Notes on EM Algorithm in PRML

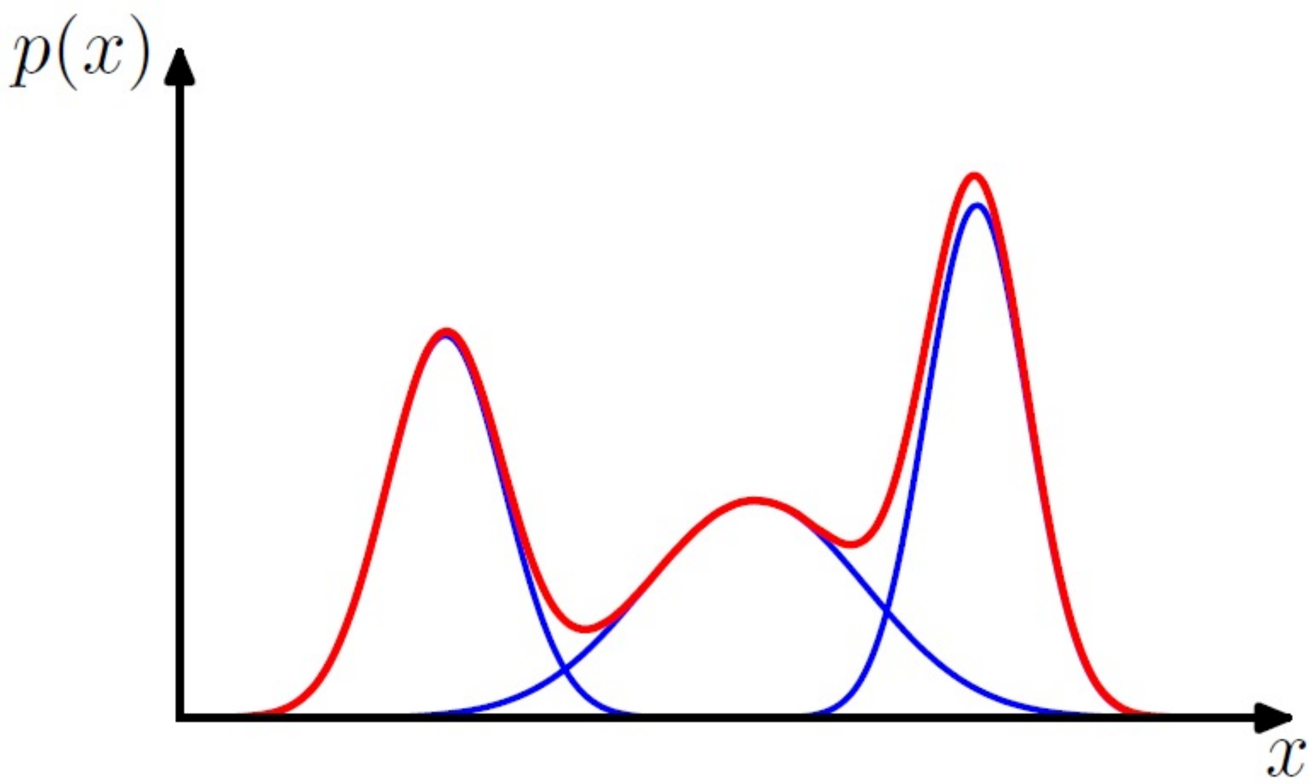
https://hustwj.github.io/notes/
https://github.com/hustwj/notes

The expectation maximization algorithm, or EM algorithm, is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables. 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

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

Here we are assuming Z is discrete, although the discussion is identical if Z comprises continuous variables or a combination of discrete and continuous variables, with summation replaced by integration as appropriate.

We shall suppose that direct optimization of $p(X; \theta)$ is difficult, but that optimization of the complete-data likelihood function $p(X, Z; \theta)$ is significantly easier. As shown in the following figure, $p(X; \theta)$ represented by the sum curve in red is relatively harder to optimize, and $p(X, Z = k; \theta), 1 \leq k \leq 3$ represented by three blue curves corresponding to three Gaussian distributions (each scaled by a coefficient) are easy to optimize respectively.



We know

$$p(Z|X;\theta) = \frac{p(X,Z;\theta)}{p(X;\theta)}$$
$$p(X;\theta) = \frac{p(X,Z;\theta)}{p(Z|X;\theta)}$$

So

$$\ln p(X;\theta) = \ln \frac{p(X,Z;\theta)}{p(Z|X;\theta)}$$

We introduce a distribution $q(Z)$ defined over the latent variables Z, and we observe that, for any choice of $q(Z)$, we always have

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

Note:
$q(Z)$ is actually a function defined over Z, and there are many different possible options for the function. For any choice of function $q(Z)$, the above equation holds.

Because $\ln p(X;\theta)$ doesn't contain Z, so it won't be affected by the distribution of $q(Z)$. For any value of X, we always get

$$\begin{aligned}\mathbb{E}_Z[\ln p(X;\theta)] &= \sum_Z q(Z) \times \ln p(X;\theta) \\ &= \ln p(X;\theta) \times \sum_Z q(Z) \\ &= \ln p(X;\theta)\end{aligned}$$

We can also get

$$\begin{aligned}&\mathbb{E}_Z[\ln \frac{p(X,Z;\theta)}{q(Z)} - \ln \frac{p(Z|X;\theta)}{q(Z)}] \\ =&\mathbb{E}_Z[\ln \frac{p(X,Z;\theta)}{q(Z)}] - \mathbb{E}_Z[\ln \frac{p(Z|X;\theta)}{q(Z)}] \\ =&\sum_Z q(Z) \times \ln \frac{p(X,Z;\theta)}{q(Z)} - \sum_Z q(Z) \times \ln \frac{p(Z|X;\theta)}{q(Z)}\end{aligned}$$

We can define

$$\begin{aligned}\mathcal{L}(q;\theta) &= \sum_Z q(Z) \times \ln \frac{p(X,Z;\theta)}{q(Z)} \\ KL(q||p) &= -\sum_Z q(Z) \times \ln \frac{p(Z|X;\theta)}{q(Z)}\end{aligned}$$

So we can get

$$\ln p(X;\theta) = \mathcal{L}(q;\theta) + KL(q||p)$$

$\mathcal{L}(q;\theta)$ is a **functional** of the distribution $q(Z)$ (because $q(Z)$ is a function as input), and a function of the parameters θ. $\mathcal{L}(q;\theta)$ contains the joint distribution of X and Z while $KL(q||p)$ contains the conditional distribution of Z given X. $KL(q||p)$ is the **Kullback-Leibler divergence** between $q(Z)$ and the posterior distribution $p(Z|X;\theta)$.

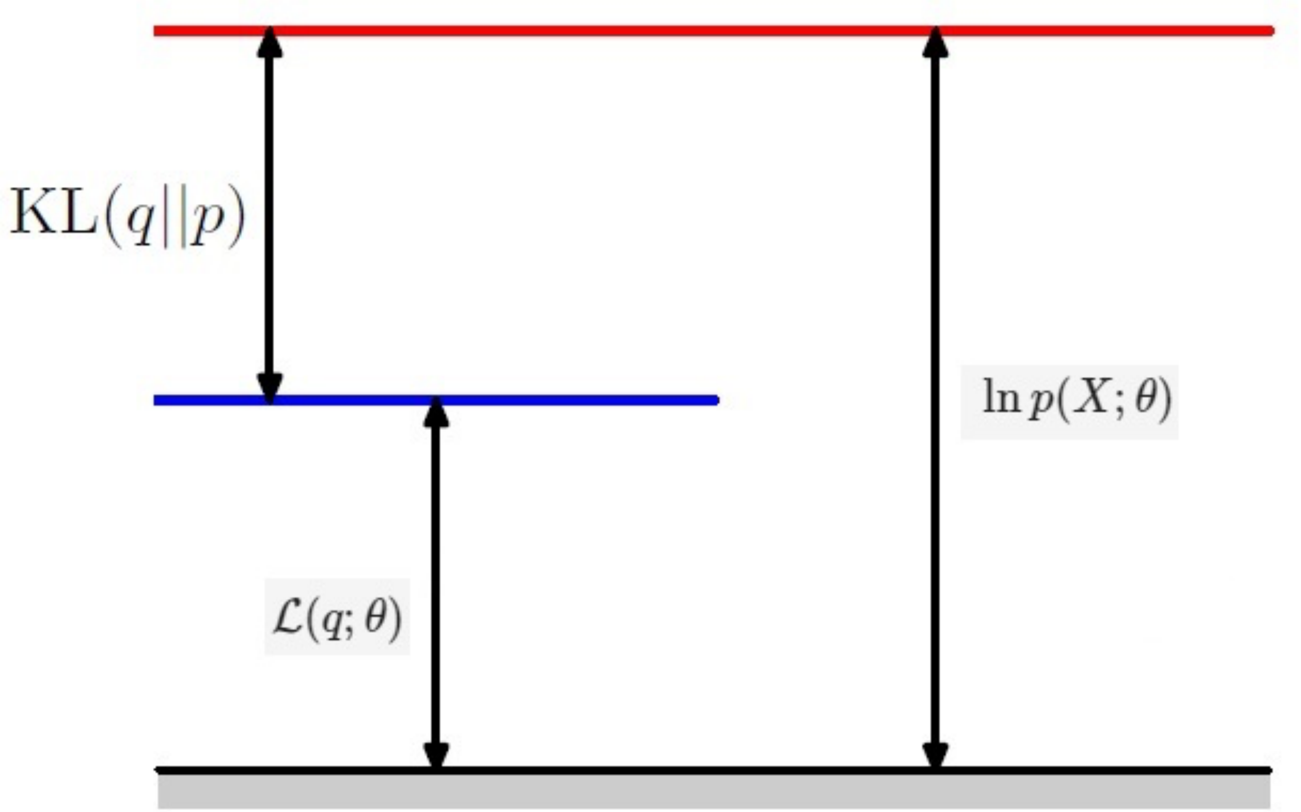
In particular, we should notice that $\mathcal{L}(q;\theta)$ and $KL(q||p)$ differ in sign.

Note:
In mathematics, and particularly in functional analysis and the calculus of variations, a functional is a function from a vector space into its underlying scalar field, or a set of functions of the real numbers. In other words, it is a function that takes a vector as its input argument, and returns a scalar. **Commonly the vector space is a space of functions, thus the functional takes a function for its input argument, then it is sometimes considered a function of a function (a higher-order function)**. Its use originates in the calculus of variations where one searches for a function that minimizes a certain functional.

We can get $KL(q||p) \geq 0$, with the equality if, and only if $q(Z) = p(Z|X;\theta)$. We can prove the conclusion based on the Jensen's inequality.

$$\begin{aligned}
KL(q||p) &= -\sum_Z q(Z) \times \ln \frac{p(Z|X;\theta)}{q(Z)} \\
&= -\mathbb{E}_{q(Z)}[\ln \frac{p(Z|X;\theta)}{q(Z)}] \\
&\geq -\ln \mathbb{E}_{q(Z)}[\frac{p(Z|X;\theta)}{q(Z)}] \\
&= -\ln \sum_Z q(Z) \times \frac{p(Z|X;\theta)}{q(Z)} \\
&= -\ln \sum_Z p(Z|X;\theta) = \ln(1) = 0
\end{aligned}$$

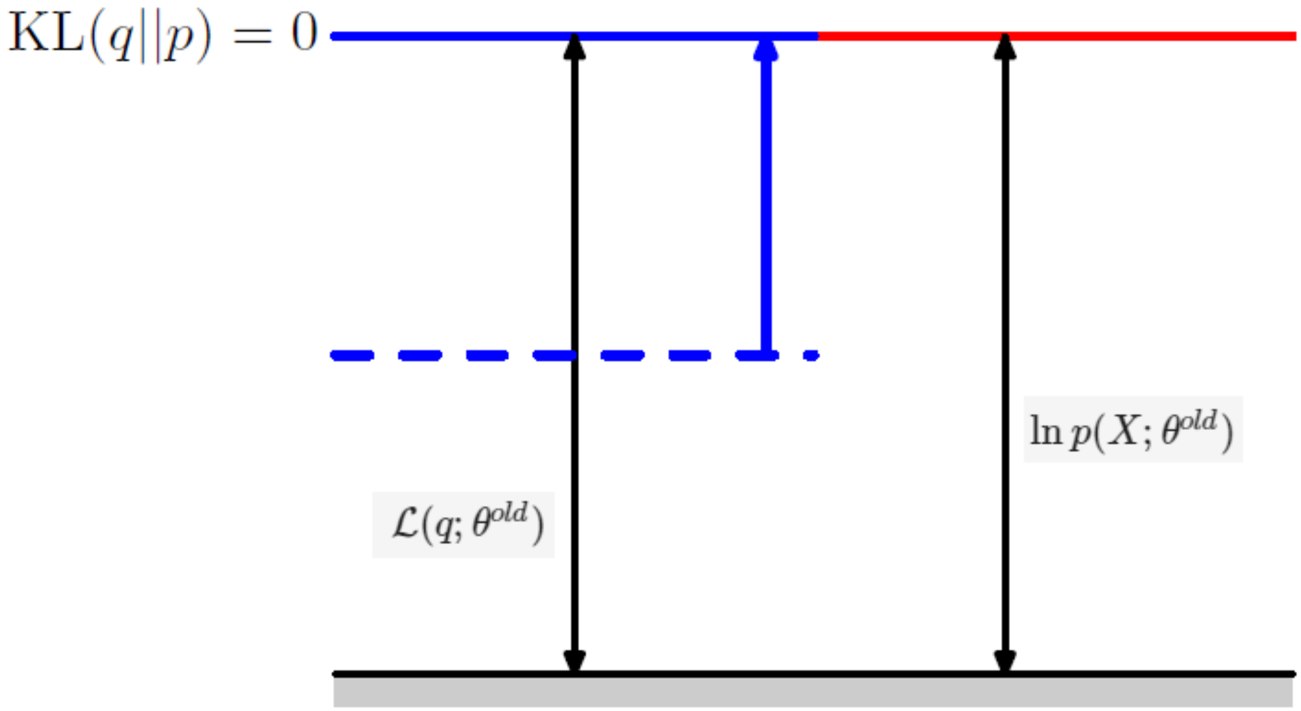
So we can know that $\mathcal{L}(q;\theta) \leq \ln p(X;\theta)$, in other words that $\mathcal{L}(q;\theta)$ is a lower bound on $\ln p(X;\theta)$ as shown in the following figure.



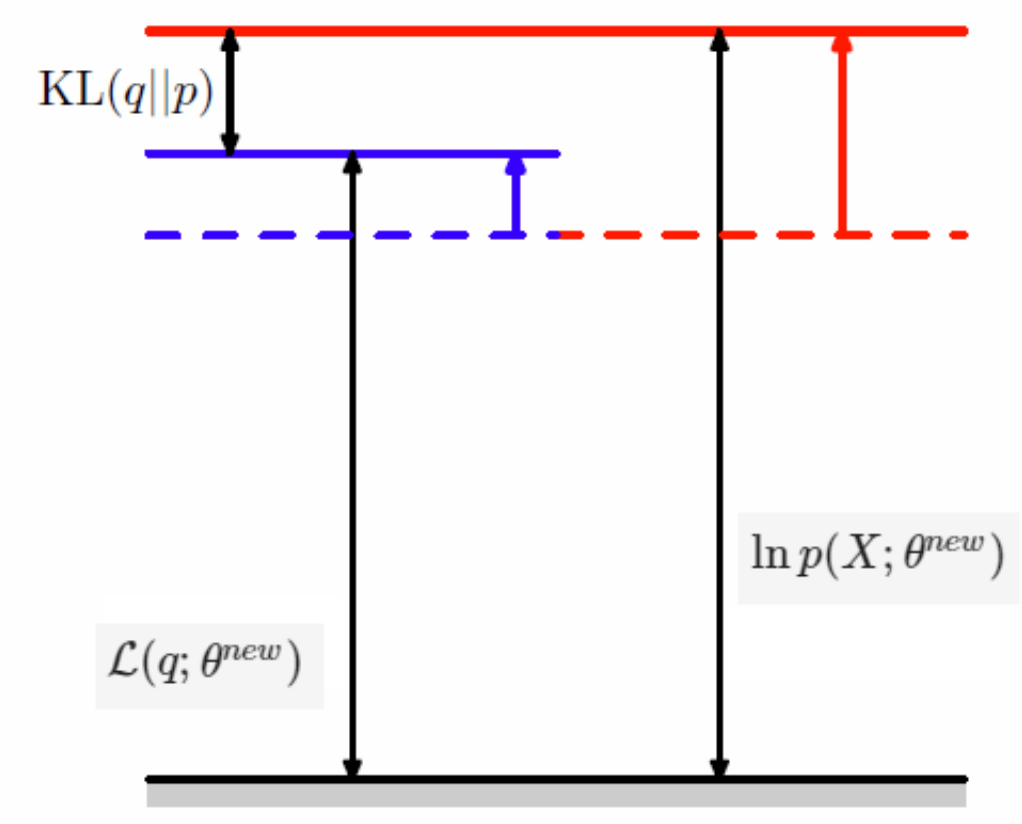
EM Steps

The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions. We can use the above decomposition 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 $\mathcal{L}(q;\theta^{old})$ 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 $\ln p(X;\theta^{old})$ does not depend on $q(Z)$ and so the largest value of $\mathcal{L}(q;\theta^{old})$ will occur when the Kullback-Leibler divergence vanishes, in other words when $q(Z)$ is equal to the posterior distribution $p(Z|X;\theta^{old})$. In this case, the lower bound will equal the log likelihood, as illustrated in the following figure.



In the subsequent M step, the distribution $q(Z)$ is held fixed and the lower bound $\mathcal{L}(q;\theta)$ is maximized with respect to θ to give some new value θ^{new}. This will cause the lower bound $\mathcal{L}$ 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 $p(Z|X;\theta^{new})$, 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 shown in the following figure.

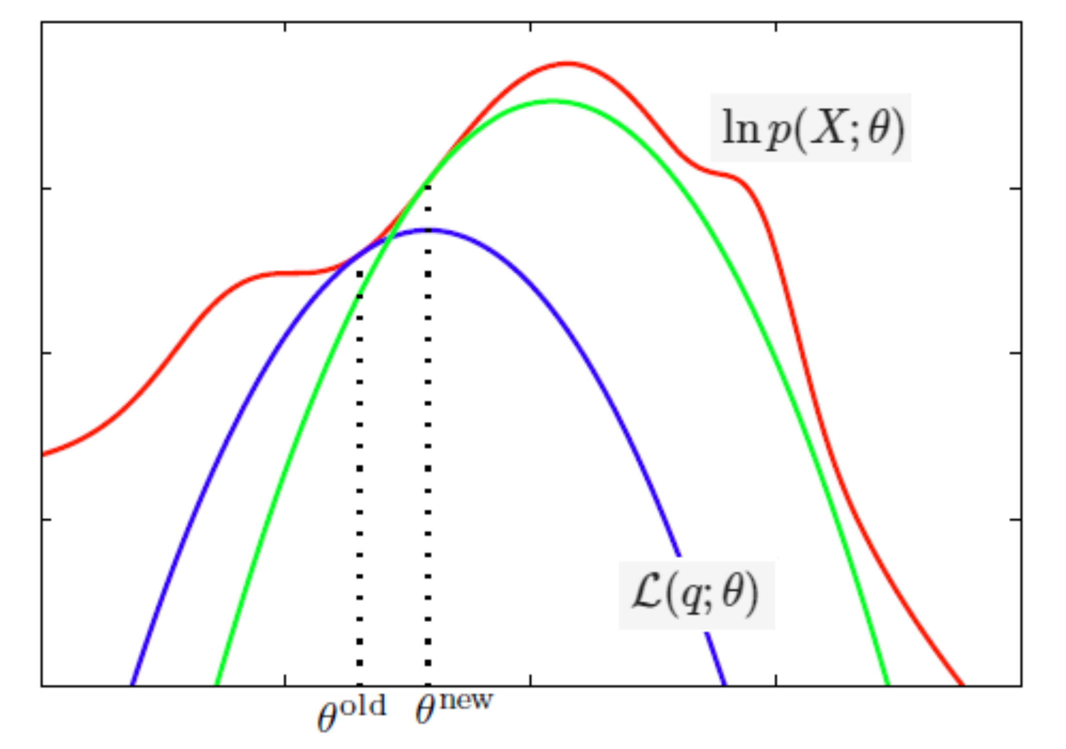


If we substitute $q(Z) = p(Z|X;\theta^{new})$, we see that, after the E step, the lower bound takes the form

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

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. Note that the variable θ over which we are optimizing appears only inside the logarithm. If the joint distribution $p(Z,X;\theta)$ comprises a member of the exponential family, or a product of such members, then we see that the logarithm will cancel the exponential and lead to an M step that will be typically much simpler than the maximization of the corresponding incomplete-data log-likelihood function $p(X;\theta)$.

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



Here the red curve depicts the (in-complete 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 Z, which gives rise to a lower bound $\mathcal{L}(q;\theta,\theta^{old})$
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.

For the particular case of an independent, identically distributed data set, X will comprise N data points $\{x_n\}$ while Z will comprise N corresponding latent variables

$\{z_n\}$, where $n = 1, \cdots, N$. From the independence assumption, we have $p(X, Z) = \prod_n p(x_n, z_n)$ and, by marginalizing over the $\{z_n\}$ we have $p(X) = \prod_n p(x_n)$. Using the sum and product rules, we see that the posterior probability that is evaluated in the E step takes the form

$$\begin{aligned} p(Z|X;\theta) &= \frac{p(X,Z;\theta)}{\sum_Z p(X,Z;\theta)} \\ &= \frac{\prod_{n=1}^N p(x_n, z_n;\theta)}{\sum_Z \prod_{n=1}^N p(x_n, z_n;\theta)} \\ &= \frac{\prod_{n=1}^N p(x_n, z_n;\theta)}{\prod_{n=1}^N p(x_n;\theta)} \\ &= \prod_{n=1}^N p(z_n|x_n;\theta) \end{aligned}$$

and so the posterior distribution also factorizes with respect to n. In the case of the Gaussian mixture model this simply says that the responsibility that each of the mixture components takes for a particular data point x_n depends only on the value of x_n and on the parameters θ of the mixture components, not on the values of the other data points.

We have seen that both the E and the M steps of the EM algorithm are increasing the value of a well-defined bound on the log likelihood function and that the complete EM cycle will change the model parameters in such a way as to cause the log likelihood to increase (unless it is already at a maximum, in which case the parameters remain unchanged).

We can also use the EM algorithm to maximize the posterior distribution $p(\theta|X)$ for models in which we have introduced a prior $p(\theta)$ over the parameters.

Note:
Here, θ is also a random variable instead of some fix value point. For example, we will use $p(X|\theta)$ instead of $p(X;\theta)$ used in the previous sections.

To see this, we note that as a function of θ, we have $p(\theta|X) = \frac{p(\theta,X)}{p(X)}$ and so

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

We have

$$\ln p(X|\theta) = \mathcal{L}(q;\theta) + KL(q||p)$$

So

$$\begin{aligned}
\ln p(\theta|X) &= \mathcal{L}(q;\theta) + KL(q||p) + \ln p(\theta) - \ln p(X) \\
&\geq \mathcal{L}(q;\theta) + \ln p(\theta) - \ln p(X)
\end{aligned}$$

where $\ln p(X)$ is a constant. We can again optimize the right-hand side alternately with respect to q and θ. The optimization with respect to q gives rise to the same E step equations as for the standard EM algorithm, because q only appears in $\mathcal{L}(q;\theta)$. The M-step equations are modified through the introduction of the prior term $\ln p(\theta)$, which typically requires only a small modification to the standard maximum likelihood M-step equations.

Reference

Christopher M. Bishop, **Pattern Recognition and Machine Learning**