also been demonstrated by Xu [23], who proposed an EM algorithm for logistic regression with individual-level predictors and aggregate-level responses. In high-dimensional statistical inference, Zhang et al. [24] introduced an improved EM-based Bayesian algorithm for genome-wide association studies. Finally, Ritz et al. [25] explored evolutionary maximum-likelihood estimation for discrete latent-variable models, offering an alternative optimization strategy when the M-step becomes analytically difficult.

B. Limitations of Existing Works

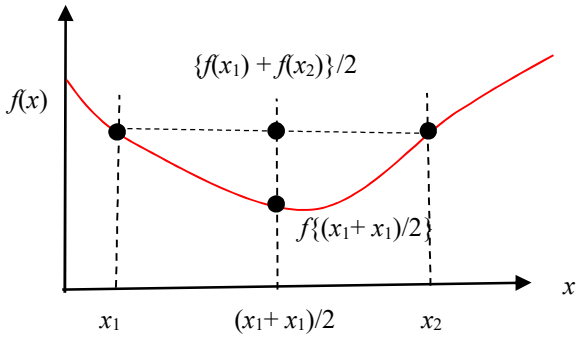
Overall, the existing literature demonstrates that the EM algorithm is both theoretically important and practically versatile, with applications spanning process identification, signal processing, image classification, wireless communications, reliability analysis, and high-dimensional statistical inference. However, much of the available work is either highly application-specific or mathematically advanced, making it difficult for new readers to connect the theoretical derivation of EM with its practical implementation. Therefore, a tutorial-style review that integrates the mathematical basis of EM, its lower-bound interpretation, step-by-step algorithmic procedure, and representative applications remains valuable.

.

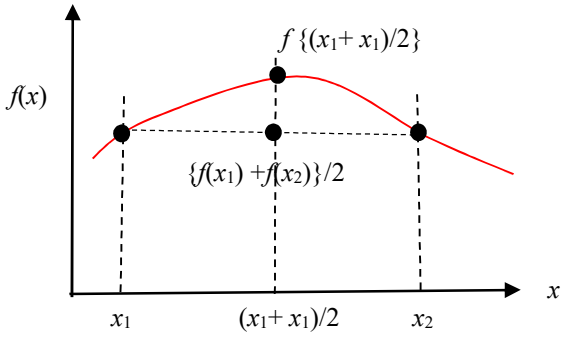
III. Basic Theory

A. Jensen's inequalities

The 'Jensen's inequalities' is mathematical model found in [11]-[12], which is used in EM algorithm. It provides two inequalities regarding concave and convex curves. For a convex function f we can write, $f\left(\frac{x_1+x_2}{2}\right) \leq \frac{f(x_1)+f(x_2)}{2}$ like fig.1 (a); again, for a concave function, $f\left(\frac{x_1+x_2}{2}\right) \geq \frac{f(x_1)+f(x_2)}{2}$ shown in fig.1 (b).



(a) Convex function



(b) Concave function

Fig. 1. Geometry of convex and concave curves

For convex function, $f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2)$; where $\alpha_1 + \alpha_2 = 1$, $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$.

In generalize form, $f(\sum \alpha_i x_i) \leq \sum \alpha_i f(x_i)$, where $\sum \alpha_i = 1$and $\alpha_i \geq 0$.

Similarly, for concave function, $f(\sum \alpha_i x_i) \geq \sum \alpha_i f(x_i)$, where $\sum \alpha_i = 1$and $\alpha_i \geq 0$.

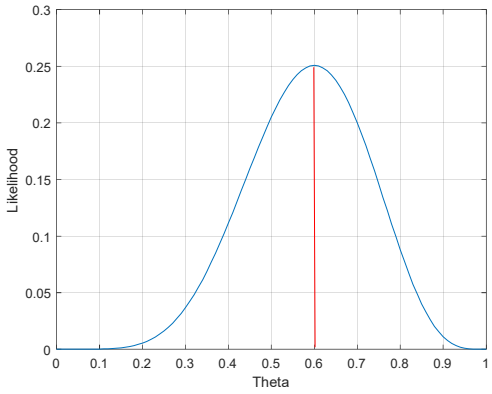
Since log is a concave function, therefore, $log(\sum \alpha_i x_i) \geq \sum \alpha_i \, log(x_i)$, where $\sum \alpha_i = 1$and $\alpha_i \geq 0$.

B. Basic of the likelihood function

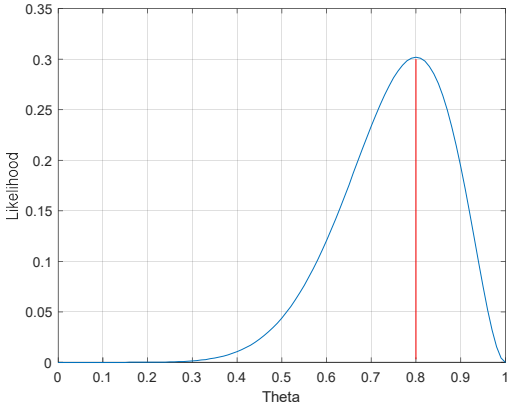
Consider a biased coin is tossed n times and we expect to get heads k times. If the biasing of the coin to the head i.e., probability of appearing head is θ (unknown parameter), then the likelihood of getting k heads from the n tosses is found from Binomial pdf,

$$L(\theta|n,k) = \binom{n}{k} \times \theta^k \times (1-\theta)^{n-k} \quad (1)$$

Taking, $n = 10$, $k = 6$, we get the pdf like fig.2 (a) and for $k = 8$ pdf is shown in fig. 2(b).



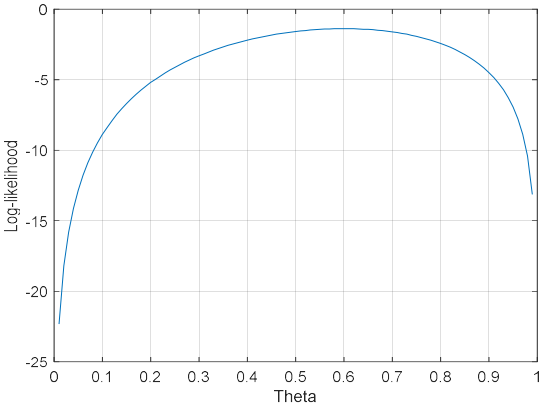
(a) K=6



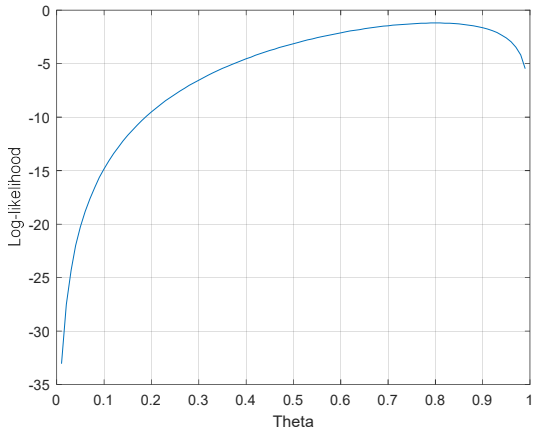
(b) K= 8

Fig. 2. The likelihood of Binomial pdf

The likelihood is found to be maximum 0.25 at $\theta = 0.6$ for $k = 6$, and at $\theta = 0.8$ for $k = 8$, the value of the likelihood is 0.3. The data samples at $k = 6$ and $k = 8$ are made most likely at $\theta = 0.6$ and $\theta = 0.8$, respectively. The logarithm of the likelihood function is called the log-likelihood function, and the plot of the log-likelihood function of the above cases is shown in fig.3.



(a) K = 6



(b) K= 8

Fig. 3. The log-likelihood of Binomial pdf

For the independent random variables, the joint pdf is the product of the individual random variables expressed as,

$$f(x;\theta) = \prod_{i=1}^{n} f_i(x_i;\theta) \quad (2)$$

In this case, it is convenient to take the natural logarithm of the above likelihood function to convert the multiplication to addition.

C. Log-likelihood in EM

Now, consider a latent (hidden or missing or uncovered) variable z and a visible or observed variable x. Let q(z) be the distribution of latent variable z, where both z and q(z) are unknown. From the definition of marginal probability,

$p(x_i|\theta) = \sum_Z p(x_i, z_i|\theta)$, where θ is an unknown parameter that parameterizes the joint probability.

The log-likelihood [13]-[15],

$$L(q,\theta) = log\, p\,(x_i|\theta) = log \sum_Z p(x_i, z_i|\theta)$$

Taking, $\alpha_i = q(z_i)$,

$$= log \sum_Z q(z_i) \frac{p(x_i,z_i|\theta)}{q(z_i)} \geq \sum_Z q(z_i)\, log \frac{p(x_i,z_i|\theta)}{q(z_i)} = L'(q,\theta), \quad (3)$$

where $L'(q,\theta)$ is the lower bound of log-likelihood $L(q,\theta)$. Again, we can express the lower bound $L'(q, \theta)$ in a different way like,

$$L'(q,\theta) = \sum_{z_i} q(z_i)\, log \frac{p(x_i,z_i|\theta)}{q(z_i)}$$

$$= \sum_{z_i} q(z_i)\, log \frac{p(z_i|x_i,\theta)p(x_i|\theta)}{q(z_i)}$$

$$= \sum_{z_i} q(z_i) \left\{ log \frac{p(z_i|x_i,\theta)}{q(z_i)} + log\, p\,(x_i|\theta) \right\}$$

$$= \sum_{z_i} q(z_i)\, log \frac{p(z_i|x_i,\theta)}{q(z_i)} + \sum_{z_i} q(z_i)\, log\, p\,(x_i|\theta)$$

$$= -\sum_{z_i} q(z_i)\, log \frac{q(z_i)}{p(z_i|x_i,\theta)} + log\, p\,(x_i|\theta) \sum_{z_i} q(z_i)$$

$$= -\sum_{z_i} q(z_i)\, log \frac{q(z_i)}{p(z_i|x_i,\theta)} + log\, p\,(x_i|\theta)$$

$$= -\sum_{z_i} q(z_i)\, log \frac{q(z_i)}{p(z_i|x_i,\theta)} + L(q,\theta)$$

$$\Rightarrow L(q,\theta) = \sum_{z_i} q(z_i)\, log \frac{q(z_i)}{p(z_i|x_i,\theta)} + L'(q,\theta)$$

For an estimated value model parameter θ^t at time instant t, put $q(z_j) = p(z_j|x_i,\theta^t)$,

$$\Rightarrow L(q,\theta^t) = \sum_{z_i} q(z_i)\, log \frac{p(z_j|x_i,\theta^t)}{p(z_j|x_i,\theta^t)} + L'(q,\theta^t) = 0 +$$

$$L'(q,\theta^t) = L'(q,\theta^t) \quad (4)$$

Therefore, log likelihood $L(q,\theta)$ and its lower bound $L'(q,\theta)$ touches at a known estimated value θ^t as shown in fig.4.

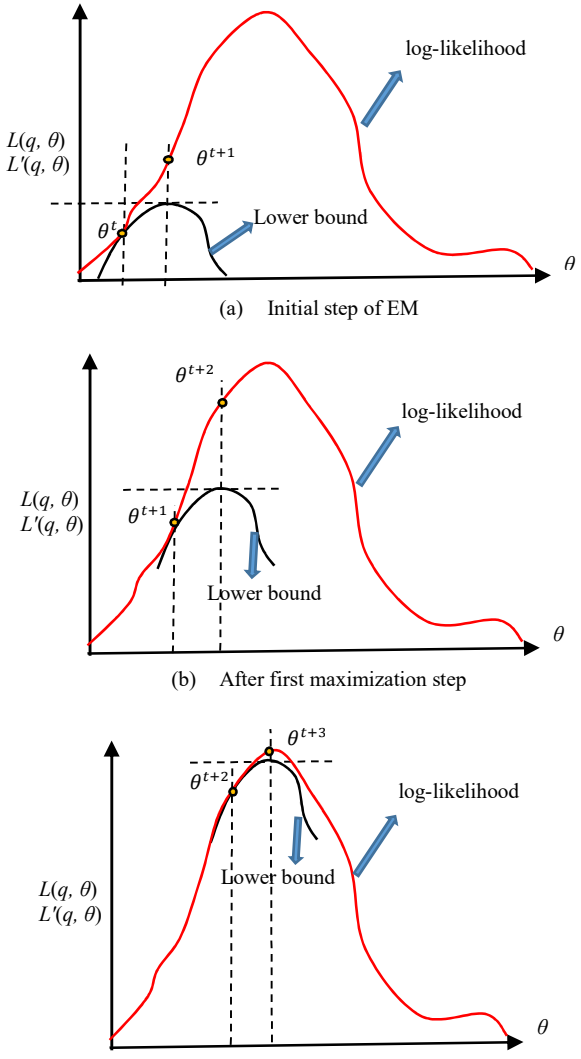


Fig. 4. Graphical representation of likelihood and its lower bound

D. The lower bound of log-likelihood and its expected value

Putting $q(z_i) = p(z_i|x_i,\theta^t)$ in the inequality of (1), the lower bound of log-likelihood is found as,

$$log \sum_{Z} p(z_i|x_i,\theta^t)\frac{p(x_i,z_i|\theta)}{p(z_i|x_i,\theta^t)} \geq \sum_{Z} p(z_i|x_i,\theta^t)\, log\frac{p(x_i,z_i|\theta)}{p(z_i|x_i,\theta^t)}$$

$$= E_{z_i|x_i,\theta^t}\left[log\frac{p(x_i,z_i|\theta)}{p(z_i|x_i,\theta^t)}\right] = U(\theta,\theta^t) = (L'(q,\theta) \quad (5)$$

The EM algorithm consist of two steps: (a) expected step and (b) maximization steps. The lower bound $L'(q,\theta)$ is related to expected value of log-likelihood function. The expected step of the algorithm is nothing but calculation of lower bound$L'(q,\theta)$. The maximization step is related to finding the model parameter θ^t at time instant t from the maxima of $L'(q,\theta)$.

Lemma-1:

If the visible variables are $x_1, x_2, x_3, \ldots, x_N$ and latent variables are $z_1, z_2, z_3, \ldots, z_M$ then considering summation of observed variable x_i on log-likelihood derive the lower bound,

$\sum_{i=1}^{N} E[log\, p\,(x_i,z_i|\theta)] + \sum_{i=1}^{N} H(q)$, and

$L'(q,\theta) = E[log\, p\,(x_i,z_i|\theta)] + H(q_i)$; where the entropy, $H(q) = \sum_{z_i} q(z_j)\, log\frac{1}{q(z_j)}$.

Ans.

The log-likelihood,

$$l(q,\theta) = \sum_{i=1}^{N} log \sum_{z_i} p(x_i,z_i|\theta)$$

$$= \sum_{i=1}^{N} log \sum_{z_i} q(z_i)\, log\frac{p(x_i,z_i|\theta)}{q(z_i)} \geq \sum_{i=1}^{N}\sum_{z_i} q(z_i)\, log\frac{p(x_i,z_i|\theta)}{q(z_i)} = Q(q,\theta)$$

where $Q(q,\theta)$is the lower bound of $l(q,\theta)$.

Now,

$$Q(q,\theta) = \sum_{i=1}^{N}\sum_{z_i} q(z_i)\, log\frac{p(x_i,z_i|\theta)}{q(z_i)}$$

$$= \sum_{i=1}^{N}\sum_{z_i} q(z_i)\left\{log\, p\,(x_i,z_i|\theta) + log\frac{1}{q(z_i)}\right\}$$

$$\Rightarrow Q(q,\theta) = \sum_{i=1}^{N}\sum_{z_i} q(z_i)\, log\, p\,(x_i,z_i|\theta) + \sum_{i=1}^{N}\sum_{z_i} q(z_i)\, log\frac{1}{q(z_j)}$$

$\Rightarrow Q(q,\theta) = \sum_{i=1}^{N}\sum_{z_i} q(z_i)\, log\, p\,(x_i,z_i|\theta) + \sum_{i=1}^{N} H(q)$;
Where the entropy, $H(q) = \sum_{z_i} q(z_j)\, log\frac{1}{q(z_j)}$

$$\Rightarrow Q(q,\theta) = \sum_{i=1}^{N} E[log\, p\,(x_i,z_i|\theta)] + \sum_{i=1}^{N} H(q), \quad (1)$$

where the expectation is, $E[log\, p\,(x_i,z_i|\theta)] = \sum_{j=1}^{M} q(z_i)\, log\, p\,(x_i,z_i|\theta)$and the second term of $Q(q,\theta)$ is independent of θ. Elimination the summation over i we get the lower bound as,

$$L'(q,\theta) = E[log\, p\,(x_i,z_i|\theta)] + H(q_i) \quad (2)$$

, which is a function of θ and q. The first term is independent of q and the second term is independent of θ.

Lemma-2:

Derive the relation, $L'(q,\theta) = E_Z[Log - likelihood] + H$; where H is the entropy of $q(z_i)$.

Ans.

The lower bound of likelihood function,

$$L'(q,\theta) = \sum_{Z} q(z_i)\, log\frac{p(x_i,z_i|\theta)}{q(z_i)}$$

$$= \sum_{Z}\left\{q(z_i)\, log\, p(x_i,z_i|\theta) + q(z_i)\, log\frac{1}{q(z_i)}\right\}$$

$$= \sum_{Z} q(z_i)\, log\, p(x_i,z_i|\theta) + \sum_{Z} q(z_i)\, log\frac{1}{q(z_i)};$$

In above expression the first part is the expected value of log-likelihood, and the second part is the entropy H of $q(z)$. Therefore,

$$L'(q,\theta) = E_Z[Log - likelihood] + H;$$

IV. EM ALGORITHM

The EM algorithm was first introduced in [16]. This section provides the algorithm in several steps not like the pseudo code form, where the few steps may contain repeating works but it is done to make the readers to perceive it conveniently. The EM algorithm operates in two steps: expectation Step or E-step and Maximization Step or M-step. The details of both steps is found in [17-21].

Initial parameters: select initial model parameter θ^t

Step-1

Derive the expression of the lower bound $L'(q,\theta)$ using the model parameter θ^t known as the expectation step. The initial model parameter θ^t is the touching point of $L'(q,\theta)$ and $L(q,\theta)$.

Step-2

Determine the maxima of $L'(q,\theta)$ using $\frac{\delta}{\theta}L'(q,\theta) = 0$, the parameter θ corresponding to the maxima is θ^{t+1} is the updated version of θ^t, which is known as the maximization step shown in fig.4(a).

Step-3

Reconstruct $L'(q,\theta)$ as the expectation step, considering θ^{t+1} as the touching point as shown in fig. 4(b). The new value of $\theta = \theta^{t+1}$ is now closer to the peak of $L(q,\theta)$.

Step-4

Determine θ^{t+2} corresponding to maxima of $L'(q,\theta)$ then reconstruct $L'(q,\theta)$ (as the expectation step) such that log-likelihood and lower bound touches at θ^{t+2} as shown in fig.4(b). The new value of $\theta = \theta^{t+2}$ is now closer to the peak of $L(q,\theta)$ compare to previous step.

Step-5

Reconstruct $L'(q,\theta)$ as the expectation step, considering θ^{t+2} as the touching point as shown in fig. 4(c). The new value of $\theta = \theta^{t+2}$ is now closer to the peak of $L(q,\theta)$.

Step-6

Determine θ^{t+3} corresponding to maxima of $L'(q,\theta)$ as shown in fig.4(c).

Step-7

Determine, θ^{t+4}, θ^{t+5}, θ^{t+6},… based on alternate E-step and M-step.

To determine the convergence of the algorithm, continue iteration till, $|\theta^{t+k} - \theta^{t+k+1}| \leq \varepsilon$ i.e. peaks of both $L'(q,\theta)$ and $L'(q,\theta)$ converge.

The complete flowchart of the EM is shown in fig. 5.

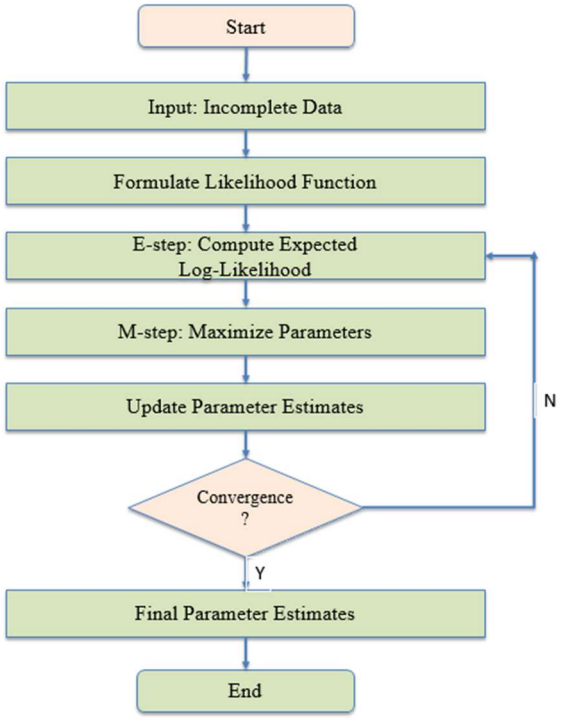


Fig. 5. Basic flowchart of EM algorithm

A. The explanation of E-step and M-step of EM algorithm from the lower bound $L'(q,\theta)$

Expectation step:

a) E-step of the EM algorithm, the lower bound $L'(\theta, q)$ is maximized w.r.t. q (the distribution of hidden or latent variable z) and θ is kept fixed. In this step, we have to select a $q = q^{\text{new}}$ that maximizes $L'(\theta, q)$. We can write, $q^{new} = \underset{q}{argmax} L'(\theta^{old}, q)$

b) Another explanation of the E-step is, at (t+1) th step, we assume the unknown variable $\theta(t)$ of the previous step is available. Compute the expected value $E[log\ p\ (x_i, z_i|\theta(t))]$ of $L'(q,\theta)$.

Maximization step:

In the M-step, the lower bound $L'(q,\theta) = E[log\ p\ (x_i, z_i|\theta)] + H(q_i)$ is maximized w.r.t. θ and q is kept fixed. We have to select θ^{new} which maximizes $L'(\theta, q)$, using $\frac{\delta}{\delta\theta} L'(q,\theta(t)) = 0$. We can write,

$$\theta^{new} = \underset{\theta}{argmax} L'(\theta, q^{new})$$

V. NUMERICAL EXAMPLE

A. Example 1

Consider n animals are distributed into four categories, following a multinomial pdf with probabilities: $p_1 = \frac{1}{2} + \frac{\theta}{4}$, $p_2 = \frac{1-\theta}{4}$, $p_3 = \frac{1-\theta}{4}$, and $p_4 = \frac{\theta}{4}$. The observed number of four categories is: x_1, x_2, x_3, and x_4; therefore, $x_1 + x_2 + x_3 + x_4 = n$. The samples of the observed variable, $\mathbf{x} = (x_1, x_2, x_3, x_4)$ has the likelihood,

$$L(\mathbf{x}:\theta) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!}\left(+\frac{\theta}{4}\right)^{x_1}\left(\frac{1-\theta}{4}\right)^{x_2}\left(\frac{1-\theta}{4}\right)^{x_3}\left(\frac{\theta}{4}\right)^{x_4}$$

The log-likelihood,

$$ln(L(\boldsymbol{x}:\theta) = l(\boldsymbol{x}:\theta)$$

$$= ln\left(\frac{n!}{x_1!\,x_2!\,x_3!\,x_4!}\right) + x_1 ln\left(\frac{1}{2} + \frac{\theta}{4}\right) + x_2 ln\left(\frac{1-\theta}{4}\right) + x_3 ln\left(\frac{1-\theta}{4}\right) + x_4 ln\left(\frac{\theta}{4}\right)$$

$$= C + x_1 ln\left(\frac{1}{2} + \frac{\theta}{4}\right) + (x_2 + x_3) ln(1-\theta) + x_4 ln(\theta)$$

Consider the samples of the complete data variable,

$\mathbf{y} = (y_{11}, y_{12}, x_2, x_3, x_4)$ i.e. x_1 is split into two parts: y_{11} and y_{12} has the probabilities:

$$p_{11} = \frac{1}{2}, p_{12} = \frac{\theta}{4}, p_2 = \frac{1-\theta}{4}, p_3 = \frac{1-\theta}{4}, \text{ and } p_4 = \frac{\theta}{4}$$

Someone can think that the first sample x_1 has few goats (y_{11}) and few rams (y_{12}), but they were visualized in the combined form of a small animal.

The log-likelihood,

$$\ln(L(\mathbf{y}:\theta) = l(\mathbf{y}:\theta)$$

$$= C + y_{12} ln(\theta) + (x_2 + x_3) ln(1-\theta) + x_4 ln(\theta)$$

Here, the probability of y_{11} is independent of θ but y_{12} depends on θ considered as the latent variable.

Now,

$$P(y_{12}|y_{11}+y_{12},\theta)=\frac{P(y_{12}\cap y_{11}+y_{12})}{P(y_{11}+y_{12})}$$
$$=\frac{P(y_{12})}{P(y_{11})+P(y_{12})}=\frac{\theta/4}{\frac{1}{2}+\theta/4}$$

, since of y_{11} and y_{12} are independent.

The conditional PDF of $y_{11}+y_{12}$ follows a Binomial distribution,

$$f(\,y_{12}|\mathbf{x},\theta)=B\left(y_{11}+y_{12},\ \frac{\theta/4}{1/2+\theta/4}\right)$$

Therefore, the mean value of the latent variable, $E[y_{12}]=(y_{11}+y_{12}).\frac{\theta/4}{1/2+\theta/4}$

E-Step:

$$Q(\theta;\theta_{old})=E[l(\theta;\mathbf{x},\mathbf{y})|\mathbf{x},\theta_{old}]$$
$$=C+E[y_{12}ln(\theta)|\mathbf{x},\theta_{old}]+(x_2+x_3)ln(1-\theta)+x_4ln(\theta)$$
$$=C+(y_{11}+y_{12}).\frac{\theta_{old}/4}{1/2+\theta_{old}/4}ln(\theta)+(x_2+x_3)ln(1-\theta)+x_4ln(\theta)$$

M-step:

$$\frac{dQ(\theta;\theta_{old})}{d\theta}=\frac{(y_{11}+y_{12})}{\theta}.\frac{\theta_{old}/4}{1/2+\theta_{old}/4}-\frac{x_2+x_3}{1-\theta}+\frac{x_4}{\theta}$$
$$=0$$

Solving,

$$\theta_{new}=\frac{x_4+(y_{11}+y_{12})\frac{\theta_{old}/4}{1/2+\theta_{old}/4}}{x_2+x_3+x_4+(y_{11}+y_{12})\frac{\theta_{old}/4}{1/2+\theta_{old}/4}}$$

Now, apply iteration on the above E-steps and M-step.

B. Example-2

Consider two biased coins A and B, where the biasing of A to the head (probability of getting a head) is $\theta_A$and that of B is θ_B. We will determine θ_A and θ_B from several experiments of coin tossing. Consider coin A is selected 3 times and B is chosen 2 times for the experiment of tossing. A coin is tossed 10 times under each experiment, and the corresponding results are shown in Table I.

TABLE I: COIN TOSSING EXPERIMENT WITH KNOWN PARAMETERS

COIN SELECTED	RESULTS	COIN A	COIN B
B	HHTTHHHTTHHHTHH		10H, 5T
A	HHTHHHTHHHHTHHHH	12H, 3T	
A	HTTTHTTTHHHTTHH	7H, 8T	
B	HHTTTTTTHTTHTTH		5H, 10T
A	HHHTTTHHHTTHTTH	8H,7T	
SUM		27H, 18T	15H, 15T

Now, $\theta_A=\frac{Number\ of\ Head}{Number\ toss}=\frac{27}{18+2}=0.6$ and $\theta_B=\frac{Number\ of\ Head}{Number\ toss}=\frac{15}{15+15}=0.5$

Now, consider the situation that, results are known, but which coin is selected is unknown to us. Now, consider under each experiment a coin is tossed 15 times and the tabular results are shown in Table II.

TABLE II: COIN TOSSING EXPERIMENT WITH UNKNOWN COIN

COIN SELECTED	RESULTS	NUMBER OF HEAD
?	HHTTHHHTTHHHTHH	10
?	HHTHHTHHHHTHHHH	12
?	HTTTHTTTHHHTTHH	7
?	HHTTTTTTHTTHTTH	5
?	HHHTTTHHHTTHTTH	8
SUM		

In this case, the bias of each coin to the head can be determined using EM algorithm. First of all, we have to select some initial random probability $\theta_A^{(0)}=0.6$ and $\theta_B^{(0)}=0.5$, which will be updated using EM algorithm consists of two steps: E-step and M-step under each iteration.

E-step

For the first experiment,

$\boldsymbol{P}(HHTTHHHTTHHHTHH|\text{Coin A})$

$=\binom{n}{k}\times\theta_A^k\times(1-\theta_A)^{n-k}$, n is the number of head and k is number of tail

$=\binom{15}{10}\times0.4^{10}\times0.6^5=0.0245$

$\boldsymbol{P}(HHTTHHHTTHHHTHH|\text{Coin B})$

$=\binom{n}{k}\times\theta_B^k\times(1-\theta_B)^{n-k}$

$=\binom{15}{10}\times0.5^{10}\times0.5^5=0.0916$

Now the probability of biasness of coin A under this experiment E is,

$$\boldsymbol{P}(\text{Coin A}|E)=\frac{0.0245}{0.0916+0.0245}=0.2108$$

Now the probability of biasness of coin B under this experiment E is,

$$\boldsymbol{P}(\text{Coin B}|E)=\frac{0.0916}{0.0916+0.0245}=0.7892$$

For the coin A, the expected value of head is 0.2108*10H = 2.1085H and that of tail is 0.2108*5T =1.0542T.

For the coin B, the expected value of head is 0.7892 ×10H = 7.892H H and that of tail is 0.7892 ×5T =3.9458T.

For the second experiment,

$$\boldsymbol{P}(\text{HHTHHTHHHHTHHHH}|\text{Coin A})$$

$$= \binom{n}{k} \times \theta_A^k \times (1-\theta_A)^{n-k}$$

$$= \binom{10}{9} \times 0.4^{12} \times 0.6^3 = 0.0016$$

$$\boldsymbol{P}(\text{HHTHHTHHHHTHHHH}|\text{Coin B})$$

$$= \binom{n}{k} \times \theta_B^k \times (1-\theta_B)^{n-k}$$

$$= \binom{10}{9} \times 0.5^{12} \times 0.5^3 = 0.0139$$

Now the probability of biasness of coin A under this experiment E is,

$$\boldsymbol{P}(\text{Coin A}|E) = \frac{0.0016}{0.0016 + 0.0139} = 0.1061$$

Now the probability of biasness of coin B under this experiment E is,

$$\boldsymbol{P}(\text{Coin B}|E) = \frac{0.0139}{0.0016 + 0.0139} = 0.8939$$

For the coin A, the expected value of head is 0.1061 ×12H = 1.2737H and that of tail is 0.1061 ×3T =0.3184T.

For the coin B, the expected value of head is 0.8939 ×12H = 10.7263T and that of tail is 0.8939 ×3T = 2.6816T.

Doing the similar job for the rest of the other three experiments and the results are shown in Table III. Here, P(X|Y) is conditional probability of coin X given Y and E(X) is expected value of Coin X.

TABLE III: RESULTS OF EXPERIMENT

COIN	RESULTS	P(A|E)	P(B|E)	E(A)	E(B)
?	*HHTTHHHT*	0.2108	0.7892	2.1085H, 1.0542T	7.8915H, 3.9458T
?	HHTHHTHH	0.1061	0.8939	1.2737H, 0.3184T	10.7263T , 2.6816T
?	HTTTHTT THHHTTH H	0.4742	0.5258	3.3192H, 3.7933T	3.6808H, 4.2067T
?	HHTTTTT THTTHTT H	0.6698	0.3302	3.3492H, 6.6985T	1.6508H, 3.3015T
?	HHHTTTH HHTTHTT H	0.3755	0.6245	3.0036H, 2.6282T	4.9964H, 4.3718T
SUM				13.0542H, 14.4926T	28.9458H , 18.5074T

The probability, $\boldsymbol{P}(\text{Coin A}|E_i)$ and $\boldsymbol{P}(\text{Coin B}|E_i)$, for five experiments i =1, 2, 3, 4, 5 are evaluated in E-step.

M-step

Determine the sum of expected value of head and tail of both coins.

Update the biasing probability,

$$\theta_A^{(1)} = \frac{13.0542}{13.0542+14.4926} = 0.4739;\ \theta_B^{(1)} = \frac{28.9458}{28.9458+18.507} = 0.61$$

Using the biasing probability of first iteration as the input, evaluate $\theta_A^{(2)}$ and $\theta_A^{(2)}$. Continue the process until $\left|\theta_A^{(i)} - \theta_A^{(i+1)}\right| \le \varepsilon$ and $\left|\theta_B^{(i)} - \theta_B^{(i+1)}\right| \le \varepsilon$. For example, after 12 iterations no significant change in probability is found and the result is, $\theta_A^{(12)} = 0.4536$ and $\theta_B^{(12)} = 0.6776$. The variation of biasing angle against the number of iterations is shown in fig.6.

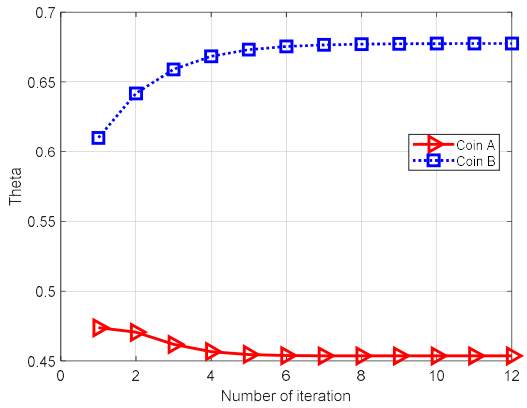


Fig. 6. Variation of biasing angle

C. Explanation of numerical example from EM algorithm

In E-step the algorithm determines the expected value of hidden variables (expected value of head and tail of coin A and B) using observed data (sequence of H and T of all experiments) and current value of parameters: $\theta_A^{(i)}$ and $\theta_B^{(i)}$.

In M-step the model parameters $\theta_A^{(i)}$ and $\theta_B^{(i)}$ are updated as: $\theta_A^{(i+1)}$ and $\theta_B^{(i+1)}$ with the condition of maximizing likelihood (or lower bound of likelihood) of the expected data found in E-step. The entire operation of EM is shown in fig.7

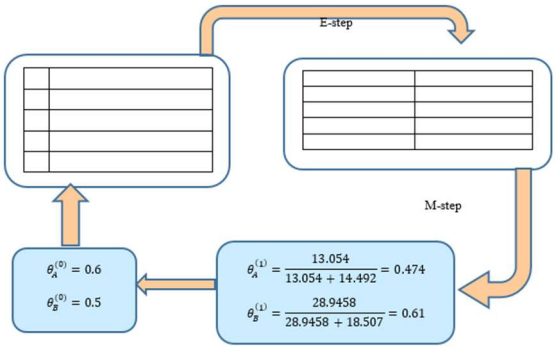


Fig. 7. Cycle of EM algorithm

The complete algorithm of coin tossing experiment is shown in Appendix.

VI. Conclusions

In this paper, a theoretical analysis of the EM algorithm with a simple numerical example is shown explicitly. In practical life EM algorithm is used to determine some unknown parameters of the incomplete data set. The hidden variables are determined from EM algorithm then is applied in other MLs like GMM and HMM (Hidden Markov Model) for data classification. In the future a composite model of EM+GMM or EM+HMM will be applied in the classification of tabular data or image set.

Appendix

Algorithm of Coin Tossing Experiment

Initialization: n - Size of Head-Tail sequence
K - Number of experiments
N - The number of iterations
Output: θ_{aU} and θ_{bU} %Updated theta
for s←1: K
*Read k(s) %*The number of heads of sth head-tail sequence*
end
for j←1:N, %Number of iteration
for i=1:K, %Number of experiment
H=k(i); T=n-k(i);
%Coin A
if count==1
theta_a=0.4; % initial value
else
theta_a=u_theta_a; %updated value
end
*y_a=$\binom{n}{k(i)}$ *(theta_a^k(i))*((1-theta_a)^(n-k(i)));*
%Coin B
if count==1
theta_b=0.5; % initial value
else
theta_b=u_theta_b; %updated value
end
y_b=$\binom{n}{k(i)}$(theta_b^k(i))*((1-theta_b)^(n-k(i)));*
P_con_a = y_a/(y_a+ y_b);
P_con_b = y_b/(y_a+ y_b);
H_a(i)=H P_con_a;*
T_a(i)=T P_con_a;*
H_b(i)=H P_con_b;*
T_b(i)=T P_con_b;*
end
θ_a(j)=sum(H_a)/(sum(H_a)+sum(T_a)); %Updated theta
θ_b(j)==sum(H_b)/(sum(H_b)+sum(T_b));
count=count+1;
end.

References

[1] Nima Sammaknejad, Yujia Zhao, Biao Huang, A review of the Expectation Maximization algorithm in data-driven process identification,' Journal of Process Control, Volume 73, , Pages 123-136, Jan'2019, https://doi.org/10.1016/j.jprocont.2018.12.010.

[2] N. K. Verma, S. Dwivedi and R. K. Sevakula, 'Expectation maximization algorithm made fast for large scale data,' *2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI)*, 2015, pp. 1-7, 14-17 December 2015, Kanpur, India, doi: 10.1109/WCI.2015.7495515

[3] P. Qian, Y. Guo and N. Li, 'Multitarget Localization with Inaccurate Sensor Locations via Variational EM Algorithm,' in *IEEE Sensors Letters*, vol. 3, no. 2, pp. 1-4, Feb. 2019, Art no. 7000204, doi: 10.1109/LSENS.2018.2889817

[4] H. Sawada, R. Ikeshita and T. Nakatani, 'Experimental Analysis of EM and MU Algorithms for Optimizing Full-rank Spatial Covariance Model,' *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 885-889, 18-21 January 2021 Amsterdam, Netherlands, doi: 10.23919/Eusipco47968.2020.9287336

[5] M. Przyborowski and D. Ślęzak, 'Approximation of the expectation-maximization algorithm for Gaussian mixture models on big data,' *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 6256-6260,17-20 December 2022, Osaka, Japan, doi: 10.1109/BigData55660.2022.10020450

[6] Y. T. Wu, H. J. Song, J. Zhou and Z. Lu, 'Reliability Analysis for Mixture Weibull Distribution with Progressively Censored Data Based on Stochastic Expectation-maximization Method,' *2022 13th International Conference on Reliability, Maintainability, and Safety (ICRMS)*, pp. 50-54, 21-24 August 2022, Kowloon, Hong Kong, doi: 10.1109/ICRMS55680.2022.9944579

[7] L. Zhuang, L. Gao, L. Ni and B. Zhang, 'An improved Expectation Maximization algorithm for hyperspectral image classification,' *2013 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1-4, 26-28 June 2013, Gainesville, FL, USA, 2013, doi: 10.1109/WHISPERS.2013.8080631.

[8] L. Yan *et al*., 'EM-Based Algorithm for Unsupervised Clustering of Measurements from a Radar Sensor Network,' in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, no. 1, pp. 787-801, Feb. 2025, doi: 10.1109/TAES.2024.3448390

[9] X. Wang, G. Wang, R. Fan and B. Ai, 'Channel Estimation with Expectation Maximization and Historical Information Based Basis Expansion Model for Wireless Communication Systems on High Speed Railways,' in *IEEE Access*, vol. 6, pp. 72-80, Feb' 2018, doi: 10.1109/ACCESS.2017.2745708

[10] D. Ron and J. -R. Lee, 'Expectation Maximization Based Power-Saving Method in Wi-Fi Direct,' in *IEEE Access*, vol. 8, pp. 158600-158611, August 2020, doi: 10.1109/ACCESS.2020.3014673

[11] S. Abramovich, B. Mond, J.E. Pečarić, 'Sharpening Jensen's Inequality and a Majorization Theorem,' Journal of Mathematical Analysis and Applications, vol. 214, Issue 2, pp.721-728, October 1997, https://doi.org/10.1006/jmaa.1997.5527

[12] Nisar, S., Zafar, F., Alamri, H. 'Improved Bounds for Integral Jensen's Inequality Through Fifth-Order differentiable Convex Functions and Applications,' *Axioms* 2025, *14(8)*, 602, pp.1-22, August 2025 https://doi.org/10.3390/axioms14080602

[13] da Silva, A.; Quintino, F.; Almeida, F.; Aguiar, D. Bias Reduction of Modified Maximum Likelihood Estimates for a Three-Parameter Weibull Distribution. *Entropy* 2025, *27(5)*, 485, pp.1-22, March 2025, https://doi.org/10.3390/e27050485

[14] Stijn van Lierop, Daniel Ramos, Marjan Sjerps, Rolf Ypma, 'An overview of log likelihood ratio cost in forensic science – Where is it used and what values can we expect?' Forensic Science International: Synergy, Volume 8, pp.1-12, March 2024, https://doi.org/10.1016/j.fsisyn.2024.100466

[15] Gemeay, A.M., Abd El-Bar, A.M.T., 'Bias Reduction of Maximum Likelihood Estimation in the Unit Odd Lindley Half-Logistic Model with Applications to Medicine and Geology Real Datasets,' Journal of Statistical Theory and Applications (2025), Springer, August 2025, https://doi.org/10.1007/s44199-025-00120-3

[16] A. P. Dempster; N. M. Laird; D. B. Rubin, 'Maximum Likelihood from Incomplete Data via the EM Algorithm,' Journal of the Royal

Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38
[17] S. Hua-Yan, L. Ye-Li, Z. Yun-Fei and H. Xu, 'Accelerating EM Missing Data Filling Algorithm Based on the K-Means,' 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC), 19-21 April 2018, pp. 401-406, Wuhan, China, doi: 10.1109/ICNISC.2018.00088
[18] Michiko Watanabe and Kazunori Yamaguchi, 'The EM Algorithm and Related Statistical Models,' STATISTICS: A DEKKER Series of TEXTBOOKS and MONOGRAPHS, Marcel Dekker, Inc. 2004, ISBN: 0-8247-4701-1
[19] Neal, R.M., Hinton, G.E. (1998). A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants. In: Jordan, M.I. (eds) Learning in Graphical Models. NATO ASI Series, vol 89. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-5014-9_12
[20] Thiesson, B., Meek, C. & Heckerman, D., 'Accelerating EM for Large Databases. *Machine Learning,'* vol. 45, pp.279–299, Dec' 2001, https://doi.org/10.1023/A:1017986506241
[21] José Silva, Paulo Vaz, Pedro Martins and Luís Ferreira, 'Reliability Estimation Using EM Algorithm with Censored Data: A Case Study on Centrifugal Pumps in an Oil Refinery,' *Appl. Sci.2023,13*(13), 7736, pp.1-11, June 2023, https://doi.org/10.3390/app13137736
[22] Caprio and A. M. Johansen, "Fast convergence of the expectation-maximization algorithm under a logarithmic Sobolev inequality," Biometrika, vol. 112, no. 4, pp. 1–12, Aug. 2025, doi: 10.1093/biomet/asaf061.
[23] Z. Xu, "An expectation-maximization algorithm for logistic regression based on individual-level predictors and aggregate-level response", Statistics in Transition, vol. 26, no. 1, pp. 9–28, 2025, doi: 10.59139/stattrans-2025-002.
[24] G. Zhang et al.**,** "An improved expectation–maximization Bayesian algorithm for GWAS," *Mathematics*, vol. 12, no. 13, Jun. 23, 2024, doi**:** 10.3390/math12131944
[25] M. Ritz and collaborators, "Maximum likelihood estimation for discrete latent variable models via evolutionary algorithms," Statistics and Computing, vol. 34, no. 62, Jan. 10, 2024, doi: 10.1007/s11222-023-10358-5.

Biography

Mr. Masum Bhuiyan is a Lecturer in the Department of Computer Science and Engineering at Jahangirnagar University, teaching and researching AI/ML-based real-world applications. With a distinguished background as a former Software Engineer at Samsung R&D Institute, where he contributed to the development of the Samsung Galaxy Watch and worked on a patent-pending technology graded A1(Highest Grade). Recognized for his outstanding contributions, he earned the Icon Award 2022. Academically, he has excelled with both M.Sc and B.Sc degrees in Computer Science and Engineering from Jahangirnagar University, graduating at the top of his class. As an innovator and leader, he founded Open AIR, a research community dedicated to fostering innovation. His top achievements include being the 2nd Runner-up in the National AI for Bangla 2.0 competition and securing a Special Innovation Fund grant from the ICT Ministry.

Samsun Nahar Khandakar received her B.Sc. (Honors) and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh in 2017 and 2018 respectively. Previously, she worked as an Assistant Professor in the Department of Computer Science and Engineering, University of Information Technology and Science, Dhaka, Bangladesh. Currently, she is working as a Lecturer in the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh. Her research interest is focused on Artificial Intelligence, Machine Learning, Deep Learning and IoT and Network Security.

Nadia Afrin Ritu received her B.Sc. (Honors) and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh in 2016 and 2017 respectively. Previously, she worked as an Assistant Professor in the Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Previously, she also worked as a lecturer in the Department of Computer Science and Engineering, Bangladesh University of Business and Technology (BUBT), Dhaka, Bangladesh. Currently, she is working as a lecturer in the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh. Her research interest is focused on Artificial Intelligence, Machine Learning and Expert System, and Data Mining.

Md. Imdadul Islam has completed his B.Sc. and M.Sc Engineering in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh in 1993 and 1998 respectively and has completed his Ph.D degree from the Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh in the field of network traffic in 2010. He is now working as a Professor at the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. Previously, he worked as an Assistant Engineer in Sheba Telecom (Pvt.) LTD (A joint venture company between Bangladesh and Malaysia, for Mobile cellular and WLL), from Sept.1994 to July 1996. Dr Islam has a very good field experience in installation and design of mobile cellular network, Radio Base Stations