patern recognition and machine learning　Bishop

EM(Expectation Maximization)

Consider a probabilistic model in which we collectively denote all of the observed variables by X and all of the hidden variables by Z. The joint distribution $p(X, Z|\theta)$ is governed by a set of parameters denoted θ. Our goal is to maximize the likelihood function that is given by

$$p(X|\theta) = \sum_Z p(X, Z|\theta)$$

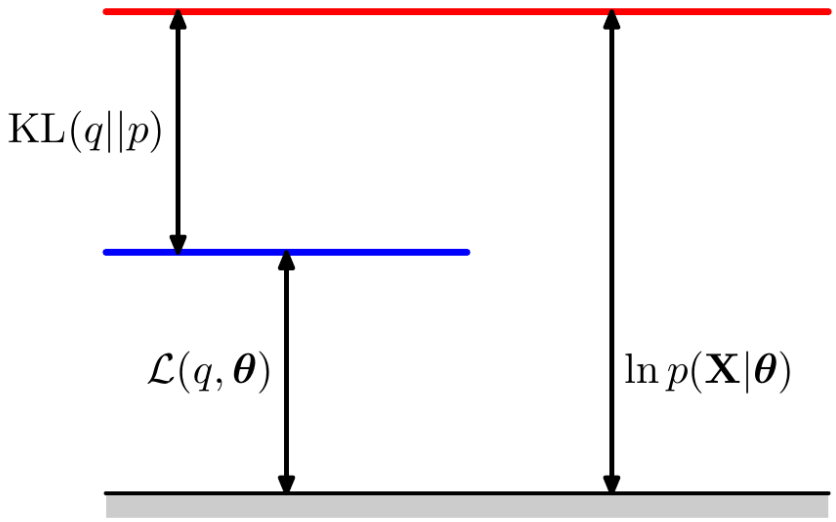
Next we introduce a distribution q(Z) defined over the latent variables, and we observe that, for any choice of q(Z), the following decomposition holds

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

where we have defined

$$L(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\} \tag{2}$$

$$\mathrm{KL}(q\|p) = -\sum_Z q(Z) \ln \left\{ \frac{p(Z|X, \theta)}{q(Z)} \right\} \tag{3}$$

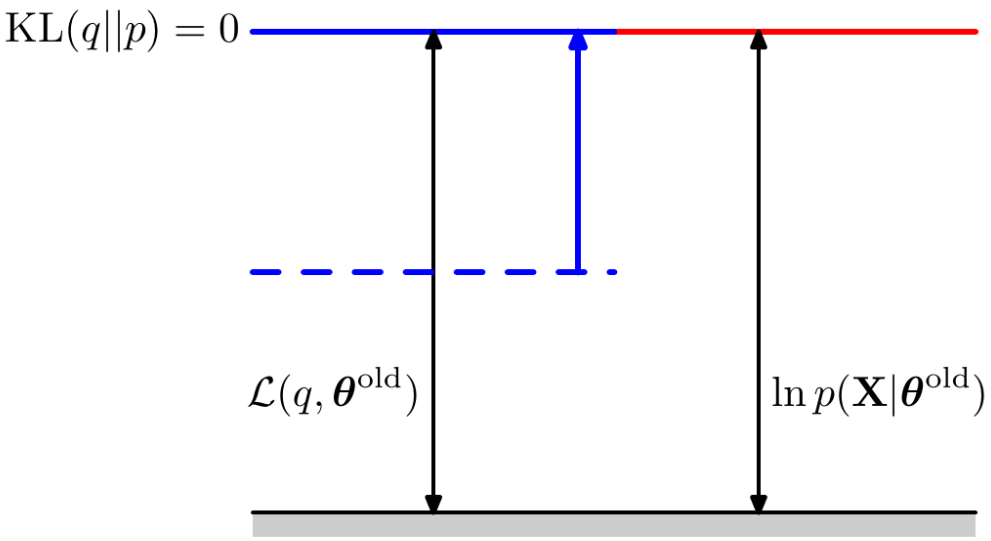


where the Kullback-Leibler divergence satisfies $\mathrm{KL}(q\|p) > 0$, the quantity $L(q, \theta)$ is a lower bound on the log likelihood function $\ln p(X|\theta)$

We can use the decomposition (1) to define the EM algorithm and to demonstrate that it does indeed maximize the log likelihood.

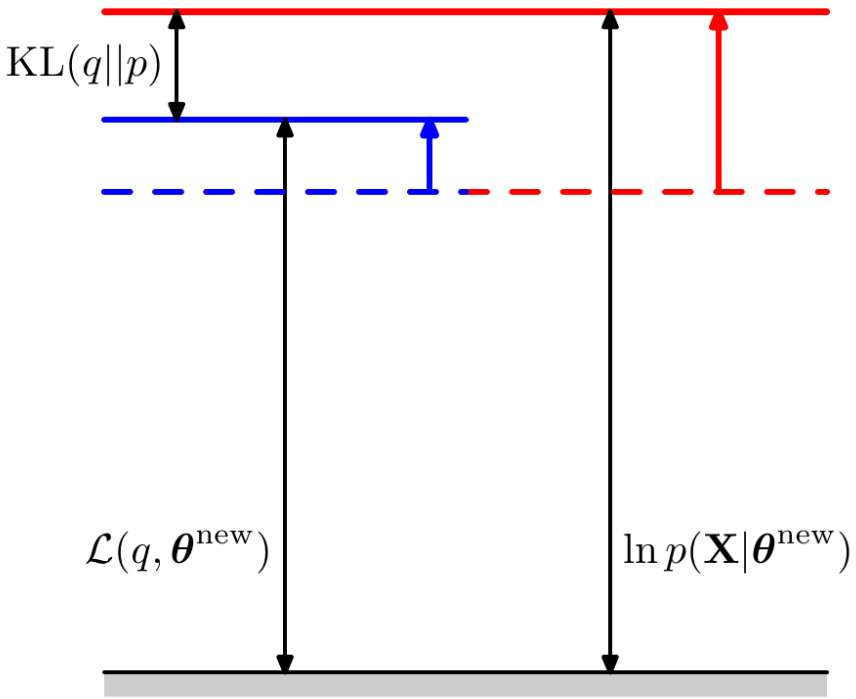
Suppose that the current value of the parameter vector is θ^{old} . In the E step, the lower bound $L(q, \theta^{\mathrm{old}})$ is maximized with respect to $q(Z)$ while holding θ^{old} fixed. The solution to this maximization problem is easily seen by noting that the value of $\ln p(X|\theta^{\mathrm{old}})$ does not depend on $q(Z)$ and so the largest value of $L(q, \theta^{\mathrm{old}})$ will occur when the Kullback-Leibler divergence vanishes,

in other words when $q(Z)$ is equal to the posterior distribution $p(Z|X, \theta^{\text{old}})$. In this case, the lower bound will equal the log likelihood, as illustrated in the figure:



In the subsequent M step, the distribution $q(Z)$ is held fixed and the lower bound $L(q, \theta)$ is maximized with respect to θ to give some new value θ new . This will cause the lower bound L to increase (unless it is already at a maximum), which will necessarily cause the corresponding log likelihood function to increase.

Because the distribution q is determined using the old parameter values rather than the new values and is held fixed during the M step, it will not equal the new posterior distribution p(Z|X, ϑ new), and hence there will be a nonzero KL divergence. The increase in the log likelihood function is therefore greater than the increase in the lower bound, as

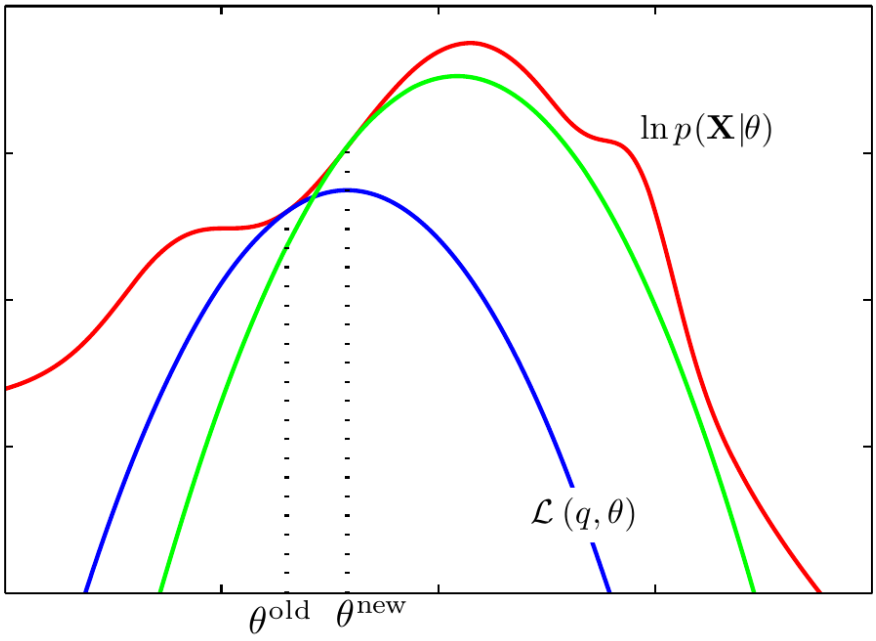


If we substitute $q(Z)=p(Z|X,\theta^{\text{old}})$ into (2), we see that, after the E step, the lower bound takes the form

$$\begin{aligned} L(q,\theta) &= \sum_Z p(Z|X,\theta^{\text{old}})\ln p(X,Z|\theta) - \sum_Z p(Z,X,\theta^{\text{old}})\ln p(Z|X,\theta^{\text{old}}) \\ &= \mathcal{Q}(\theta,\theta^{\text{old}}) + \text{cons} \end{aligned}$$

where the constant is simply the negative entropy of the q distribution and is therefore independent of θ. Thus in the M step, the quantity that is being maximized is the expectation of the complete-data log likelihood, as we saw earlier in the case of mixtures of Gaussians.

The operation of the EM algorithm can also be viewed in the space of parameters, as illustrated schematically in figure



Here the red curve depicts the (incomplete data) log likelihood function whose value we wish to maximize. We start with some initial parameter value θ^{old}, and in the first E step we evaluate the posterior distribution over latent variables, which gives rise to a lower bound $L(\theta,\theta^{\text{old}})$ whose value equals the log likelihood at θ^{old}, as shown by the blue curve. Note that the bound makes a tangential contact with the log likelihood at θ^{old}, so that both curves have the same gradient. This bound is a convex function having a unique maximum (for mixture components from the exponential family). In the M step, the bound is maximized giving the value θ^{new}, which gives a larger value of log likelihood than θ^{old}. The subsequent E step then constructs a bound that is tangential at θ^{new} as shown by the green curve.

Note 1.

EM 算法将$Z\sim p(z|\theta,X)$与θ看作两个参数，交替进行优化。

E步时固定$\theta=\theta^{\text{old}}$，将$q(Z)$取为$p(Z|\theta^{\text{old}},X)$，得

$L(\theta)=E_{Z\sim p(Z|\theta^{\text{old}},X)}\ln\frac{p(X,Z|\theta)}{p(Z|\theta^{\text{old}},X)}=E_{Z\sim p(Z|\theta^{\text{old}},X)}p(X,Z|\theta)+\text{const}$

M步时对Z步得到的期望进行优化，

$\theta^{\text{new}}=\arg\max_\theta L(\theta)=\arg\max_\theta E_{Z\sim p(Z|\theta^{\text{old}},X)}p(X,Z|\theta)$

可以不使用$q(Z)$，直接按照交替优化的思路分析EM算法。

$$\begin{aligned} p(X|\theta)p(Z|X,\theta) &= p(X,Z|\theta) \\ \ln p(X|\theta) &= \ln p(X,Z|\theta) - \ln p(Z|X,\theta) \end{aligned}$$

由于$p(X|\theta)$中没有Z,得

$$\ln p(X|\theta) = E_Z \ln p(X,Z|\theta) - E_Z \ln p(Z|X,\theta) \tag{4}$$

其中$Z \sim p(Z|X,\theta)$。

交替优化。

E步：

首先假设当前$\theta = \theta^{\text{old}}$，得$Z \sim p(Z|X,\theta^{\text{old}})$，由(4)得

$$\begin{aligned} \ln p(X|\theta) &= E_Z \ln \frac{p(X,Z|\theta)}{p(Z|X,\theta^{\text{old}})} - E_Z \ln \frac{p(Z|X,\theta)}{p(Z|X,\theta^{\text{old}})} \\ &= L(\theta) + K(\theta) \\ L(\theta) &= E_Z \ln \frac{p(X,Z|\theta)}{p(Z|X,\theta^{\text{old}})} \\ K(\theta) &= -E_Z \ln \frac{p(Z|X,\theta)}{p(Z|X,\theta^{\text{old}})} \end{aligned}$$

其中$Z \sim p(Z|X,\theta^{\text{old}})$,且有$\ln p(X|\theta^{\text{old}}) = L(\theta^{\text{old}})$。下面分析$\ln p(X|\theta)$与$L(\theta)$之间的关系。

$$\begin{aligned} \ln(x) &\leqslant x - 1 \\ E_Z \ln \frac{p(Z|X,\theta)}{p(Z|X,\theta^{\text{old}})} &\leqslant E_Z \frac{p(Z|X,\theta)}{p(Z|X,\theta^{\text{old}})} - 1 \\ &\leqslant \int \frac{p(Z|X,\theta)}{p(Z|X,\theta^{\text{old}})} p(Z|X,\theta^{\text{old}}) \mathrm{d}Z - 1 \\ &\leqslant \int p(Z|X,\theta) \mathrm{d}Z - 1 \\ &\leqslant 0 \end{aligned}$$

因此 $K(\theta) \geqslant 0$,得

$$\ln p(X|\theta) \geqslant L(\theta)$$

可知当$L(\theta) > L(\theta^{\text{old}})$时，有

$$\ln p(X|\theta) > \ln p(X|\theta^{\text{old}}) = L(\theta^{\text{old}})$$

M步

优化θ，由于

$$\begin{aligned} L(\theta) &= E_Z \ln \frac{p(X,Z|\theta)}{p(Z|X,\theta^{\text{old}})} \\ &= E_Z \ln p(X,Z|\theta) - E_Z \ln p(Z|X,\theta^{\text{old}}) \\ &= E_Z \ln p(X,Z|\theta) - \text{const} \end{aligned}$$

因此

$$\theta^{\text{new}} = \arg\max_{\theta} E_Z \ln p(X,Z|\theta)$$