

Bishop 9.4 Annotated

Montag, 26. Juni 2023 10:50

$$\alpha_i^{\text{new}} = \frac{1}{m_i^2 + \Sigma_{ii}} \quad (9.67)$$

$$(\beta^{\text{new}})^{-1} = \frac{\|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \beta^{-1} \sum_i \gamma_i}{N} \quad (9.68)$$

Exercise 9.23

These re-estimation equations are formally equivalent to those obtained by direct maximization.

9.4. The EM Algorithm in General

Section 10.1

The *expectation maximization* algorithm, or EM algorithm, is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997). Here we give a very general treatment of the EM algorithm and in the process provide a proof that the EM algorithm derived heuristically in Sections 9.2 and 9.3 for Gaussian mixtures does indeed maximize the likelihood function (Csiszár and Tusnády, 1984; Hathaway, 1986; Neal and Hinton, 1999). Our discussion will also form the basis for the derivation of the variational inference framework.

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

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta). \quad (9.69)$$

Here we are assuming \mathbf{Z} is discrete, although the discussion is identical if \mathbf{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 $p(\mathbf{X}|\theta)$ is difficult, but that optimization of the complete-data likelihood function $p(\mathbf{X}, \mathbf{Z}|\theta)$ is significantly easier. Next we introduce a distribution $q(\mathbf{Z})$ defined over the latent variables, and we observe that, for any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p) \quad (9.70)$$

where we have defined

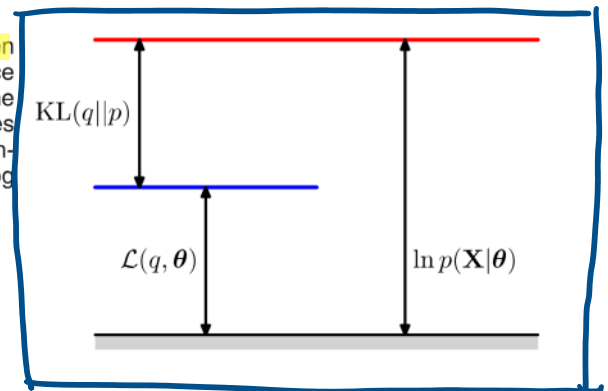
$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\} \quad (9.71)$$

$$\text{KL}(q||p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}. \quad (9.72)$$

Note that $\mathcal{L}(q, \theta)$ is a functional (see Appendix D for a discussion of functionals) of the distribution $q(\mathbf{Z})$, and a function of the parameters θ . It is worth studying

$$* \mathcal{L}(q, \theta) = \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})} \left[\ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right] = \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})} [\ln p(\mathbf{X}, \mathbf{Z}|\theta)] - \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})} [\ln q(\mathbf{Z})]$$

Figure 9.11 Illustration of the decomposition given by (9.70), which holds for any choice of distribution $q(\mathbf{Z})$. Because the Kullback-Leibler divergence satisfies $\text{KL}(q\|p) \geq 0$, we see that the quantity $\mathcal{L}(q, \theta)$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\theta)$.



carefully the forms of the expressions (9.71) and (9.72), and in particular noting that they differ in sign and also that $\mathcal{L}(q, \theta)$ contains the joint distribution of \mathbf{X} and \mathbf{Z} while $\text{KL}(q\|p)$ contains the conditional distribution of \mathbf{Z} given \mathbf{X} . To verify the decomposition (9.70), we first make use of the product rule of probability to give

$$\ln p(\mathbf{X}, \mathbf{Z}|\theta) = \ln p(\mathbf{Z}|\mathbf{X}, \theta) + \ln p(\mathbf{X}|\theta) \quad (9.73)$$

which we then substitute into the expression for $\mathcal{L}(q, \theta)$. This gives rise to two terms, one of which cancels $\text{KL}(q\|p)$ while the other gives the required log likelihood $\ln p(\mathbf{X}|\theta)$ after noting that $q(\mathbf{Z})$ is a normalized distribution that sums to 1.

From (9.72), we see that $\text{KL}(q\|p)$ is the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \theta)$. Recall that the Kullback-Leibler divergence satisfies $\text{KL}(q\|p) \geq 0$, with equality if, and only if, $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$. It therefore follows from (9.70) that $\mathcal{L}(q, \theta) \leq \ln p(\mathbf{X}|\theta)$, in other words that $\mathcal{L}(q, \theta)$ is a lower bound on $\ln p(\mathbf{X}|\theta)$. The decomposition (9.70) is illustrated in Figure 9.11.

The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions. We can use the decomposition (9.70) 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 $\mathcal{L}(q, \theta^{\text{old}})$ is maximized with respect to $q(\mathbf{Z})$ while holding θ^{old} fixed. The solution to this maximization problem is easily seen by noting that the value of $\ln p(\mathbf{X}|\theta^{\text{old}})$ does not depend on $q(\mathbf{Z})$ and so the largest value of $\mathcal{L}(q, \theta^{\text{old}})$ will occur when the Kullback-Leibler divergence vanishes, in other words when $q(\mathbf{Z})$ is equal to the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$. In this case, the lower bound will equal the log likelihood, as illustrated in Figure 9.12.

In the subsequent M step, the distribution $q(\mathbf{Z})$ 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 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{new}})$, 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

Ex 7, T2 i)

Section 1.6.1

iterative opt.
in EM algo:

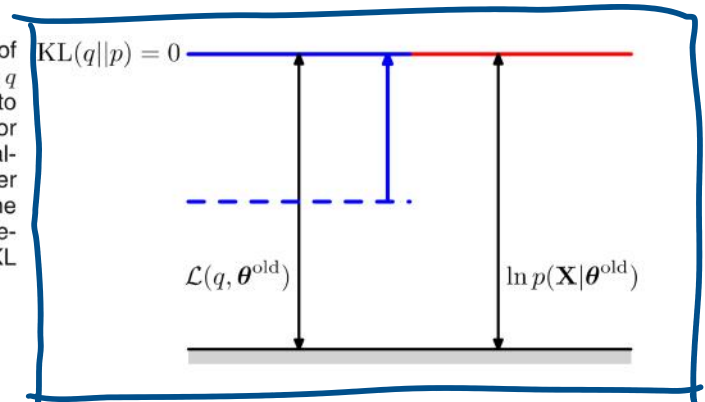
E-step:
 $\max_{q(\mathbf{Z})} \mathcal{L}(q, \theta^{\text{old}})$
wrt. $q(\mathbf{Z})$

M-step:
 $\max_{\theta} \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}), \theta)$
wrt. θ

→ why call it
"expectation"-step?
→ "calculates the
expected log-likelihood"
[NML, p. 332]

E-step

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 θ^{old} , causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.



shown in Figure 9.13. If we substitute $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ into (9.71), we see that, after the E step, the lower bound takes the form

$$\begin{aligned} \underset{\theta}{\text{argmax}} \quad \mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) - \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})}_{\text{const. w.r.t } \theta} \\ &= Q(\theta, \theta^{\text{old}}) + \text{const} \end{aligned} \quad (9.74)$$

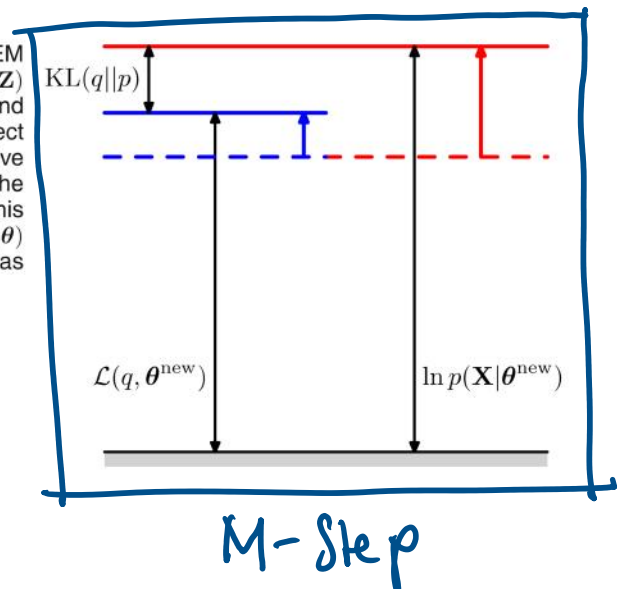
$Q \equiv \underset{\theta}{\text{argmax}}$

opt. of Q is
nice for exp.

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. Note that the variable θ over which we are optimizing appears only inside the logarithm. If the joint distribution $p(\mathbf{Z}, \mathbf{X}|\theta)$ 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 $p(\mathbf{X}|\theta)$.

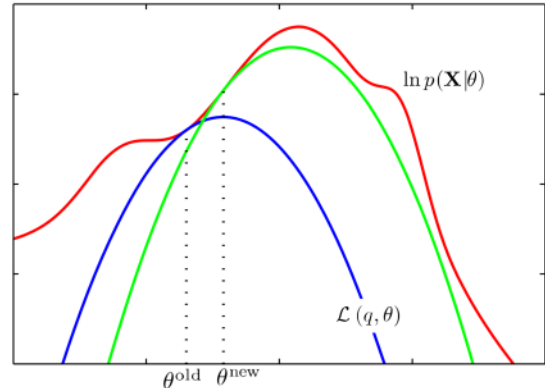
The operation of the EM algorithm can also be viewed in the space of parameters, as illustrated schematically in Figure 9.14. Here the red curve depicts the (in-

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



M-Step

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.



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, which gives rise to a lower bound $\mathcal{L}(\theta, \theta^{old})$ 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.

Exercise 9.25

Ex7, T2 iv)

posterior factorizes over datapoints

For the particular case of an independent, identically distributed data set, \mathbf{X} will comprise N data points $\{\mathbf{x}_n\}$ while \mathbf{Z} will comprise N corresponding latent variables $\{\mathbf{z}_n\}$, where $n = 1, \dots, N$. From the independence assumption, we have $p(\mathbf{X}, \mathbf{Z}) = \prod_n p(\mathbf{x}_n, \mathbf{z}_n)$ and, by marginalizing over the $\{\mathbf{z}_n\}$ we have $p(\mathbf{X}) = \prod_n p(\mathbf{x}_n)$. Using the sum and product rules, we see that the posterior probability that is evaluated in the E step takes the form

Ex7, T2 v)

$$p(\mathbf{Z}|\mathbf{X}, \theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)} = \frac{\prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_n|\theta)}{\sum_{\mathbf{Z}} \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_n|\theta)} = \prod_{n=1}^N p(\mathbf{z}_n|\mathbf{x}_n, \theta) \quad (9.75)$$

i.e.

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 \mathbf{x}_n depends only on the value of \mathbf{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