EM Algorithm Writeup

Woojeong Kim

November 17, 2021

1 High-Level Concepts

Expectation–Maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between two steps:

1. Expectation (E) step: which determines the distribution of the latent variables with the current estimate for the parameters
2. Maximization (M) step: which computes parameters maximizing the expected log-likelihood found on the E step.

2 Method

2.1 Log-likelihood

Consider a probabilistic model with observed variables $\mathbf{X}$ and hidden variables $\mathbf{Z}$. The joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ is parameterized by θ. Our goal is to maximize the likelihood which is given by,

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta). \tag{1}$$

The implicit assumption underlying the EM algorithm is that it is difficult to optimize $p(\mathbf{X}|\theta)$ with respect to θ but that it is much easier to optimize $p(\mathbf{X}, \mathbf{Z}|\theta)$.

Now we introduce arbitrary distribution $q(\mathbf{Z})$ over latent variables $\mathbf{Z}$. Then log-likelihood would be

$$\ln p(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} = \ln \mathbb{E}_{q(\mathbf{Z})} \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}. \tag{2}$$

Applying Jensen's inequality to Equation 2 ($\ln x$ is a concave function),

$$\ln p(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\} = L(q, \theta), \tag{3}$$

we can get a lower bound of log-likelihood and denote it as $L(q, \theta)$. Note that $L(q, \theta)$ is a functional of the distribution $q(\mathbf{Z})$, and a function of the parameters θ.

Now we derive the difference between log-likelihood and its lower bound.

$$\begin{aligned}
\ln p(\mathbf{X}|\theta) - L(q,\theta) &= \ln p(\mathbf{X}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\} \\
&= \ln p(\mathbf{X}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)p(\mathbf{X}|\theta)}{q(\mathbf{Z})} \right\} \\
&= \ln p(\mathbf{X}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \right\} - \ln p(\mathbf{X}|\theta) \sum_{\mathbf{Z}} q(\mathbf{Z}) \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\theta)} \right\} \\
&= \mathrm{KL}\left(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\theta)\right) \\
&= \mathrm{KL}\left(q||p\right).
\end{aligned} \tag{4}$$

To sum up, log-likelihood can be divided into two parts.

$$\ln p(\mathbf{X}|\theta) = L(q,\theta) + \mathrm{KL}\left(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\theta)\right). \tag{5}$$

KL divergence in the second term is zero if $p = q$; otherwise it's strictly positive. Choosing $q(\mathbf{Z})$ that minimizes KL is equivalent to maximizing $L(q,\theta)$. Our goal is to maximize $L(q,\theta)$ to be equal to the log-likelihood by appropriately choosing an arbitrary $q(\mathbf{Z})$ distribution.

2.2 EM Algorithm

2.2.1 E-step

Given a time step t, suppose the current value of the parameter is $\theta^{(t)}$. In E-step, we choose $q(\mathbf{Z})$ which maximizes $L(q,\theta^{(t)})$, while holding $\theta^{(t)}$ fixed. We have seen in Equation 5 that the largest value occurs when KL divergence is equal to 0, so $p = q$. In other words, q is set equal to the posterior distribution for the current parameter values $\theta^{(t)}$; $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\theta^{(t)})$.

2.2.2 M-step

In M-step, $q(\mathbf{Z})$ is fixed and lower bound $L(q,\theta)$ is maximized with respect to θ to obtain new parameters $\theta^{(t+1)}$. This will cause the lower bound $L(q,\theta)$ to increase, which will necessarily cause the corresponding log likelihood function to increase.

If we substitute $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\theta^{(t)})$ into Equation 3, after the E step, the lower bound takes the form below.

$$\begin{aligned}
L(q,\theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{(t)}) \ln p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{(t)}) \ln p(\mathbf{Z}|\mathbf{X},\theta^{(t)}) \\
&= Q(\theta,\theta^{(t)}) + \mathcal{C}
\end{aligned} \tag{6}$$

Since we maximize $L(q,\theta)$ with respect to θ, the second term of Equation 6 is a nonnegative constant. We denote the first term as $Q(\theta,\theta^{(t)})$ and maximize it instead of $L(q,\theta)$. Note that $Q(\theta,\theta^{(t)}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta^{(t)})} \ln p(\mathbf{X},\mathbf{Z}|\theta)$.

3 Reasoning Behind the Math

In order to apply EM algorithm, the following three properties must be held.

1. There has to be a latent variable, $\mathbf{Z}$
2. There has to be a parameter, θ, we don't know.
3. We need to be able to compute the posterior probability, $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\theta^{(t)})$, efficiently.

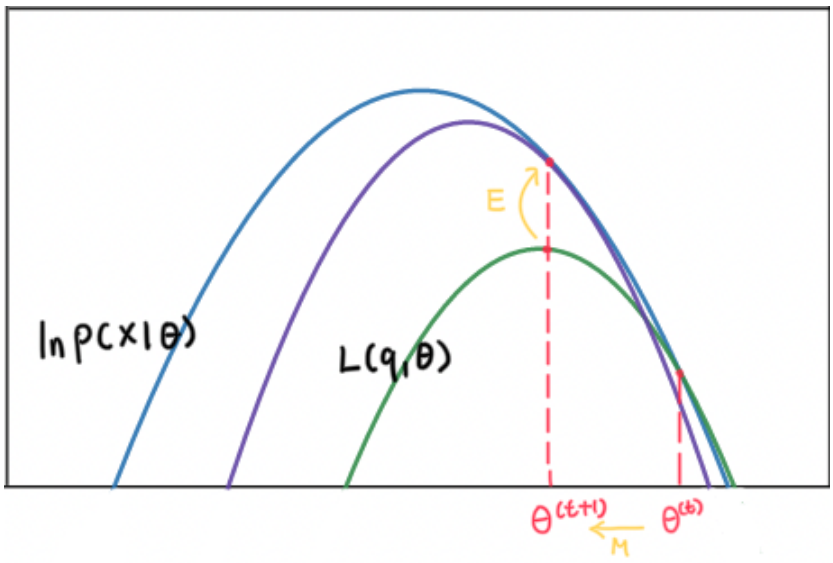

Figure 1: Overview of EM algorithm. Blue curve is log-likelihood function $\ln p(\mathbf{X}|\theta)$ we want to maximize. In E step, we evaluate the posterior distribution over latent variables, giving rise to a lower bound $Ł(q, \theta)$. In the M step, the bound (green curve) is maximized giving the value $\theta^{(t+1)}$, which gives a larger value of log likelihood than $\theta^{(t)}$.

The general EM algorithm is summarized as below.

1. E step: With fixed $\theta^{(t)}$, evaluate the posterior $p(\mathbf{Z}|\mathbf{X}, \theta^{(t)})$
2. M step: Evaluate $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$

EM algorithm involves alternately computing a lower bound on the log-likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. Figure 3 illustrates this procedure in terms of the parameter θ. We can see that alternating E and M steps change the model parameters in a way to increase log-likelihood. (Can I say "EM is guaranteed to increase log-likelihood if $\ln p(\mathbf{X}|\theta)$ is concave"?)