Expectation-Maximization (EM) is a method used to find the maximizer of a likelihood function. EM is designed for latent variable models, where the observed variable y can be regarded as 'incomplete data' where the 'complete' data (x, y) includes a latent variable x also. The joint distribution of x, y is assumed to depend on some parameter $\theta \in \Theta$, hence shown as $p_\theta(x, y)$. Since we observe y only, a maximum likelihood estimation procedure for θ seeks to find

$$\theta_{\text{MLE}} = \arg\max_{\theta \in \Theta} p_\theta(y),$$

where

$$p_\theta(y) = \int p_\theta(x, y) dx.$$

I am using the integral sign in a general sense here. If x is a discrete random variable, the above integral would reduce to $p_\theta(y) = \sum_x p_\theta(x, y)$.

EM is an iterative algorithm, producing iterates $\theta^{(1)}, \theta^{(2)}, \ldots$ starting from an initial point $\theta^{(0)}$. One iteration of EM can be expressed as

$$\theta^{(j+1)} = \arg\max_{\theta \in \Theta} Q(\theta^{(j)}, \theta), \quad j \geq 0.$$

Here, $Q(\theta^{(j)}, \theta)$ is called the intermediate function of EM and it is an expectation of the form

$$Q(\theta', \theta) = \int \log p_\theta(x, y) p_{\theta'}(x|y) dx, \quad \theta, \theta' \in \Theta.$$

The evaluation of the above expectation is called the E-step and its maximization is called the M-step.

EM guarantees that $p_{\theta^{(j+1)}}(y) \geq p_{\theta^{(j)}}(y)$. Why? Firstly, observe that

$$\log p_\theta(y) = \log p_\theta(x, y) - \log p_\theta(x|y).$$

Taking the integral of both sides with respect to $p_{\theta'}(x|y)$, we get

$$\log p_\theta(y) = \int \log p_\theta(y) p_{\theta'}(x|y) dx = \int (\log p_\theta(x, y) - \log p_\theta(x|y)) p_{\theta'}(x|y) dx.$$

Reorganizing, we have

$$\begin{aligned} \log p_\theta(y) &= \int \log p_\theta(x, y) p_{\theta'}(x|y) dx - \int \log p_\theta(x|y) p_{\theta'}(x|y) dx \\ &= Q(\theta', \theta) - \int \log p_\theta(x|y) p_{\theta'}(x|y) dx. \end{aligned}$$

Note that this holds for any $\theta, \theta' \in \Theta$. Now, let $\theta = \theta^{(j+1)}$ and $\theta' = \theta^{(j)}$ and rewrite the above as

$$\log p_{\theta^{(j+1)}}(y) = Q(\theta^{(j)}, \theta^{(j+1)}) - \int \log p_{\theta^{(j+1)}}(x|y) p_{\theta^{(j)}}(x|y) dx$$

Again, let $\theta = \theta^{(j)}$ and $\theta' = \theta^{(j)}$ and rewrite the equality as

$$\log p_{\theta^{(j)}}(y) = Q(\theta^{(j)}, \theta^{(j)}) - \int \log p_{\theta^{(j)}}(x|y) p_{\theta^{(j)}}(x|y) dx.$$

Subtracting the second equation from the first, we get

$$\begin{aligned}\log p_{\theta^{(j+1)}}(y) - \log p_{\theta^{(j)}}(y) &= Q(\theta^{(j)}, \theta^{(j+1)}) - Q(\theta^{(j)}, \theta^{(j)}) - \int [\log p_{\theta^{(j+1)}}(x|y) - \log p_{\theta^{(j)}}(x|y)] p_{\theta^{(j)}}(x|y) dx \\ &= Q(\theta^{(j)}, \theta^{(j+1)}) - Q(\theta^{(j)}, \theta^{(j)}) - \int \log \frac{p_{\theta^{(j+1)}}(x|y)}{p_{\theta^{(j)}}(x|y)} p_{\theta^{(j)}}(x|y) dx\end{aligned}$$

Lets look at the differences $Q(\theta^{(j)}, \theta^{(j+1)}) - Q(\theta^{(j)}, \theta^{(j)})$ and $-\int \log \frac{p_{\theta^{(j+1)}}(x|y)}{p_{\theta^{(j)}}(x|y)} p_{\theta^{(j)}}(x|y) dx$. Since $\theta^{(j+1)}$ is the maximizer of $Q(\theta^{(j)}, \theta)$ over θ, it holds necessarily that

$$Q(\theta^{(j)}, \theta^{(j+1)}) - Q(\theta^{(j)}, \theta^{(j)}) \geq 0.$$

The second difference is also positive. This is because logarithm is a concave function, hence minus logarithm is a convex function. A convex function $f : \mathbb{R} \mapsto \mathbb{R}$, Jensen's inequality states that $\mathbb{E}[f(U)] \geq f(\mathbb{E}[U])$ for any distribution for U. Applying it to $f(u) = -\log u$ and $U = \frac{p_{\theta^{(j)}}(X|y)}{p_{\theta^{(j+1)}}(X|y)}$, we get

$$\int -\log \frac{p_{\theta^{(j)}}(x|y)}{p_{\theta^{(j+1)}}(x|y)} p_{\theta^{(j+1)}}(x|y) dx \geq -\log \int \frac{p_{\theta^{(j)}}(x|y)}{p_{\theta^{(j+1)}}(x|y)} p_{\theta^{(j+1)}}(x|y) dx = -\log 1 = 0.$$

Combining both inequalities, we finally obtain $\log p_{\theta^{(j+1)}}(y) \geq \log p_{\theta^{(j)}}(y)$.