Theory of EM algorithm

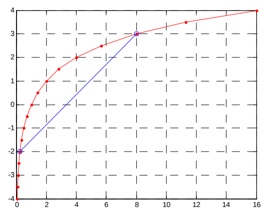
Reference to the book "Pattern Recognition and Machine Learning" by C. M. Bishop.

$\ln p(X|\theta) = L(q,\theta) + KL(q\|p)$, (1)

Where $L(q,\theta) = \sum_z q(z) \ln \frac{p(X,z|\theta)}{q(z)}$ and $KL(q\|p) = -\sum_z q(z) \ln \frac{p(z|X,\theta)}{q(z)}$

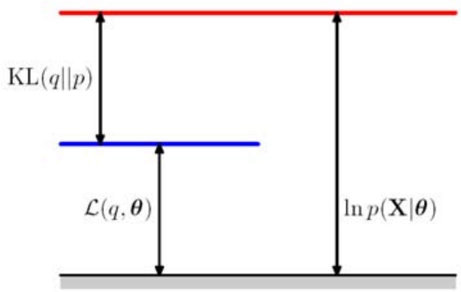
By examining $KL(q\|p)$ which is a Kullback–Leibler divergence of two distributions, we find out $L(q,\theta)$ is always a lower bound of our objective function $\ln p(X|\theta)$!

Specifically, $KL(q\|p) \geq 0$ and equality occurs iff $p = q$, and we prove it as follows. We first note that $\ln x$ is a strictly concave function.



i.e. $\sum_i \alpha_i \ln(x_i) \leq \ln\left(\sum_i \alpha_i x_i\right)$, for $\sum_i \alpha_i = 1$, and equality holds iff $x_i = 1, \forall i$. So we have $KL(q\|p) = -\sum_z q(z) \ln \frac{p(z|X,\theta)}{q(z)} \geq -\ln\left(\sum_z p(z|X,\theta)\right) = 0$, equality hold iff $p = q$.

No matter how we chose the distribution $q(z)$, (1) always holds. So if we set $q(z) = p(z|X,\theta)$, then we have $\ln p(X|\theta) = L(q,\theta)$.



Let we have a previously estimated parameters θ_{old}. From (1) we have

$$\ln p(X|\theta_{old}) = L(q,\theta_{old}) + KL(q\|p) = \sum_z q(z) \ln \frac{p(X,z|\theta_{old})}{q(z)} + \left(-\sum_z q(z) \ln \frac{p(z|X,\theta_{old})}{q(z)}\right)$$

If we set $q^*(z) = p(z|X,\theta_{old})$. We have

$$\ln p(\mathrm{X} \mid \theta_{old}) = L(q^*, \theta_{old}) = \sum_z q^*(z) \ln \tfrac{p(\mathrm{X}, z \mid \theta_{old})}{q^*(z)}$$

Setting $\theta_{new} = \arg\max_\theta L(q^*, \theta)$. We have a new lower bound $L(q^*, \theta_{new}) \geq L(q^*, \theta_{old})$.

And we have the following relation:

$$\ln p(\mathrm{X} \mid \theta_{old}) = L(q^*, \theta_{old}) \leq L(q^*, \theta_{new}) \leq L(q^*, \theta_{new}) + KL(q^* \| p) = \ln p(\mathrm{X} \mid \theta_{new}).$$

θ_{new} increases the objective function now! Again, setting $q' = p(z \mid \mathrm{X}, \theta_{new})$ and maximizing $L(q', \theta)$, we could have a sequence of non-decreasing objective function values.

We conclude two main steps here:

(1) E-step: Setting $q^*(z) = p(z \mid \mathrm{X}, \theta_{old})$ which induces to the lower bound $L(q^*, \theta) = \sum_z q^*(z) \ln \frac{p(\mathrm{X}, z \mid \theta)}{q^*(z)}$ having equality with objective function while $\theta = \theta_{old}$ (i.e. $\ln p(\mathrm{X} \mid \theta_{old}) = L(q^*, \theta_{old})$).

(2) M-step:

$$\theta_{new} = \arg\max_\theta L(q^*, \theta) = \arg\max_\theta \sum_z q^*(z) \ln \frac{p(\mathrm{X}, z \mid \theta)}{q^*(z)} = \arg\max_\theta \sum_z q^*(z) \ln p(\mathrm{X}, z \mid \theta)$$

An alternative interpretation of M-step is finding the parameters that maximize the complete data log-likelihood under the expectation of missing variables.

Figure 9.12 Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values $\boldsymbol{\theta}^{\text{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

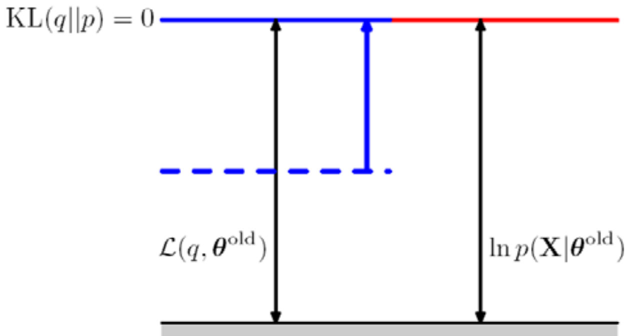


Figure 9.13 Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.

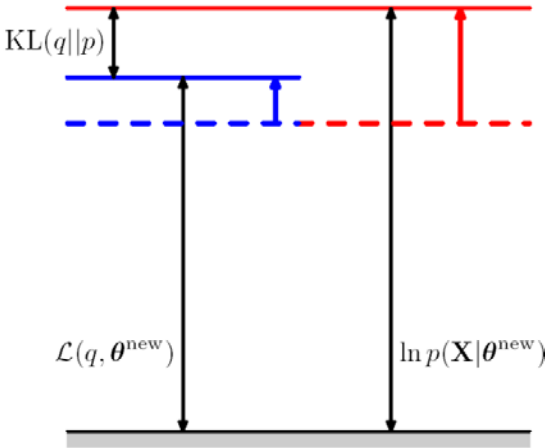
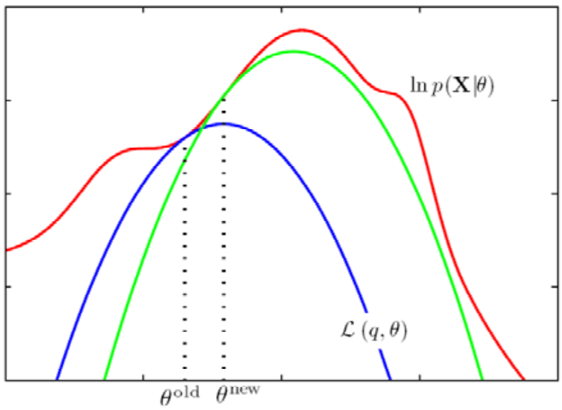


Figure 9.14 The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.



Maximum a posterior could also be achieved by EM. To see this, we note that $\ln p(\theta | \mathrm{X}) = \ln p(\mathrm{X} | \theta) + \ln p(\theta) - \ln p(\mathrm{X})$, so we have

$$\ln p(\theta | \mathrm{X}) = \ln p(\mathrm{X} | \theta) + \ln p(\theta) - \ln p(\mathrm{X})$$
$$= L(q, \theta) + \ln p(\theta) + KL(q \| p) - \ln p(\mathrm{X})$$

The only difference of MAP EM and ML EM is that the M-step is involved maximizing $L(q, \theta) + \ln p(\theta)$

Also in practice, we usually have multiple observations $\mathrm{X} = \{x_1 \ldots x_N\}$. The EM algorithm leads to maximize $\ln p(\mathrm{X} | \theta) = \ln \prod_i p(x_i | \theta) = \sum_i \ln p(x_i | \theta)$. So we adopt EM for each observation $\ln p(x_i | \theta)$, and the whole EM cycle is the same as before except we have a summation for all observations from the outside.