CS5487 - EM algorithm as optimizing a lower bound

Antoni Chan
Department of Computer Science
City University of Hong Kong

In this handout, we will interpret the EM algorithm as maximizing a lower-bound function on the data log-likeilhood.

1 Lower-bound to data log-likelihood

Let X be the observed variables, Z be the hidden variables, and $p(X, Z|\theta)$ be the joint log-likelihood with parameters θ. The data likelihood is

$$p(X|\theta) = \sum_Z p(X, Z|\theta) \tag{1}$$

We first decompose the data log-likelihood into two terms as follows (we don't write θ to remove clutter),

$$\log p(X, Z) = \log p(Z|X) + \log p(X) \tag{2}$$

$$\log p(X) = \log \frac{p(X, Z)}{p(Z|X)} \tag{3}$$

We next introduce a distribution $q(Z)$ over the hidden variables Z. This is sometimes called a *variational distribution* because we are allowed to optimize it for our purpose. Then for any choice of distribution q, we have

$$\log p(X) = \log \frac{p(X, Z)q(Z)}{p(Z|X)q(Z)} \tag{4}$$

$$q(Z) \log p(X) = q(Z) \log \frac{p(X, Z)q(Z)}{p(Z|X)q(Z)} \tag{5}$$

$$\sum_Z q(Z) \log p(X) = \sum_Z q(Z) \log \frac{p(X, Z)q(Z)}{p(Z|X)q(Z)} \tag{6}$$

$$\log p(X) = \sum_Z q(Z) \log \frac{p(X, Z)q(Z)}{p(Z|X)q(Z)}, \tag{7}$$

where the last line follows from $\sum_Z q(Z) = 1$. Finally,

$$\log p(X) = \sum_Z q(Z) \log \frac{p(X, Z)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{p(Z|X)} \tag{8}$$

$$= \mathcal{L}(q, \theta) + \mathrm{KL}(q||p), \tag{9}$$

where we define the two terms:

$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \frac{p(X, Z)}{q(Z)}, \tag{10}$$

$$\mathrm{KL}(q||p) = \sum_Z q(Z) \log \frac{q(Z)}{p(Z|X)}. \tag{11}$$

The 2nd term is the Kullback-Leibler (KL) divergence between two distributions q and p, which is a measure of dissimilarity (see PRML Ch. 1.6). An important property of KL divergence is that it is non-negative, $\mathrm{KL}(q||p) \geq 0$.[1] Furthermore, $\mathrm{KL}(q||p) = 0$ if and only if $q = p$.

Since $\mathrm{KL}(q||p) \geq 0$, then from (9) we have

$$\log p(X|\theta) \geq \mathcal{L}(q, \theta). \tag{12}$$

In other words, $\mathcal{L}(q, \theta)$ is a lower-bound function of the data log-likelihood. This bound holds for any choice of the variational distribution q, with equality when $q = p$.

2 Optimizing the lower bound

Now we consider maximizing the lower bound $\mathcal{L}(q, \theta)$ as a proxy for maximizing the data log-likelihood $\log p(X|\theta)$. There are now two things to optimize in $\mathcal{L}$, the variational distribution q and the parameters θ, We will optimize these in an alternating way: 1) hold θ fixed and optimize q; 2) hold q fixed and optimize θ.

1. Suppose the current parameters are $\theta^{(\mathrm{old})}$, our goal is to optimize w.r.t. q,

$$q^* = \operatorname*{argmax}_q \mathcal{L}(q, \theta^{(\mathrm{old})}) \tag{13}$$

   From (12), we know that the maximum value of $\mathcal{L}(q, \theta)$ is $\log p(X|\theta)$, and this occurs when $\mathrm{KL}(q||p) = 0$. Thus the maximum occurs at $q^*(Z) = p(Z|X, \theta^{(\mathrm{old})})$ and $\mathcal{L}(q^*, \theta^{(\mathrm{old})}) = \log p(X|\theta^{(\mathrm{old})})$.

2. Next suppose we fix q^* according to the above, our goal is to optimize w.r.t. θ,

$$\theta^{(\mathrm{new})} = \operatorname*{argmax}_\theta \mathcal{L}(q^*, \theta) \tag{14}$$

   Substituting q^* into $\mathcal{L}(q, \theta)$,

$$\begin{aligned}
\mathcal{L}(q^*, \theta) &= \sum_Z p(Z|X, \theta^{(\mathrm{old})}) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta^{(\mathrm{old})})} && (15)\\
&= \sum_Z p(Z|X, \theta^{(\mathrm{old})}) \log p(X, Z|\theta) - \sum_Z p(Z|X, \theta^{(\mathrm{old})}) \log p(Z|X, \theta^{(\mathrm{old})}) && (16)\\
&= \mathbb{E}_{Z|X,\theta^{(\mathrm{old})}}[\log p(X, Z)] + H[p(Z|X, \theta^{(\mathrm{old})})], && (17)
\end{aligned}$$

   where $H[p]$ is the entropy of p. Note that $p(Z|X, \theta^{(\mathrm{old})})$ is not a function of θ, and thus $H[p(Z|X, \theta^{(\mathrm{old})})]$ is a constant. The first expectation term is the $\mathcal{Q}$ function from EM, and thus

$$\mathcal{L}(q^*, \theta) = \mathcal{Q}(\theta; \theta^{(\mathrm{old})}) + \mathrm{const}, \tag{18}$$

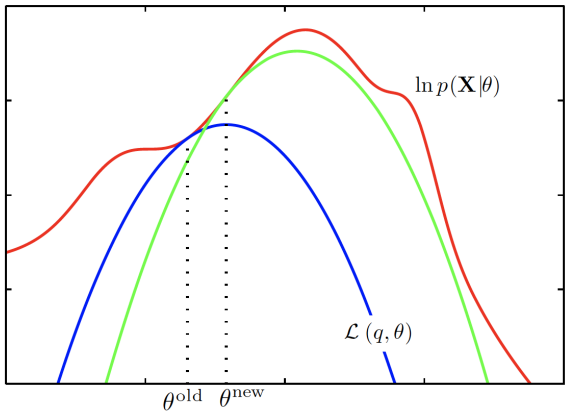
   and the optimization problem becomes:

$$\theta^{(\mathrm{new})} = \operatorname*{argmax}_\theta \mathcal{Q}(\theta; \theta^{(\mathrm{old})}). \tag{19}$$

Interestingly, (19) is exactly the M-step of the EM algorithm! The E-step is to compute $p(Z|X, \theta^{(\mathrm{old})})$ in Step 1, and substitute into $\mathcal{L}$ in (17). Therefore, the EM algorithm can be interpreted as maximizing the lower bound $\mathcal{L}(q, \theta)$ of the data log likelihood $\log p(X|\theta)$. Given a $\theta^{(\mathrm{old})}$, the E-step constructs a lower bound function $\mathcal{L}(q^*, \theta)$, which touches $\log p(X|\theta^{(\mathrm{old})})$ at $\theta^{(\mathrm{old})}$ (since at $\mathcal{L}(q^*, \theta^{(\mathrm{old})}) = \log p(X|\theta^{(\mathrm{old})})$). Then the M-step maximizes the constructed bound $\mathcal{L}(q^*, \theta)$, w.r.t, θ. This gives a new point $\theta^{(\mathrm{new})}$ for constructing a new lower bound in the next E-step. This is illustrated in the below figure from PRML.

[1] Note that KL divergence is not symmetric, and the version $\mathrm{KL}(q||p)$ where the "approximate" distribution is in the left position and the "true" distribution is in the right position is sometimes called the "reverse" KL. The "forward" KL swaps the the two arguments, $\mathrm{KL}(p||q)$.

Figure 9.14 The 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. See the text for a full discussion.



From the analysis, we see that each step of EM is maximizing the lower-bound, and thus EM should converge to a maximum. Furthermore, at the maximum, we have $\mathcal{L}(q^*, \theta^{(\text{old})}) = \log p(X|\theta^{(\text{old})})$, and thus we have also maximized the data log-likelihood. More details can be found in Chapter 9.4 in PRML.

The above analysis also shows that in each step we can partially optimize the lower-bound. For example, in the E-step, we can use a different solution for the variational distribution (suppose $p(Z|X)$ doesn't have a tractable form), and we will still construct a lower-bound, but not tangent to $\log p(X)$. In the M-step, we do not need to fully optimize θ, which means we only move partially up the contructed lower bound (this is called *generalized EM*).