## 10.5. Local Variational Methods

The variational framework discussed in Sections 10.1 and 10.2 can be considered a 'global' method in the sense that it directly seeks an approximation to the full posterior distribution over all random variables. An alternative 'local' approach involves finding bounds on functions over individual variables or groups of variables within a model. For instance, we might seek a bound on a conditional distribution $p(y|x)$, which is itself just one factor in a much larger probabilistic model specified by a directed graph. The purpose of introducing the bound of course is to simplify the resulting distribution. This local approximation can be applied to multiple variables in turn until a tractable approximation is obtained, and in Section 10.6.1 we shall give a practical example of this approach in the context of logistic regression. Here we focus on developing the bounds themselves.

We have already seen in our discussion of the Kullback-Leibler divergence that the convexity of the logarithm function played a key role in developing the lower bound in the global variational approach. We have defined a (strictly) convex function as one for which every chord lies above the function. Convexity also plays a central role in the local variational framework. Note that our discussion will apply equally to concave functions with 'min' and 'max' interchanged and with lower bounds replaced by upper bounds.

Let us begin by considering a simple example, namely the function $f(x) = \exp(-x)$, which is a convex function of $x$, and which is shown in the left-hand plot of Figure 10.10. Our goal is to approximate $f(x)$ by a simpler function, in particular a linear function of $x$. From Figure 10.10, we see that this linear function will be a lower bound on $f(x)$ if it corresponds to a tangent. We can obtain the tangent line $y(x)$ at a specific value of $x$, say $x = \xi$, by making a first order Taylor expansion

$$y(x) = f(\xi) + f'(\xi)(x - \xi) \tag{10.125}$$

so that $y(x) \leqslant f(x)$ with equality when $x = \xi$. For our example function $f(x) =$

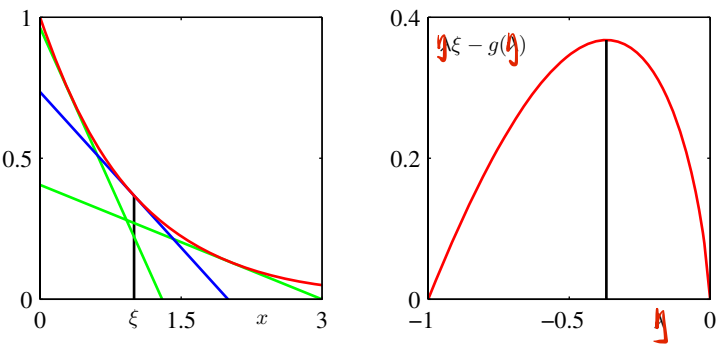



**Figure 10.10** In the left-hand figure the red curve shows the function $\exp(-x)$, and the blue line shows the tangent at $x = \xi$ defined by (10.125) with $\xi = 1$. This line has slope $\lambda = f'(\xi) = -\exp(-\xi)$. Note that any other tangent line, for example the ones shown in green, will have a smaller value of $y$ at $x = \xi$. The right-hand figure shows the corresponding plot of the function $\lambda\xi - g(\lambda)$, where $g(\lambda)$ is given by (10.131), versus $\lambda$ for $\xi = 1$, in which the maximum corresponds to $\lambda = -\exp(-\xi) = -1/e$.

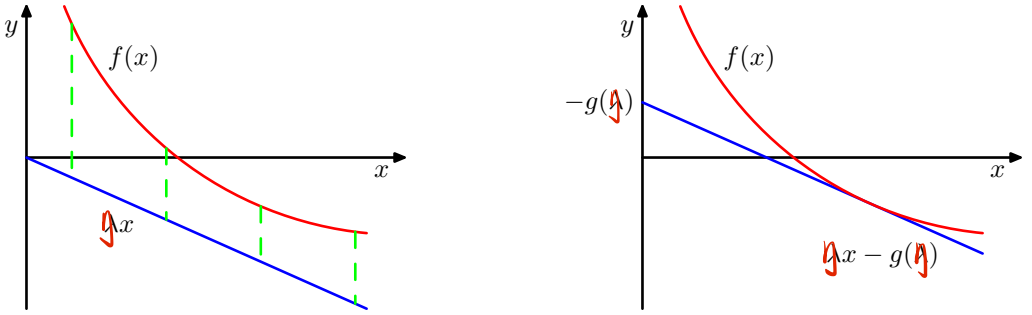



**Figure 10.11** In the left-hand plot the red curve shows a convex function $f(x)$, and the blue line represents the linear function $\lambda x$, which is a lower bound on $f(x)$ because $f(x) > \lambda x$ for all $x$. For the given value of slope $\lambda$ the contact point of the tangent line having the same slope is found by minimizing with respect to $x$ the discrepancy (shown by the green dashed lines) given by $f(x) - \lambda x$. This defines the dual function $g(\lambda)$, which corresponds to the (negative of the) intercept of the tangent line having slope $\lambda$.

$\exp(-x)$, we therefore obtain the tangent line in the form

$$y(x) = \exp(-\xi) - \exp(-\xi)(x - \xi) \tag{10.126}$$

which is a linear function parameterized by $\xi$. For consistency with subsequent discussion, let us define $\lambda = -\exp(-\xi)$ so that

$$y(x, \lambda) = \lambda x - \lambda + \lambda \ln(-\lambda). \tag{10.127}$$

Different values of $\lambda$ correspond to different tangent lines, and because all such lines are lower bounds on the function, we have $f(x) \geqslant y(x, \lambda)$. Thus we can write the function in the form

$$f(x) = \max_{\lambda} \{\lambda x - \lambda + \lambda \ln(-\lambda)\}. \tag{10.128}$$

We have succeeded in approximating the convex function $f(x)$ by a simpler, linear function $y(x, \lambda)$. The price we have paid is that we have introduced a variational parameter $\lambda$, and to obtain the tightest bound we must optimize with respect to $\lambda$.

We can formulate this approach more generally using the framework of *convex duality* (Rockafellar, 1972; Jordan *et al.*, 1999). Consider the illustration of a convex function $f(x)$ shown in the left-hand plot in Figure 10.11. In this example, the function $\lambda x$ is a lower bound on $f(x)$ but it is not the best lower bound that can be achieved by a linear function having slope $\lambda$, because the tightest bound is given by the tangent line. Let us write the equation of the tangent line, having slope $\lambda$ as $\lambda x - g(\lambda)$ where the (negative) intercept $g(\lambda)$ clearly depends on the slope $\lambda$ of the tangent. To determine the intercept, we note that the line must be moved vertically by an amount equal to the smallest vertical distance between the line and the function, as shown in Figure 10.11. Thus

$$\begin{aligned} g(\lambda) &= -\min_{x} \{f(x) - \lambda x\} \\ &= \max_{x} \{\lambda x - f(x)\}. \end{aligned} \tag{10.129}$$

Now, instead of fixing $\eta$ and varying $x$, we can consider a particular $x$ and then adjust $\eta$ until the tangent plane is tangent at that particular $x$. Because the $y$ value of the tangent line at a particular $x$ is maximized when that value coincides with its contact point, we have

$$f(x) = \max_{\eta} \{\eta x - g(\eta)\}. \tag{10.130}$$

We see that the functions $f(x)$ and $g(\eta)$ play a dual role, and are related through (10.129) and (10.130).

Let us apply these duality relations to our simple example $f(x) = \exp(-x)$. From (10.129) we see that the maximizing value of $x$ is given by $\xi = -\ln(-\eta)$, and back-substituting we obtain the conjugate function $g(\eta)$ in the form

$$g(\eta) = \eta - \eta \ln(-\eta) \tag{10.131}$$

as obtained previously. The function $\eta\xi - g(\eta)$ is shown, for $\xi = 1$ in the right-hand plot in Figure 10.10. As a check, we can substitute (10.131) into (10.130), which gives the maximizing value of $\eta = -\exp(-x)$, and back-substituting then recovers the original function $f(x) = \exp(-x)$.

For concave functions, we can follow a similar argument to obtain upper bounds, in which 'max' is replaced with 'min', so that

$$f(x) = \min_{\eta} \{\eta x - g(\eta)\} \tag{10.132}$$

$$g(\eta) = \min_{x} \{\eta x - f(x)\}. \tag{10.133}$$

If the function of interest is not convex (or concave), then we cannot directly apply the method above to obtain a bound. However, we can first seek invertible transformations either of the function or of its argument which change it into a convex form. We then calculate the conjugate function and then transform back to the original variables.

An important example, which arises frequently in pattern recognition, is the logistic sigmoid function defined by

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{10.134}$$

As it stands this function is neither convex nor concave. However, if we take the logarithm we obtain a function which is concave, as is easily verified by finding the second derivative. From (10.133) the corresponding conjugate function then takes the form

*Exercise 10.30*

$$g(\eta) = \min_{x} \{\eta x - f(x)\} = -\eta \ln \eta - (1 - \eta) \ln(1 - \eta) \tag{10.135}$$

which we recognize as the binary entropy function for a variable whose probability of having the value 1 is $\eta$. Using (10.132), we then obtain an upper bound on the log sigmoid

*Appendix B*

$$\ln \sigma(x) \leqslant \eta x - g(\eta) \tag{10.136}$$

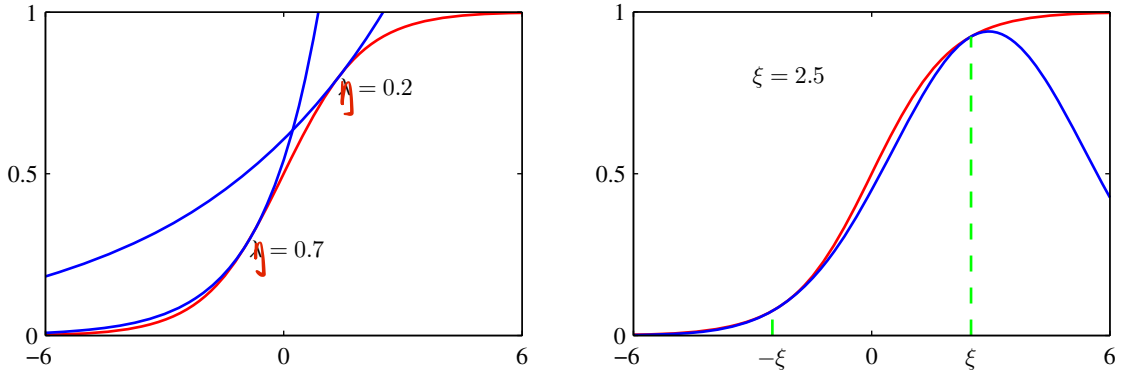



**Figure 10.12** The left-hand plot shows the logistic sigmoid function $\sigma(x)$ defined by (10.134) in red, together with two examples of the exponential upper bound (10.137) shown in blue. The right-hand plot shows the logistic sigmoid again in red together with the Gaussian lower bound (10.144) shown in blue. Here the parameter $\xi = 2.5$, and the bound is exact at $x = \xi$ and $x = -\xi$, denoted by the dashed green lines.

and taking the exponential, we obtain an upper bound on the logistic sigmoid itself of the form

$$\sigma(x) \leqslant \exp(\eta x - g(\eta)) \tag{10.137}$$

which is plotted for two values of $\eta$ on the left-hand plot in Figure 10.12.

上面构造了σ(x)的local上界，下面开始构造其local下界

We can also obtain a lower bound on the sigmoid having the functional form of a Gaussian. To do this, we follow Jaakkola and Jordan (2000) and make transformations both of the input variable and of the function itself. First we take the log of the logistic function and then decompose it so that

$$\begin{aligned} \ln \sigma(x) &= -\ln(1 + e^{-x}) = -\ln\left\{e^{-x/2}(e^{x/2} + e^{-x/2})\right\} \\ &= x/2 - \ln(e^{x/2} + e^{-x/2}). \end{aligned} \tag{10.138}$$

*Exercise 10.31*

We now note that the function $f(x) = -\ln(e^{x/2} + e^{-x/2})$ is a convex function of the variable $x^2$, as can again be verified by finding the second derivative. This leads to a lower bound on $f(x)$, which is a linear function of $x^2$ whose conjugate function is given by

$$g(\eta) = \max_{x^2} \left\{\eta x^2 - f\left(\sqrt{x^2}\right)\right\}. \tag{10.139}$$

The stationarity condition leads to

$$0 = \eta - \frac{dx}{dx^2}\frac{d}{dx} f(x) = \eta + \frac{1}{4x}\tanh\left(\frac{x}{2}\right). \tag{10.140}$$

If we denote this value of $x$, corresponding to the contact point of the tangent line for this particular value of $\eta$, by $\xi$, then we have

$$\eta = \lambda(\xi) = -\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right) = -\frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]. = -\lambda(\xi) \tag{10.141}$$

where we have defined $\lambda = -\eta$ to maintain consistency with Jaakkola and Jordan (2000).

Instead of thinking of $\lambda$ as the variational parameter, we can let $\xi$ play this role as this leads to simpler expressions for the conjugate function, which is then given by

$$g(\lambda) = -\lambda(\xi)\xi^2 - f(\xi) = -\lambda(\xi)\xi^2 + \ln(e^{\xi/2} + e^{-\xi/2}). \tag{10.142}$$

Hence the bound on $f(x)$ can be written as

$$f(x) \geqslant -\lambda x^2 - g(\lambda) = -\lambda x^2 + \lambda\xi^2 - \ln(e^{\xi/2} + e^{-\xi/2}). \tag{10.143}$$

The bound on the sigmoid then becomes

$$\sigma(x) \geqslant \sigma(\xi) \exp\left\{(x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2)\right\} \tag{10.144}$$

where $\lambda(\xi)$ is defined by (10.141). This bound is illustrated in the right-hand plot of Figure 10.12. We see that the bound has the form of the exponential of a quadratic function of $x$, which will prove useful when we seek Gaussian representations of posterior distributions defined through logistic sigmoid functions.

*Section 4.5*

上述推导并不能直接推广到softmax函数中

*Section 4.3*

The logistic sigmoid arises frequently in probabilistic models over binary variables because it is the function that transforms a log odds ratio into a posterior probability. The corresponding transformation for a multiclass distribution is given by the softmax function. Unfortunately, the lower bound derived here for the logistic sigmoid does not directly extend to the softmax. Gibbs (1997) proposes a method for constructing a Gaussian distribution that is conjectured to be a bound (although no rigorous proof is given), which may be used to apply local variational methods to multiclass problems.

下凸函数变分下界的应用举例

We shall see an example of the use of local variational bounds in Sections 10.6.1. For the moment, however, it is instructive to consider in general terms how these bounds can be used. Suppose we wish to evaluate an integral of the form

$$I = \int \sigma(a)p(a)\,\mathrm{d}a \tag{10.145}$$

where $\sigma(a)$ is the logistic sigmoid, and $p(a)$ is a Gaussian probability density. Such integrals arise in Bayesian models when, for instance, we wish to evaluate the predictive distribution, in which case $p(a)$ represents a posterior parameter distribution. Because the integral is intractable, we employ the variational bound (10.144), which we write in the form $\sigma(a) \geqslant f(a, \xi)$ where $\xi$ is a variational parameter. The integral now becomes the product of two exponential-quadratic functions and so can be integrated analytically to give a bound on $I$

$$I \geqslant \int f(a, \xi)p(a)\,\mathrm{d}a = F(\xi). \tag{10.146}$$

We now have the freedom to choose the variational parameter $\xi$, which we do by finding the value $\xi^\star$ that maximizes the function $F(\xi)$. The resulting value $F(\xi^\star)$ represents the tightest bound within this family of bounds and can be used as an approximation to $I$. This optimized bound, however, will in general not be exact.

Although the bound $\sigma(a) \geqslant f(a, \xi)$ on the logistic sigmoid can be optimized exactly, the required choice for $\xi$ depends on the value of $a$, so that the bound is exact for one value of $a$ only. Because the quantity $F(\xi)$ is obtained by integrating over all values of $a$, the value of $\xi^\star$ represents a compromise, weighted by the distribution $p(a)$.

## 10.6. Variational Logistic Regression

We now illustrate the use of local variational methods by returning to the Bayesian logistic regression model studied in Section 4.5. There we focussed on the use of the Laplace approximation, while here we consider a variational treatment based on the approach of Jaakkola and Jordan (2000). Like the Laplace method, this also leads to a Gaussian approximation to the posterior distribution. However, the greater flexibility of the variational approximation leads to improved accuracy compared to the Laplace method. Furthermore (unlike the Laplace method), the variational approach is optimizing a well defined objective function given by a rigourous bound on the model evidence. Logistic regression has also been treated by Dybowski and Roberts (2005) from a Bayesian perspective using Monte Carlo sampling techniques.

### 10.6.1 Variational posterior distribution

Here we shall make use of a variational approximation based on the local bounds introduced in Section 10.5. This allows the likelihood function for logistic regression, which is governed by the logistic sigmoid, to be approximated by the exponential of a quadratic form. It is therefore again convenient to choose a conjugate Gaussian prior of the form (4.140). For the moment, we shall treat the hyperparameters $\mathbf{m}_0$ and $\mathbf{S}_0$ as fixed constants. In Section 10.6.3, we shall demonstrate how the variational formalism can be extended to the case where there are unknown hyperparameters whose values are to be inferred from the data.

In the variational framework, we seek to maximize a lower bound on the marginal likelihood. For the Bayesian logistic regression model, the marginal likelihood takes the form

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w} = \int \left[\prod_{n=1}^{N} p(t_n|\mathbf{w})\right] p(\mathbf{w})\,\mathrm{d}\mathbf{w}. \tag{10.147}$$

We first note that the conditional distribution for $t$ can be written as

$$\begin{aligned} p(t|\mathbf{w}) &= \sigma(a)^t \left\{1-\sigma(a)\right\}^{1-t} \\ &= \left(\frac{1}{1+e^{-a}}\right)^t \left(1-\frac{1}{1+e^{-a}}\right)^{1-t} \\ &= e^{at}\frac{e^{-a}}{1+e^{-a}} = e^{at}\sigma(-a) \end{aligned} \tag{10.148}$$

where $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$. In order to obtain a lower bound on $p(\mathbf{t})$, we make use of the variational lower bound on the logistic sigmoid function given by (10.144), which

we reproduce here for convenience

$$\sigma(z) \geqslant \sigma(\xi) \exp\left\{(z-\xi)/2 - \lambda(\xi)(z^2-\xi^2)\right\} \tag{10.149}$$

where

$$\lambda(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]. \tag{10.150}$$

We can therefore write

$$p(t|\mathbf{w}) = e^{at}\sigma(-a) \geqslant e^{at}\sigma(\xi)\exp\left\{-(a+\xi)/2 - \lambda(\xi)(a^2-\xi^2)\right\}. \tag{10.151}$$

Note that because this bound is applied to each of the terms in the likelihood function separately, there is a variational parameter $\xi_n$ corresponding to each training set observation $(\boldsymbol{\phi}_n, t_n)$. Using $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$, and multiplying by the prior distribution, we obtain the following bound on the joint distribution of $\mathbf{t}$ and $\mathbf{w}$

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \geqslant h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w}) \tag{10.152}$$

where $\boldsymbol{\xi}$ denotes the set $\{\xi_n\}$ of variational parameters, and

$$\begin{aligned} h(\mathbf{w}, \boldsymbol{\xi}) &= \prod_{n=1}^{N} \sigma(\xi_n) \exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n - (\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \xi_n)/2 \right. \\ &\quad \left. - \lambda(\xi_n)([\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n]^2 - \xi_n^2)\right\}. \end{aligned} \tag{10.153}$$

Evaluation of the exact posterior distribution would require normalization of the left-hand side of this inequality. Because this is intractable, we work instead with the right-hand side. Note that the function on the right-hand side cannot be interpreted as a probability density because it is not normalized. Once it is normalized to give a variational posterior distribution $q(\mathbf{w})$, however, it no longer represents a bound.

Because the logarithm function is monotonically increasing, the inequality $A \geqslant B$ implies $\ln A \geqslant \ln B$. This gives a lower bound on the log of the joint distribution of $\mathbf{t}$ and $\mathbf{w}$ of the form

$$\begin{aligned} \ln\{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\} &\geqslant \ln p(\mathbf{w}) + \sum_{n=1}^{N}\left\{\ln\sigma(\xi_n) + \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n \right. \\ &\quad \left. - (\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \xi_n)/2 - \lambda(\xi_n)([\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n]^2 - \xi_n^2)\right\}. \end{aligned} \tag{10.154}$$

Substituting for the prior $p(\mathbf{w})$, the right-hand side of this inequality becomes, as a function of $\mathbf{w}$

$$\begin{aligned} &-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ &+\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n(t_n - 1/2) - \lambda(\xi_n)\mathbf{w}^{\mathrm{T}}(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}})\mathbf{w}\right\} + \text{const}. \end{aligned} \tag{10.155}$$

注：式(10.152)右边是下界，但是一个概率分布，而一但归一化为概率分布，则其不再为下界。

最终推得参数w的变分近似后验q(w)注意，这里是近似而不是下界，只有这一近似是由下界推导而来的

This is a quadratic function of $\mathbf{w}$, and so we can obtain the corresponding variational approximation to the posterior distribution by identifying the linear and quadratic terms in $\mathbf{w}$, giving a Gaussian variational posterior of the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{10.156}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N}(t_n - 1/2)\boldsymbol{\phi}_n\right) \tag{10.157}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}. \tag{10.158}$$

q(w)带有变分参数{ξn}因此相比Laplace approximation能提供更灵活更精确的近似

As with the Laplace framework, we have again obtained a Gaussian approximation to the posterior distribution. However, the additional flexibility provided by the variational parameters $\{\xi_n\}$ leads to improved accuracy in the approximation (Jaakkola and Jordan, 2000).

Here we have considered a batch learning context in which all of the training data is available at once. However, Bayesian methods are intrinsically well suited to sequential learning in which the data points are processed one at a time and then discarded. The formulation of this variational approach for the sequential case is straightforward. 扩展到sequential learning（结合10.6.2{ξn}的学习）

*Exercise 10.32*

多类问题情况

Note that the bound given by (10.149) applies only to the two-class problem and so this approach does not directly generalize to classification problems with $K > 2$ classes. An alternative bound for the multiclass case has been explored by Gibbs (1997).

### 10.6.2 Optimizing the variational parameters

q(w)实际上condition on {ξn}, m0, S0, 它们是超参，本小节在给定m0, S0的条件下优化变分超参{ξn}

We now have a normalized Gaussian approximation to the posterior distribution, which we shall use shortly to evaluate the predictive distribution for new data points. First, however, we need to determine the variational parameters $\{\xi_n\}$ by maximizing the lower bound on the marginal likelihood.

To do this, we substitute the inequality (10.152) back into the marginal likelihood to give

$$\ln p(\mathbf{t}) = \ln \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w} \geqslant \ln \int h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w})\,\mathrm{d}\mathbf{w} = \mathcal{L}(\boldsymbol{\xi}). \tag{10.159}$$

As with the optimization of the hyperparameter $\alpha$ in the linear regression model of Section 3.5, there are two approaches to determining the $\xi_n$. ① In the first approach, we recognize that the function $\mathcal{L}(\boldsymbol{\xi})$ is defined by an integration over $\mathbf{w}$ and so we can view $\mathbf{w}$ as a latent variable and invoke the EM algorithm. ② In the second approach, we integrate over $\mathbf{w}$ analytically and then perform a direct maximization over $\boldsymbol{\xi}$. Let us begin by considering the EM approach.

① The EM algorithm starts by choosing some initial values for the parameters $\{\xi_n\}$, which we denote collectively by $\boldsymbol{\xi}^{\mathrm{old}}$. In the E step of the EM algorithm,

这里的EM算法并不严谨：h(w,ξ)虽然是p(t|w)的近似，但却并不是一个概率分布，因此严格来讲这里并不符合EM算法的使用场景，这里只是硬套EM的计算过程罢了。

we then use these parameter values to find the posterior distribution over $\mathbf{w}$, which is given by (10.156). In the M step, we then maximize the expected complete-data log likelihood which is given by

$$Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{\text{old}}) = \mathbb{E}\left[\ln h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w})\right] \tag{10.160}$$

where the expectation is taken with respect to the posterior distribution $q(\mathbf{w})$ evaluated using $\boldsymbol{\xi}^{\text{old}}$. Noting that $p(\mathbf{w})$ does not depend on $\boldsymbol{\xi}$, and substituting for $h(\mathbf{w}, \boldsymbol{\xi})$ we obtain

$$Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{\text{old}}) = \sum_{n=1}^{N} \left\{ \ln \sigma(\xi_n) - \xi_n/2 - \lambda(\xi_n)(\boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] \boldsymbol{\phi}_n - \xi_n^2) \right\} + \text{const} \tag{10.161}$$

where 'const' denotes terms that are independent of $\boldsymbol{\xi}$. We now set the derivative with respect to $\xi_n$ equal to zero. A few lines of algebra, making use of the definitions of $\sigma(\xi)$ and $\lambda(\xi)$, then gives

$$0 = \lambda'(\xi_n)(\boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] \boldsymbol{\phi}_n - \xi_n^2). \tag{10.162}$$

We now note that $\lambda'(\xi)$ is a monotonic function of $\xi$ for $\xi \geqslant 0$, and that we can restrict attention to nonnegative values of $\xi$ without loss of generality due to the symmetry of the bound around $\xi = 0$. Thus $\lambda'(\xi) \neq 0$, and hence we obtain the following re-estimation equations

*Exercise 10.33*

$$(\xi_n^{\text{new}})^2 = \boldsymbol{\phi}_n^{\mathrm{T}} \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] \boldsymbol{\phi}_n = \boldsymbol{\phi}_n^{\mathrm{T}} \left( \mathbf{S}_N + \mathbf{m}_N \mathbf{m}_N^{\mathrm{T}} \right) \boldsymbol{\phi}_n \tag{10.163}$$

where we have used (10.156).

EM算法计算过程的总结：Let us summarize the EM algorithm for finding the variational posterior distribution. We first initialize the variational parameters $\boldsymbol{\xi}^{\text{old}}$. In the E step, we evaluate the posterior distribution over $\mathbf{w}$ given by (10.156), in which the mean and covariance are defined by (10.157) and (10.158). In the M step, we then use this variational posterior to compute a new value for $\boldsymbol{\xi}$ given by (10.163). The E and M steps are repeated until a suitable convergence criterion is satisfied, which in practice typically requires only a few iterations.

② 解析计算下界再求导优化，可得到与EM算法相同之迭代优化式

*Exercise 10.34*

An alternative approach to obtaining re-estimation equations for $\boldsymbol{\xi}$ is to note that in the integral over $\mathbf{w}$ in the definition (10.159) of the lower bound $\mathcal{L}(\boldsymbol{\xi})$, the integrand has a Gaussian-like form and so the integral can be evaluated analytically. Having evaluated the integral, we can then differentiate with respect to $\xi_n$. It turns out that this gives rise to exactly the same re-estimation equations as does the EM approach given by (10.163).

如何解析计算 $\mathcal{L}(\xi)$

*Exercise 10.35*

As we have emphasized already, in the application of variational methods it is useful to be able to evaluate the lower bound $\mathcal{L}(\boldsymbol{\xi})$ given by (10.159). The integration over $\mathbf{w}$ can be performed analytically by noting that $p(\mathbf{w})$ is Gaussian and $h(\mathbf{w}, \boldsymbol{\xi})$ is the exponential of a quadratic function of $\mathbf{w}$. Thus, by completing the square and making use of the standard result for the normalization coefficient of a Gaussian distribution, we can obtain a closed form solution which takes the form

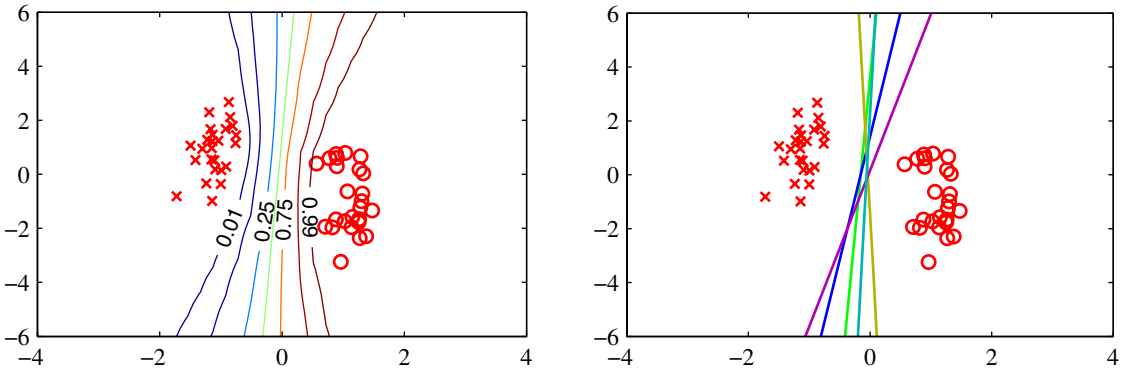



**Figure 10.13** Illustration of the Bayesian approach to logistic regression for a simple linearly separable data set. The plot on the left shows the predictive distribution obtained using variational inference. We see that the decision boundary lies roughly mid way between the clusters of data points, and that the contours of the predictive distribution splay out away from the data reflecting the greater uncertainty in the classification of such regions. The plot on the right shows the decision boundaries corresponding to five samples of the parameter vector $\mathbf{w}$ drawn from the posterior distribution $p(\mathbf{w}|\mathbf{t})$.

$$\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{2}\ln\frac{|\mathbf{S}_N|}{|\mathbf{S}_0|} + \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N}\left\{\ln\sigma(\xi_n) - \frac{1}{2}\xi_n + \lambda(\xi_n)\xi_n^2\right\}. \tag{10.164}$$

如何进行 sequential learning, 实际上就是 Exercise 10.32

This variational framework can also be applied to situations in which the data is arriving sequentially (Jaakkola and Jordan, 2000). In this case we maintain a Gaussian posterior distribution over $\mathbf{w}$, which is initialized using the prior $p(\mathbf{w})$. As each data point arrives, the posterior is updated by making use of the bound (10.151) and then normalized to give an updated posterior distribution.

预测分布

The predictive distribution is obtained by marginalizing over the posterior distribution, and takes the same form as for the Laplace approximation discussed in Section 4.5.2. Figure 10.13 shows the variational predictive distributions for a synthetic data set. This example provides interesting insights into the concept of 'large margin', which was discussed in Section 7.1 and which has qualitatively similar behaviour to the Bayesian solution.

### 10.6.3 Inference of hyperparameters

So far, we have treated the hyperparameter $\alpha$ in the prior distribution as a known constant. We now extend the Bayesian logistic regression model to allow the value of this parameter to be inferred from the data set. This can be achieved by combining the global and local variational approximations into a single framework, so as to maintain a lower bound on the marginal likelihood at each stage. Such a combined approach was adopted by Bishop and Svensén (2003) in the context of a Bayesian treatment of the hierarchical mixture of experts model.

注意，10.6.1 和 10.6.2 是一脉相承的：$\mathbf{m}_0, \mathbf{S}_0$ 是超参，直接给定，$\{\xi_n\}$ 是变分超参，在 10.6.2 小节中进行了优化，参数化后验 $q(\mathbf{w})$ 实际上 condition on $\mathbf{m}_0, \mathbf{S}_0, \{\xi_n\}$；在 10.6.1 小节给出。而 10.6.3 小节则自成体系：前文中视 $\mathbf{m}_0, \mathbf{S}_0$ 时应该是参数 $\alpha$，并引入先验分布 $p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0)$，参数化后验此时为 $q(\mathbf{w}, \alpha)$，多了个参数 $\alpha$；超参数 $a_0, b_0$ 直接给定，剩下的变分超参 $\{\xi_n\}$ 则依然通过优化得到。

Specifically, we consider once again a simple isotropic Gaussian prior distribution of the form

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}). \tag{10.165}$$

Our analysis is readily extended to more general Gaussian priors, for instance if we wish to associate a different hyperparameter with different subsets of the parameters $w_j$. As usual, we consider a conjugate hyperprior over $\alpha$ given by a gamma distribution

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0) \tag{10.166}$$

governed by the constants $a_0$ and $b_0$.

The marginal likelihood for this model now takes the form

$$p(\mathbf{t}) = \iint p(\mathbf{w}, \alpha, \mathbf{t})\, \mathrm{d}\mathbf{w}\, \mathrm{d}\alpha \tag{10.167}$$

where the joint distribution is given by

$$p(\mathbf{w}, \alpha, \mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha). \tag{10.168}$$

We are now faced with an analytically intractable integration over $\mathbf{w}$ and $\alpha$, which we shall tackle by using both the local and global variational approaches in the same model.

上面给出了背景条件，下面开始推导

To begin with, we introduce a variational distribution $q(\mathbf{w}, \alpha)$, and then apply the decomposition (10.2), which in this instance takes the form

$$\ln p(\mathbf{t}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{10.169}$$

where the lower bound $\mathcal{L}(q)$ and the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$ are defined by

$$\mathcal{L}(q) = \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{p(\mathbf{w}, \alpha, \mathbf{t})}{q(\mathbf{w}, \alpha)} \right\} \mathrm{d}\mathbf{w}\, \mathrm{d}\alpha \tag{10.170}$$

$$\mathrm{KL}(q\|p) = -\iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{p(\mathbf{w}, \alpha|\mathbf{t}))}{q(\mathbf{w}, \alpha)} \right\} \mathrm{d}\mathbf{w}\, \mathrm{d}\alpha. \tag{10.171}$$

At this point, the lower bound $\mathcal{L}(q)$ is still intractable due to the form of the likelihood factor $p(\mathbf{t}|\mathbf{w})$. We therefore apply the local variational bound to each of the logistic sigmoid factors as before. This allows us to use the inequality (10.152) and place a lower bound on $\mathcal{L}(q)$, which will therefore also be a lower bound on the log marginal likelihood

第①个≥由global variational approximation引入，即变分推断(VI)；第②个≥则由local variational approximation引入。

$$\ln p(\mathbf{t}) \overset{①}{\geqslant} \mathcal{L}(q) \overset{②}{\geqslant} \widetilde{\mathcal{L}}(q, \boldsymbol{\xi}) = \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w}|\alpha)p(\alpha)}{q(\mathbf{w}, \alpha)} \right\} \mathrm{d}\mathbf{w}\, \mathrm{d}\alpha. \tag{10.172}$$

Next we assume that the variational distribution factorizes between parameters and hyperparameters so that

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha). \tag{10.173}$$

With this factorization we can appeal to the general result (10.9) to find expressions for the optimal factors. Consider first the distribution $q(\mathbf{w})$. Discarding terms that are independent of $\mathbf{w}$, we have

$$\begin{aligned}\ln q(\mathbf{w}) &= \mathbb{E}_\alpha \left[\ln \{h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w}|\alpha) p(\alpha)\}\right] + \text{const} \\ &= \ln h(\mathbf{w}, \boldsymbol{\xi}) + \mathbb{E}_\alpha \left[\ln p(\mathbf{w}|\alpha)\right] + \text{const}.\end{aligned}$$

We now substitute for $\ln h(\mathbf{w}, \boldsymbol{\xi})$ using (10.153), and for $\ln p(\mathbf{w}|\alpha)$ using (10.165), giving

$$\ln q(\mathbf{w}) = -\frac{\mathbb{E}[\alpha]}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N}\left\{(t_n - 1/2)\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n - \lambda(\xi_n)\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{w}\right\} + \text{const}.$$

We see that this is a quadratic function of $\mathbf{w}$ and so the solution for $q(\mathbf{w})$ will be Gaussian. Completing the square in the usual way, we obtain

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \tag{10.174}$$

where we have defined

$$\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N = \sum_{n=1}^{N}(t_n - 1/2)\boldsymbol{\phi}_n \tag{10.175}$$

$$\boldsymbol{\Sigma}_N^{-1} = \mathbb{E}[\alpha]\mathbf{I} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}. \tag{10.176}$$

Similarly, the optimal solution for the factor $q(\alpha)$ is obtained from

$$\ln q(\alpha) = \mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{w}|\alpha)\right] + \ln p(\alpha) + \text{const}.$$

Substituting for $\ln p(\mathbf{w}|\alpha)$ using (10.165), and for $\ln p(\alpha)$ using (10.166), we obtain

$$\ln q(\alpha) = \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right] + (a_0 - 1)\ln\alpha - b_0\alpha + \text{const}.$$

We recognize this as the log of a gamma distribution, and so we obtain

$$q(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N) = \frac{1}{\Gamma(a_N)}b_N^{a_N}\alpha^{a_N - 1}e^{-b_N\alpha} \tag{10.177}$$

where

$$a_N = a_0 + \frac{M}{2} \tag{10.178}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]. \tag{10.179}$$

local variational 框架下变分参数 $\xi_n$ 的优化，又 local variational 是嵌入在 VI (global variational) 中的，因此这里得到的也是迭代公式。

We also need to optimize the variational parameters $\xi_n$, and this is also done by maximizing the lower bound $\widetilde{\mathcal{L}}(q, \boldsymbol{\xi})$. Omitting terms that are independent of $\boldsymbol{\xi}$, and integrating over $\alpha$, we have

$$\widetilde{\mathcal{L}}(q, \boldsymbol{\xi}) = \int q(\mathbf{w}) \ln h(\mathbf{w}, \boldsymbol{\xi}) \, \mathrm{d}\mathbf{w} + \text{const}. \tag{10.180}$$

Note that this has precisely the same form as (~~10.159~~ 10.160), and so we can again appeal to our earlier result (10.163), which can be obtained by direct optimization of the marginal likelihood function, leading to re-estimation equations of the form

$$(\xi_n^{\text{new}})^2 = \boldsymbol{\phi}_n^{\mathrm{T}} \left( \boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N \boldsymbol{\mu}_N^{\mathrm{T}} \right) \boldsymbol{\phi}_n. \tag{10.181}$$

总结前文，给出会用到的一些矩的计算公式 *Appendix B*

We have obtained re-estimation equations for the three quantities $q(\mathbf{w})$, $q(\alpha)$, and $\boldsymbol{\xi}$, and so after making suitable initializations, we can cycle through these quantities, updating each in turn. The required moments are given by

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N} \tag{10.182}$$

$$\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right] = \boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N^{\mathrm{T}} \boldsymbol{\mu}_N. \tag{10.183}$$

## 10.7. Expectation Propagation

KL(q||p) → KL(p||q)
VI/VB EP

We conclude this chapter by discussing an alternative form of deterministic approximate inference, known as *expectation propagation* or *EP* (Minka, 2001a; Minka, 2001b). As with the variational Bayes methods discussed so far, this too is based on the minimization of a Kullback-Leibler divergence but now of the reverse form, which gives the approximation rather different properties.

注意，EP算法并不要求 $q$ 为指数族分布，只不过当限制 $q$ 为指数族分布时，最小化式(10.198)就可以基于这里推得的 moment matching 来解。

Consider for a moment the problem of minimizing $\mathrm{KL}(p\|q)$ with respect to $q(\mathbf{z})$ when $p(\mathbf{z})$ is a fixed distribution and $q(\mathbf{z})$ is a member of the exponential family and so, from (2.194), can be written in the form

$$q(\mathbf{z}) = h(\mathbf{z}) g(\boldsymbol{\eta}) \exp\left\{ \boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{z}) \right\}. \tag{10.184}$$

As a function of $\boldsymbol{\eta}$, the Kullback-Leibler divergence then becomes

$$\mathrm{KL}(p\|q) = -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^{\mathrm{T}} \mathbb{E}_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})] + \text{const} \tag{10.185}$$

where the constant terms are independent of the natural parameters $\boldsymbol{\eta}$. We can minimize $\mathrm{KL}(p\|q)$ within this family of distributions by setting the gradient with respect to $\boldsymbol{\eta}$ to zero, giving

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})]. \tag{10.186}$$

However, we have already seen in (2.226) that the negative gradient of $\ln g(\boldsymbol{\eta})$ is given by the expectation of $\mathbf{u}(\mathbf{z})$ under the distribution $q(\mathbf{z})$. Equating these two results, we obtain

$$\mathbb{E}_{q(\mathbf{z})}[\mathbf{u}(\mathbf{z})] = \mathbb{E}_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})]. \tag{10.187}$$

We see that the optimum solution simply corresponds to matching the expected sufficient statistics. So, for instance, if $q(\mathbf{z})$ is a Gaussian $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then we minimize the Kullback-Leibler divergence by setting the mean $\boldsymbol{\mu}$ of $q(\mathbf{z})$ equal to the mean of the distribution $p(\mathbf{z})$ and the covariance $\boldsymbol{\Sigma}$ equal to the covariance of $p(\mathbf{z})$. This is sometimes called *moment matching*. An example of this was seen in Figure 10.3(a).

Now let us exploit this result to obtain a practical algorithm for approximate inference. For many probabilistic models, the joint distribution of data $\mathcal{D}$ and hidden variables (including parameters) $\boldsymbol{\theta}$ comprises a product of factors in the form

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}). \tag{10.188}$$

This would arise, for example, in a model for independent, identically distributed data in which there is one factor $f_n(\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})$ for each data point $\mathbf{x}_n$, along with a factor $f_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ corresponding to the prior. More generally, it would also apply to any model defined by a directed probabilistic graph in which each factor is a conditional distribution corresponding to one of the nodes, or an undirected graph in which each factor is a clique potential. We are interested in evaluating the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ for the purpose of making predictions, as well as the model evidence $p(\mathcal{D})$ for the purpose of model comparison. From (10.188) the posterior is given by

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta}) \tag{10.189}$$

and the model evidence is given by

$$p(\mathcal{D}) = \int \prod_i f_i(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{10.190}$$

Here we are considering continuous variables, but the following discussion applies equally to discrete variables with integrals replaced by summations. We shall suppose that the marginalization over $\boldsymbol{\theta}$, along with the marginalizations with respect to the posterior distribution required to make predictions, are intractable so that some form of approximation is required.

Expectation propagation is based on an approximation to the posterior distribution which is also given by a product of factors

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \widetilde{f}_i(\boldsymbol{\theta}) \tag{10.191}$$

in which each factor $\widetilde{f}_i(\boldsymbol{\theta})$ in the approximation corresponds to one of the factors $f_i(\boldsymbol{\theta})$ in the true posterior (10.189), and the factor $1/Z$ is the normalizing constant needed to ensure that the left-hand side of (10.191) integrates to unity. In order to obtain a practical algorithm, we need to constrain the factors $\widetilde{f}_i(\boldsymbol{\theta})$ in some way, and in particular we shall assume that they come from the exponential family. The product of the factors will therefore also be from the exponential family and so can

be described by a finite set of sufficient statistics. For example, if each of the $\widetilde{f}_i(\boldsymbol{\theta})$ is a Gaussian, then the overall approximation $q(\boldsymbol{\theta})$ will also be Gaussian.

Ideally we would like to determine the $\widetilde{f}_i(\boldsymbol{\theta})$ by minimizing the Kullback-Leibler divergence between the true posterior and the approximation given by

$$\mathrm{KL}\left(p\|q\right) = \mathrm{KL}\left(\frac{1}{p(\mathcal{D})}\prod_i f_i(\boldsymbol{\theta}) \,\middle\|\, \frac{1}{Z}\prod_i \widetilde{f}_i(\boldsymbol{\theta})\right). \tag{10.192}$$

Note that this is the reverse form of KL divergence compared with that used in variational inference. In general, this minimization will be intractable because the KL divergence involves averaging with respect to the true distribution. As a rough approximation, we could instead minimize the KL divergences between the corresponding pairs $f_i(\boldsymbol{\theta})$ and $\widetilde{f}_i(\boldsymbol{\theta})$ of factors. This represents a much simpler problem to solve, and has the advantage that the algorithm is noniterative. However, because each factor is individually approximated, the product of the factors could well give a poor approximation.

Expectation propagation makes a much better approximation by optimizing each factor in turn in the context of all of the remaining factors. It starts by initializing the factors $\widetilde{f}_i(\boldsymbol{\theta})$, and then cycles through the factors refining them one at a time. This is similar in spirit to the update of factors in the variational Bayes framework considered earlier. Suppose we wish to refine factor $\widetilde{f}_j(\boldsymbol{\theta})$. We first remove this factor from the product to give $\prod_{i\neq j}\widetilde{f}_i(\boldsymbol{\theta})$. Conceptually, we will now determine a revised form of the factor $\widetilde{f}_j(\boldsymbol{\theta})$ by ensuring that the product

$$q^{\text{new}}(\boldsymbol{\theta}) \propto \widetilde{f}_j(\boldsymbol{\theta})\prod_{i\neq j}\widetilde{f}_i(\boldsymbol{\theta}) \tag{10.193}$$

is as close as possible to

$$f_j(\boldsymbol{\theta})\prod_{i\neq j}\widetilde{f}_i(\boldsymbol{\theta}) \tag{10.194}$$

in which we keep fixed all of the factors $\widetilde{f}_i(\boldsymbol{\theta})$ for $i \neq j$. This ensures that the approximation is most accurate in the regions of high posterior probability as defined by the remaining factors. We shall see an example of this effect when we apply EP to the 'clutter problem'. To achieve this, we first remove the factor $\widetilde{f}_j(\boldsymbol{\theta})$ from the current approximation to the posterior by defining the unnormalized distribution

*Section 10.7.1*

$$q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\widetilde{f}_j(\boldsymbol{\theta})}. \tag{10.195}$$

Note that we could instead find $q^{\backslash j}(\boldsymbol{\theta})$ from the product of factors $i \neq j$, although in practice division is usually easier. This is now combined with the factor $f_j(\boldsymbol{\theta})$ to give a distribution

$$\frac{1}{Z_j}f_j(\boldsymbol{\theta})q^{\backslash j}(\boldsymbol{\theta}) \tag{10.196}$$

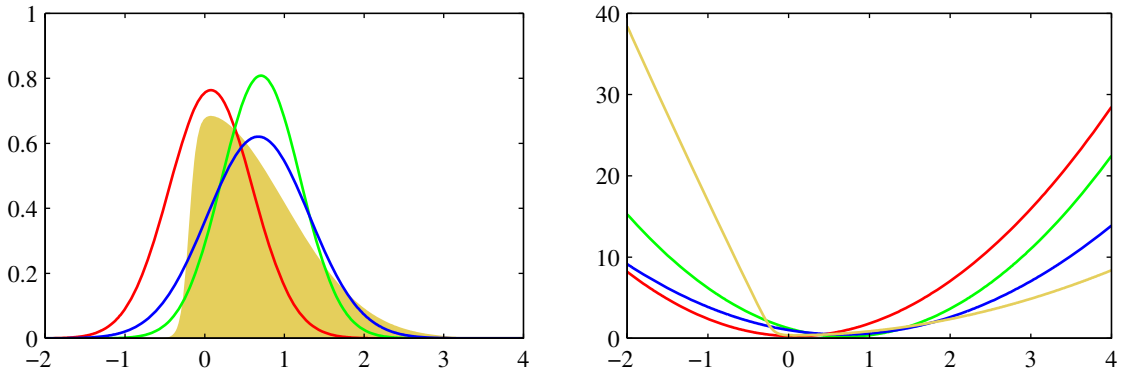



**Figure 10.14** Illustration of the expectation propagation approximation using a Gaussian distribution for the example considered earlier in Figures 4.14 and 10.1. The left-hand plot shows the original distribution (yellow) along with the Laplace (red), global variational (green), and EP (blue) approximations, and the right-hand plot shows the corresponding negative logarithms of the distributions. Note that the EP distribution is broader than that variational inference, as a consequence of the different form of KL divergence.

obtained by

where $Z_j$ is the normalization constant given by

$$Z_j = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{10.197}$$

We now determine a revised factor $\widetilde{f}_j(\boldsymbol{\theta})$ by minimizing the Kullback-Leibler divergence

$$\mathrm{KL}\left( \frac{f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta})}{Z_j} \middle\| q^{\mathrm{new}}(\boldsymbol{\theta}) \right). \tag{10.198}$$

KL( $\frac{f_j(\theta) q^{\backslash j}(\theta)}{Z_j}$ ‖ $\frac{\widetilde{f}_j(\theta) q^{\backslash j}(\theta)}{K_j}$ )

$\widetilde{f}_j(\theta)$, $K_j$ 均未知，而通过 min KL，我们能得到 $\widetilde{f}_j(\theta)/K_j$，二者如何分配实际上是任意的，为此，我们令 $K_j = Z_j$，而非方文提到的 "matching zeroth-order moments"。阶矩相等只能得到量纲意义的

$\int \frac{f_j(\theta) q^{\backslash j}(\theta)}{Z_j} d\theta = 1 = \int \frac{\widetilde{f}_j(\theta) q^{\backslash j}(\theta)}{K_j} d\theta$，我们方便得取 $K_j = Z_j$，这使得 $\widetilde{f}_j(\theta) \approx c f_j(\theta)$ 且 $c=1$。

This is easily solved because the approximating distribution $q^{\mathrm{new}}(\boldsymbol{\theta})$ is from the exponential family, and so we can appeal to the result (10.187), which tells us that the parameters of $q^{\mathrm{new}}(\boldsymbol{\theta})$ are obtained by matching its expected sufficient statistics to the corresponding moments of (10.196). We shall assume that this is a tractable operation. For example, if we choose $q(\boldsymbol{\theta})$ to be a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{\mu}$ is set equal to the mean of the (unnormalized) distribution $f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta})$, and $\boldsymbol{\Sigma}$ is set to its covariance. More generally, it is straightforward to obtain the required expectations for any member of the exponential family, provided it can be normalized, because the expected statistics can be related to the derivatives of the normalization coefficient, as given by (2.226). The EP approximation is illustrated in Figure 10.14.

From (10.193), we see that the revised factor $\widetilde{f}_j(\boldsymbol{\theta})$ can be found by taking $q^{\mathrm{new}}(\boldsymbol{\theta})$ and dividing out the remaining factors so that

$$\widetilde{f}_j(\boldsymbol{\theta}) = K_j \frac{q^{\mathrm{new}}(\boldsymbol{\theta})}{q^{\backslash j}(\boldsymbol{\theta})} \tag{10.199}$$

where we have used (10.195). The coefficient $K_j$ is determined by multiplying both

sides of (10.199) by $q^{\backslash j}(\boldsymbol{\theta})$ and integrating to give

$$K_j = \int \widetilde{f}_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{10.200}$$

where we have used the fact that $q^{\mathrm{new}}(\boldsymbol{\theta})$ is normalized. The value of $K$ can therefore be found by matching zeroth-order moments

$$\int \widetilde{f}_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{10.201}$$

Combining this with (10.197), we then see that $K_j = Z_j$ and so can be found by evaluating the integral in (10.197).

In practice, several passes are made through the set of factors, revising each factor in turn. The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ is then approximated using (10.191), and the model evidence $p(\mathcal{D})$ can be approximated by using (10.190) with the factors $f_i(\boldsymbol{\theta})$ replaced by their approximations $\widetilde{f}_i(\boldsymbol{\theta})$.

### Expectation Propagation

We are given a joint distribution over observed data $\mathcal{D}$ and stochastic variables $\boldsymbol{\theta}$ in the form of a product of factors

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}) \tag{10.202}$$

and we wish to approximate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ by a distribution of the form

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \widetilde{f}_i(\boldsymbol{\theta}). \tag{10.203}$$

We also wish to approximate the model evidence $p(\mathcal{D})$.

1. Initialize all of the approximating factors $\widetilde{f}_i(\boldsymbol{\theta})$.
2. Initialize the posterior approximation by setting

$$q(\boldsymbol{\theta}) \propto \prod_i \widetilde{f}_i(\boldsymbol{\theta}). \tag{10.204}$$

3. Until convergence:
   (a) Choose a factor $\widetilde{f}_j(\boldsymbol{\theta})$ to refine.
   (b) Remove $\widetilde{f}_j(\boldsymbol{\theta})$ from the posterior by division

$$q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\widetilde{f}_j(\boldsymbol{\theta})}. \tag{10.205}$$

(c) Evaluate the new posterior by setting the sufficient statistics (moments) of $q^{\text{new}}(\boldsymbol{\theta})$ equal to those of $q^{\backslash j}(\boldsymbol{\theta})f_j(\boldsymbol{\theta})$, including evaluation of the normalization constant

$$Z_j = \int q^{\backslash j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{10.206}$$

(d) Evaluate and store the new factor

$$\widetilde{f}_j(\boldsymbol{\theta}) = Z_j \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\backslash j}(\boldsymbol{\theta})}. \tag{10.207}$$

4. Evaluate the approximation to the model evidence

$$p(\mathcal{D}) \simeq \int \prod_i \widetilde{f}_i(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{10.208}$$

A special case of EP, known as *assumed density filtering* (ADF) or *moment matching* (Maybeck, 1982; Lauritzen, 1992; Boyen and Koller, 1998; Opper and Winther, 1999), is obtained by initializing all of the approximating factors except the first to unity and then making one pass through the factors updating each of them once. Assumed density filtering can be appropriate for on-line learning in which data points are arriving in a sequence and we need to learn from each data point and then discard it before considering the next point. However, in a batch setting we have the opportunity to re-use the data points many times in order to achieve improved accuracy, and it is this idea that is exploited in expectation propagation. Furthermore, if we apply ADF to batch data, the results will have an undesirable dependence on the (arbitrary) order in which the data points are considered, which again EP can overcome.

One disadvantage of expectation propagation is that there is no guarantee that the iterations will converge. However, for approximations $q(\boldsymbol{\theta})$ in the exponential family, if the iterations do converge, the resulting solution will be a stationary point of a particular energy function (Minka, 2001a), although each iteration of EP does not necessarily decrease the value of this energy function. This is in contrast to variational Bayes, which iteratively maximizes a lower bound on the log marginal likelihood, in which each iteration is guaranteed not to decrease the bound. It is possible to optimize the EP cost function directly, in which case it is guaranteed to converge, although the resulting algorithms can be slower and more complex to implement.

Another difference between variational Bayes and EP arises from the form of KL divergence that is minimized by the two algorithms, because the former minimizes $\mathrm{KL}(q\|p)$ whereas the latter minimizes $\mathrm{KL}(p\|q)$. As we saw in Figure 10.3, for distributions $p(\boldsymbol{\theta})$ which are multimodal, minimizing $\mathrm{KL}(p\|q)$ can lead to poor approximations. In particular, if EP is applied to mixtures the results are not sensible because the approximation tries to capture all of the modes of the posterior distribution. Conversely, in logistic-type models, EP often out-performs both local variational methods and the Laplace approximation (Kuss and Rasmussen, 2006).

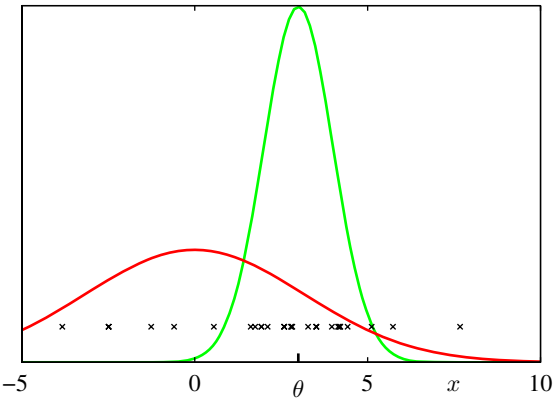



**Figure 10.15** Illustration of the clutter problem for a data space dimensionality of $D = 1$. Training data points, denoted by the crosses, are drawn from a mixture of two Gaussians with components shown in red and green. The goal is to infer the mean of the green Gaussian from the observed data.

### 10.7.1 Example: The clutter problem

Following Minka (2001b), we illustrate the EP algorithm using a simple example in which the goal is to infer the mean $\boldsymbol{\theta}$ of a multivariate Gaussian distribution over a variable $\mathbf{x}$ given a set of observations drawn from that distribution. To make the problem more interesting, the observations are embedded in background clutter, which itself is also Gaussian distributed, as illustrated in Figure 10.15. The distribution of observed values $\mathbf{x}$ is therefore a mixture of Gaussians, which we take to be of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = (1-w)\mathcal{N}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{I}) + w\mathcal{N}(\mathbf{x}|\mathbf{0}, a\mathbf{I}) \tag{10.209}$$

where $w$ is the proportion of background clutter and is assumed to be known. The prior over $\boldsymbol{\theta}$ is taken to be Gaussian

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, b\mathbf{I}) \tag{10.210}$$

注意，a,b,w均已知
我们要计算的是参数θ
证明是 及 evidence.

and Minka (2001a) chooses the parameter values $a = 10$, $b = 100$ and $w = 0.5$. The joint distribution of $N$ observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\boldsymbol{\theta}$ is given by

$$p(\mathcal{D}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}) \tag{10.211}$$

and so the posterior distribution comprises a mixture of $2^N$ Gaussians. Thus the computational cost of solving this problem exactly would grow exponentially with the size of the data set, and so an exact solution is intractable for moderately large $N$.

To apply EP to the clutter problem, we first identify the factors $f_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ and $f_n(\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})$. Next we select an approximating distribution from the exponential family, and for this example it is convenient to choose a spherical Gaussian

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, v\mathbf{I}). \tag{10.212}$$

$= \frac{1}{Z} \prod_i \tilde{f}_i(\theta)$

The factor approximations will therefore take the form of exponential-quadratic functions of the form

$$\widetilde{f}_n(\boldsymbol{\theta}) = s_n \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n\mathbf{I}) \tag{10.213}$$

where $n = 1, \ldots, N$, and we set $\widetilde{f}_0(\boldsymbol{\theta})$ equal to the prior $p(\boldsymbol{\theta})$. Note that the use of $\mathcal{N}(\boldsymbol{\theta}|\cdot,\cdot)$ does not imply that the right-hand side is a well-defined Gaussian density (in fact, as we shall see, the variance parameter $v_n$ can be negative) but is simply a convenient shorthand notation. The approximations $\widetilde{f}_n(\boldsymbol{\theta})$, for $n = 1, \ldots, N$, can be initialized to unity, corresponding to $s_n = (2\pi v_n)^{D/2}$, $v_n \to \infty$ and $\mathbf{m}_n = \mathbf{0}$, where $D$ is the dimensionality of $\mathbf{x}$ and hence of $\boldsymbol{\theta}$. The initial $q(\boldsymbol{\theta})$, defined by (10.191), is therefore equal to the prior.

We then iteratively refine the factors by taking one factor $f_n(\boldsymbol{\theta})$ at a time and applying (10.205), (10.206), and (10.207). Note that we do not need to revise the term $f_0(\boldsymbol{\theta})$ because an EP update will leave this term unchanged. Here we state the results and leave the reader to fill in the details.

*Exercise 10.37* First we remove the current estimate $\widetilde{f}_n(\boldsymbol{\theta})$ from $q(\boldsymbol{\theta})$ by division using (10.205) to give $q^{\backslash n}(\boldsymbol{\theta})$, which has mean and inverse variance given by

$$\mathbf{m}^{\backslash n} = \mathbf{m} + v^{\backslash n} v_n^{-1}(\mathbf{m} - \mathbf{m}_n) \tag{10.214}$$

$$(v^{\backslash n})^{-1} = v^{-1} - v_n^{-1}. \tag{10.215}$$

Next we evaluate the normalization constant $Z_n$ using (10.206) to give

$$Z_n = (1-w)\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n}, (v^{\backslash n}+1)\mathbf{I}) + w\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I}). \tag{10.216}$$

*Exercise 10.39* Similarly, we compute the mean and variance of $q^{\text{new}}(\boldsymbol{\theta})$ by finding the mean and variance of $q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})$ to give

$$\mathbf{m}^{\text{new}} = \mathbf{m}^{\backslash n} + \rho_n \frac{v^{\backslash n}}{v^{\backslash n}+1}(\mathbf{x}_n - \mathbf{m}^{\backslash n}) \tag{10.217}$$

$$v^{\text{new}} = v^{\backslash n} - \rho_n \frac{(v^{\backslash n})^2}{v^{\backslash n}+1} + \rho_n(1-\rho_n)\frac{(v^{\backslash n})^2\|\mathbf{x}_n - \mathbf{m}^{\backslash n}\|^2}{D(v^{\backslash n}+1)^2} \tag{10.218}$$

where the quantity

$$\rho_n = 1 - \frac{w}{Z_n}\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I}) \tag{10.219}$$

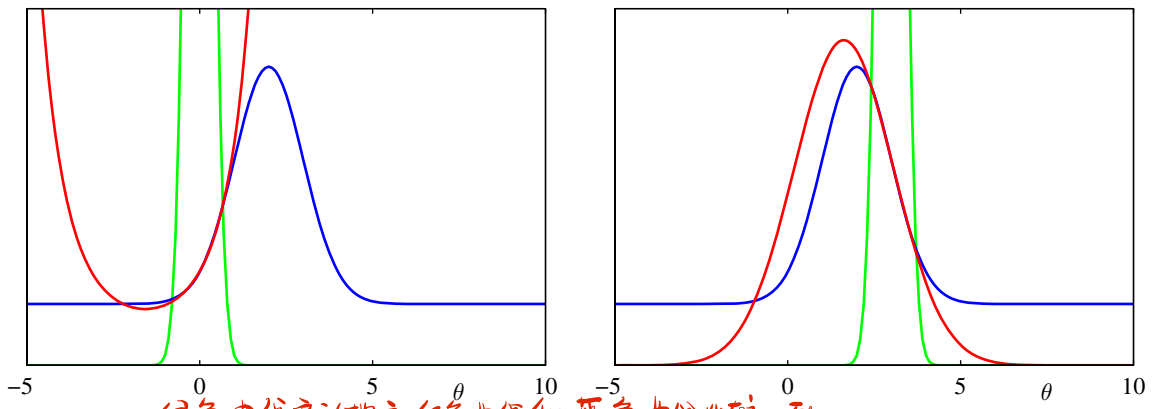
has a simple interpretation as the probability of the point $\mathbf{x}_n$ not being clutter. Then we use (10.207) to compute the refined factor $\widetilde{f}_n(\boldsymbol{\theta})$ whose parameters are given by

$$v_n^{-1} = (v^{\text{new}})^{-1} - (v^{\backslash n})^{-1} \tag{10.220}$$

$$\mathbf{m}_n = \mathbf{m}^{\backslash n} + (v_n + v^{\backslash n})(v^{\backslash n})^{-1}(\mathbf{m}^{\text{new}} - \mathbf{m}^{\backslash n}) \tag{10.221}$$

$$s_n = \frac{Z_n}{(2\pi v_n)^{D/2}\mathcal{N}(\mathbf{m}_n|\mathbf{m}^{\backslash n}, (v_n + v^{\backslash n})\mathbf{I})}. \tag{10.222}$$

This refinement process is repeated until a suitable termination criterion is satisfied, for instance that the maximum change in parameter values resulting from a complete




绿色曲线为0的地方，红色曲线和蓝色曲线比较一致

**Figure 10.16** Examples of the approximation of specific factors for a one-dimensional version of the clutter problem, showing $f_n(\theta)$ in blue, $\widetilde{f}_n(\theta)$ in red, and $q^{\backslash n}(\theta)$ in green. Notice that the current form for $q^{\backslash n}(\theta)$ controls the range of $\theta$ over which $\widetilde{f}_n(\theta)$ will be a good approximation to $f_n(\theta)$.

上面这段对应EP算法3(d)

pass through all factors is less than some threshold. Finally, we use (10.208) to evaluate the approximation to the model evidence, given by

不仅仅是针对最后一次，式(10.223)，(10.224)实际上给出了第次更新完factor $\widetilde{f}_n(\theta)$后，回正关于更新近似后验$q(\theta)$的计算公式，即$\prod_{n=1}^{N}\widetilde{f}_n(\theta)$/式(10.223)，其中$\widetilde{f}_n(\theta)$均为最新值，得到的$q(\theta)$可再被入到factor $\widetilde{f}_{j\neq n}(\theta)$的更新中。

$v^{\text{new}}, m^{\text{new}}$：最后一次更新式(10.217)，(10.218)的结果$_N$，即$q^{\text{new}}(\theta)$的均值与方差

$v$：最后一次更新前近似后验$q(\theta)$的方差

$s_n, v_n, m_n$：由最新的$\widetilde{f}_n(\theta)$给出，注意，这里的n是指前不像前文指本次更新的factor。

$$p(\mathcal{D}) \simeq (2\pi v^{\text{new}})^{D/2} \exp(B/2) \prod_{n=1}^{N} \{s_n (2\pi v_n)^{-D/2}\} \tag{10.223}$$

where

$$B = \frac{(\mathbf{m}^{\text{new}})^{\mathrm{T}}\mathbf{m}^{\text{new}}}{v} - \sum_{n=1}^{N} \frac{\mathbf{m}_n^{\mathrm{T}}\mathbf{m}_n}{v_n}. \tag{10.224}$$

最后这段对应EP算法4

Examples factor approximations for the clutter problem with a one-dimensional parameter space $\theta$ are shown in Figure 10.16. Note that the factor approximations can have infinite or even negative values for the 'variance' parameter $v_n$. This simply corresponds to approximations that curve upwards instead of downwards and are not necessarily problematic provided the overall approximate posterior $q(\boldsymbol{\theta})$ has positive variance. Figure 10.17 compares the performance of EP with variational Bayes (mean field theory) and the Laplace approximation on the clutter problem.

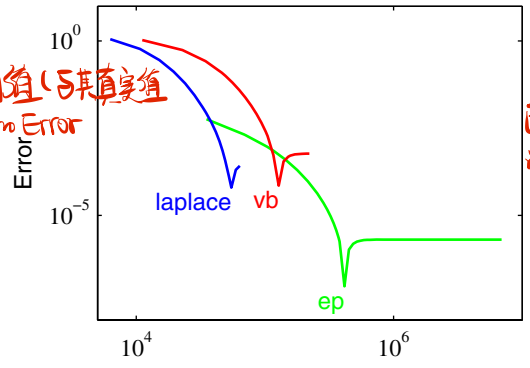
指$\widetilde{f}_n(\theta)$的形状向上弯曲而不是向下曲弯，例如Figure 10.16左图的红线（注意，这里的表述与Figure 10.17左图先下后上的趋势毫无关系）。

## 10.7.2 Expectation propagation on graphs

对方$f_i(\theta)$可进一步分解的情形，进而引用概率图模型描述。

So far in our general discussion of EP, we have allowed the factors $f_i(\boldsymbol{\theta})$ in the distribution $p(\boldsymbol{\theta})$ to be functions of all of the components of $\boldsymbol{\theta}$, and similarly for the approximating factors $\widetilde{f}(\boldsymbol{\theta})$ in the approximating distribution $q(\boldsymbol{\theta})$. We now consider situations in which the factors depend only on subsets of the variables. Such restrictions can be conveniently expressed using the framework of probabilistic graphical models, as discussed in Chapter 8. Here we use a factor graph representation because this encompasses both directed and undirected graphs.

这里对此小节内容进行梳理和说明：belief propagation (BP)算法全书并未详细给出，只是在P403中提到BP是sum-product算法的特例。我们可以认为书中提到BP算法时就是在说sum-product算法。此外，8.4.7小节介绍了loopy BP算法，它是一种近似推断算法，需要不断迭代更新，但sum-product则



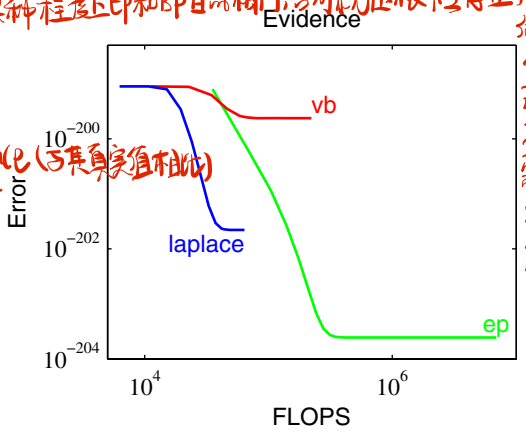




**Figure 10.17** Comparison of expectation propagation, variational inference, and the Laplace approximation on the clutter problem. The left-hand plot shows the error in the predicted posterior mean versus the number of floating point operations, and the right-hand plot shows the corresponding results for the model evidence.

We shall focus on the case in which the approximating distribution is fully factorized, and we shall show that in this case expectation propagation reduces to loopy belief propagation (Minka, 2001a). To start with, we show this in the context of a simple example, and then we shall explore the general case.

First of all, recall from (10.17) that if we minimize the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$ with respect to a factorized distribution $q$, then the optimal solution for each factor is simply the corresponding marginal of $p$.

Now consider the factor graph shown on the left in Figure 10.18, which was introduced earlier in the context of the sum-product algorithm. The joint distribution is given by

$$p(\mathbf{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4). \tag{10.225}$$

We seek an approximation $q(\mathbf{x})$ that has the same factorization, so that

$$q(\mathbf{x}) \propto \widetilde{f}_a(x_1, x_2) \widetilde{f}_b(x_2, x_3) \widetilde{f}_c(x_2, x_4). \tag{10.226}$$

Note that normalization constants have been omitted, and these can be re-instated at the end by local normalization, as is generally done in belief propagation. Now suppose we restrict attention to approximations in which the factors themselves factorize with respect to the individual variables so that

$$q(\mathbf{x}) \propto \widetilde{f}_{a1}(x_1)\widetilde{f}_{a2}(x_2)\widetilde{f}_{b2}(x_2)\widetilde{f}_{b3}(x_3)\widetilde{f}_{c2}(x_2)\widetilde{f}_{c4}(x_4) \tag{10.227}$$

which corresponds to the factor graph shown on the right in Figure 10.18. Because the individual factors are factorized, the overall distribution $q(\mathbf{x})$ is itself fully factorized.

Now we apply the EP algorithm using the fully factorized approximation. Suppose that we have initialized all of the factors and that we choose to refine factor

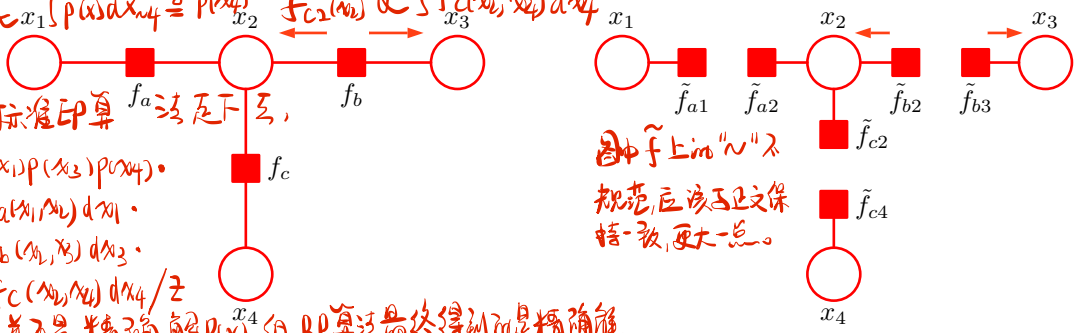



**Figure 10.18** On the left is a simple factor graph from Figure 8.51 and reproduced here for convenience. On the right is the corresponding factorized approximation.

$\tilde{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2)\tilde{f}_{b3}(x_3)$. We first remove this factor from the approximating distribution to give

$$q^{\backslash b}(\mathbf{x}) \propto \tilde{f}_{a1}(x_1)\tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2)\tilde{f}_{c4}(x_4) \tag{10.228}$$

and we then multiply this by the exact factor $f_b(x_2, x_3)$ to give

$$\widehat{p}(\mathbf{x}) = q^{\backslash b}(\mathbf{x}) f_b(x_2, x_3) = \tilde{f}_{a1}(x_1)\tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2)\tilde{f}_{c4}(x_4) f_b(x_2, x_3). \tag{10.229}$$

We now find $q^{\text{new}}(\mathbf{x})$ by minimizing the Kullback-Leibler divergence $\text{KL}(\widehat{p} \| q^{\text{new}})$. The result, as noted above, is that $q^{\text{new}}(\mathbf{z})$ comprises the product of factors, one for each variable $x_i$, in which each factor is given by the corresponding marginal of $\widehat{p}(\mathbf{x})$. These four marginals are given by

$$\widehat{p}(x_1) \propto \tilde{f}_{a1}(x_1) \tag{10.230}$$

$$\widehat{p}(x_2) \propto \tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2) \sum_{x_3} f_b(x_2, x_3) \tag{10.231}$$

$$\widehat{p}(x_3) \propto \sum_{x_2} \left\{ f_b(x_2, x_3)\tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2) \right\} \tag{10.232}$$

$$\widehat{p}(x_4) \propto \tilde{f}_{c4}(x_4) \tag{10.233}$$

and $q^{\text{new}}(\mathbf{x})$ is obtained by multiplying these marginals together. We see that the only factors in $q(\mathbf{x})$ that change when we update $\tilde{f}_b(x_2, x_3)$ are those that involve the variables in $f_b$ namely $x_2$ and $x_3$. To obtain the refined factor $\tilde{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2)\tilde{f}_{b3}(x_3)$ we simply divide $q^{\text{new}}(\mathbf{x})$ by $q^{\backslash b}(\mathbf{x})$, which gives

$$\tilde{f}_{b2}(x_2) \propto \sum_{x_3} f_b(x_2, x_3) \tag{10.234}$$

$$\tilde{f}_{b3}(x_3) \propto \sum_{x_2} \left\{ f_b(x_2, x_3)\tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2) \right\}. \tag{10.235}$$

关于EP与BP关系的详细论述请参见笔记，这里按BP的两轮传递计算得到各factor后，若仍按标准EP的步骤计算 $Q(x)$, $q(x)$ 仍然只是 $p(x)$ 的近似解，而不会像BP那样得到精确解。Figure 10.18 处备注了本例的详细结果。



*Section 8.4.4*

These are precisely the messages obtained using belief propagation in which messages from variable nodes to factor nodes have been folded into the messages from factor nodes to variable nodes. In particular, $\widetilde{f}_{b2}(x_2)$ corresponds to the message $\mu_{f_b \to x_2}(x_2)$ sent by factor node $f_b$ to variable node $x_2$ and is given by (8.81). Similarly, if we substitute (8.78) into (8.79), we obtain (10.235) in which $\widetilde{f}_{a2}(x_2)$ corresponds to $\mu_{f_a \to x_2}(x_2)$ and $\widetilde{f}_{c2}(x_2)$ corresponds to $\mu_{f_c \to x_2}(x_2)$, giving the message $\widetilde{f}_{b3}(x_3)$ which corresponds to $\mu_{f_b \to x_3}(x_3)$.

这里说的belief propagation也就是sum-product算法

这里说，若按照标准的EP算法进行更新，会同时更新 $\widetilde{f}_{b2}(x_2)$, $\widetilde{f}_{b3}(x_3)$ 两个量，对应从同一个因子结点发出的两个不同的消息，见Figure 10.18中标注的箭头。注意，这两个信息均是factor node to variable node message，是不同也上的信息，这里的计算不涉及variable node to factor node message.

This result differs slightly from standard belief propagation in that messages are passed in both directions at the same time. We can easily modify the EP procedure to give the standard form of the sum-product algorithm by updating just one of the factors at a time, for instance if we refine only $\widetilde{f}_{b3}(x_3)$, then $\widetilde{f}_{b2}(x_2)$ is unchanged by definition, while the refined version of $\widetilde{f}_{b3}(x_3)$ is again given by (10.235). If we are refining only one term at a time, then we can choose the order in which the refinements are done as we wish. In particular, for a tree-structured graph we can follow a two-pass update scheme, corresponding to the standard belief propagation schedule, which will result in exact inference of the variable and factor marginals. The initialization of the approximation factors in this case is unimportant.

tree-structure时，在与BP的计算顺序保持一致的情况下，各factor项的初始化对计算并没有影响，按步骤走一遍即可说明。

Now let us consider a general factor graph corresponding to the distribution

举例说明，从这里开始随机变量 $\boldsymbol{\theta}$ 又从 $\mathbf{x}$ 变回 $\boldsymbol{\theta}$.

$$p(\boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}_i) \tag{10.236}$$

注意，$\boldsymbol{\theta}_i$ 间可以包含相同的变量 $\theta_k$

where $\boldsymbol{\theta}_i$ represents the subset of variables associated with factor $f_i$. We approximate this using a fully factorized distribution of the form

$$q(\boldsymbol{\theta}) \propto \prod_i \prod_k \widetilde{f}_{ik}(\theta_k) \tag{10.237}$$

where $\theta_k$ corresponds to an individual variable node. Suppose that we wish to refine the particular term $\widetilde{f}_{jl}(\theta_l)$ keeping all other terms fixed. We first remove the term $\widetilde{f}_j(\boldsymbol{\theta}_j)$ from $q(\boldsymbol{\theta})$ to give

remove整个factor与消息为相得同message passing形式上相同的计算公式，即式(10.240)。接下来，计算顺序同(loopy) BP即可，并不一定同一时刻能更新一个因子factor。比如，BP的第2轮传递时，对某个因子节点，可同时更新其所有远离root结点方向的信息(因子factor)，它们的计算相互之间不复的响。

再次强调其他factor也可能含有变量 $\theta_l$

$$q^{\setminus j}(\boldsymbol{\theta}) \propto \prod_{i \neq j} \prod_k \widetilde{f}_{ik}(\theta_k) \tag{10.238}$$

and then multiply by the exact factor $f_j(\boldsymbol{\theta}_j)$. To determine the refined term $\widetilde{f}_{jl}(\theta_l)$, we need only consider the functional dependence on $\theta_l$, and so we simply find the corresponding marginal of

$$q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}_j). \tag{10.239}$$

Up to a multiplicative constant, this involves taking the marginal of $f_j(\boldsymbol{\theta}_j)$ multiplied by any terms from $q^{\setminus j}(\boldsymbol{\theta})$ that are functions of any of the variables in $\boldsymbol{\theta}_j$. Terms that correspond to other factors $\widetilde{f}_i(\boldsymbol{\theta}_i)$ for $i \neq j$ will cancel between numerator and denominator when we subsequently divide by $q^{\setminus j}(\boldsymbol{\theta})$. We therefore obtain

这里对如何推得式(10.240)叙述得很简洁，对式(10.240)本身也作详细的说明，具体推导过程同上面等价例子。在笔记中也作了详细的推导，对式(10.240)也作了更清晰的表述。

$$\widetilde{f}_{jl}(\theta_l) \propto \sum_{\theta_{m \neq l} \in \boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \prod_k \prod_{m \neq l} \widetilde{f}_{km}(\theta_m). \tag{10.240}$$

We recognize this as the sum-product rule in the form in which messages from variable nodes to factor nodes have been eliminated, as illustrated by the example shown in Figure 8.50. The quantity $\widetilde{f}_{jm}(\theta_m)$ corresponds to the message $\mu_{f_j \to \theta_m}(\theta_m)$, which factor node $j$ sends to variable node $m$, and the product over $k$ in (10.240) is over all factors that depend on the variables $\theta_m$ that have variables (other than variable $\theta_l$) in common with factor $f_j(\boldsymbol{\theta}_j)$. In other words, to compute the outgoing message from a factor node, we take the product of all the incoming messages from other factor nodes, multiply by the local factor, and then marginalize.

Thus, the sum-product algorithm arises as a special case of expectation propagation if we use an approximating distribution that is fully factorized. This suggests that more flexible approximating distributions, corresponding to partially disconnected graphs, could be used to achieve higher accuracy. Another generalization is to group factors $f_i(\boldsymbol{\theta}_i)$ together into sets and to refine all the factors in a set together at each iteration. Both of these approaches can lead to improvements in accuracy (Minka, 2001b). In general, the problem of choosing the best combination of grouping and disconnection is an open research issue.

We have seen that variational message passing and expectation propagation optimize two different forms of the Kullback-Leibler divergence. Minka (2005) has shown that a broad range of message passing algorithms can be derived from a common framework involving minimization of members of the alpha family of divergences, given by (10.19). These include variational message passing, loopy belief propagation, and expectation propagation, as well as a range of other algorithms, which we do not have space to discuss here, such as *tree-reweighted message passing* (Wainwright *et al.*, 2005), *fractional belief propagation* (Wiegerinck and Heskes, 2003), and *power EP* (Minka, 2004).

## Exercises

**10.1** (⋆) **www** Verify that the log marginal distribution of the observed data $\ln p(\mathbf{X})$ can be decomposed into two terms in the form (10.2) where $\mathcal{L}(q)$ is given by (10.3) and $\mathrm{KL}(q\|p)$ is given by (10.4).

**10.2** (⋆) Use the properties $\mathbb{E}[z_1] = m_1$ and $\mathbb{E}[z_2] = m_2$ to solve the simultaneous equations (10.13) and (10.15), and hence show that, provided the original distribution $p(\mathbf{z})$ is nonsingular, the unique solution for the means of the factors in the approximation distribution is given by $\mathbb{E}[z_1] = \mu_1$ and $\mathbb{E}[z_2] = \mu_2$.

**10.3** (⋆⋆) **www** Consider a factorized variational distribution $q(\mathbf{Z})$ of the form (10.5). By using the technique of Lagrange multipliers, verify that minimization of the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$ with respect to one of the factors $q_i(\mathbf{Z}_i)$, keeping all other factors fixed, leads to the solution (10.17).

**10.4** (⋆⋆) Suppose that $p(\mathbf{x})$ is some fixed distribution and that we wish to approximate it using a Gaussian distribution $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. By writing down the form of the KL divergence $\mathrm{KL}(p\|q)$ for a Gaussian $q(\mathbf{x})$ and then differentiating, show that

minimization of $\mathrm{KL}(p\|q)$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ leads to the result that $\boldsymbol{\mu}$ is given by the expectation of $\mathbf{x}$ under $p(\mathbf{x})$ and that $\boldsymbol{\Sigma}$ is given by the covariance.

**10.5** ($\star\star$) **www** Consider a model in which the set of all hidden stochastic variables, denoted collectively by $\mathbf{Z}$, comprises some latent variables $\mathbf{z}$ together with some model parameters $\boldsymbol{\theta}$. Suppose we use a variational distribution that factorizes between latent variables and parameters so that $q(\mathbf{z}, \boldsymbol{\theta}) = q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, in which the distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is approximated by a point estimate of the form $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is a vector of free parameters. Show that variational optimization of this factorized distribution is equivalent to an EM algorithm, in which the E step optimizes $q_{\mathbf{z}}(\mathbf{z})$, and the M step maximizes the expected complete-data log posterior distribution of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}_0$.

**10.6** ($\star\star$) The alpha family of divergences is defined by (10.19). Show that the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$ corresponds to $\alpha \to 1$. This can be done by writing $p^\epsilon = \exp(\epsilon \ln p) = 1 + \epsilon \ln p + O(\epsilon^2)$ and then taking $\epsilon \to 0$. Similarly show that $\mathrm{KL}(q\|p)$ corresponds to $\alpha \to -1$.

**10.7** ($\star\star$) Consider the problem of inferring the mean and precision of a univariate Gaussian using a factorized variational approximation, as considered in Section 10.1.3. Show that the factor $q_\mu(\mu)$ is a Gaussian of the form $\mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$ with mean and precision given by (10.26) and (10.27), respectively. Similarly show that the factor $q_\tau(\tau)$ is a gamma distribution of the form $\mathrm{Gam}(\tau|a_N, b_N)$ with parameters given by (10.29) and (10.30).

**10.8** ($\star$) Consider the variational posterior distribution for the precision of a univariate Gaussian whose parameters are given by (10.29) and (10.30). By using the standard results for the mean and variance of the gamma distribution given by (B.27) and (B.28), show that if we let $N \to \infty$, this variational posterior distribution has a mean given by the inverse of the maximum likelihood estimator for the variance of the data, and a variance that goes to zero.

**10.9** ($\star\star$) By making use of the standard result $\mathbb{E}[\tau] = a_N/b_N$ for the mean of a gamma distribution, together with (10.26), (10.27), (10.29), and (10.30), derive the result (10.33) for the reciprocal of the expected precision in the factorized variational treatment of a univariate Gaussian.

**10.10** ($\star$) **www** Derive the decomposition given by (10.34) that is used to find approximate posterior distributions over models using variational inference.

**10.11** ($\star\star$) **www** By using a Lagrange multiplier to enforce the normalization constraint on the distribution $q(m)$, show that the maximum of the lower bound (10.35) is given by (10.36).

**10.12** ($\star\star$) Starting from the joint distribution (10.41), and applying the general result (10.9), show that the optimal variational distribution $q^\star(\mathbf{Z})$ over the latent variables for the Bayesian mixture of Gaussians is given by (10.48) by verifying the steps given in the text.

**10.13** (⋆⋆) **www** Starting from (10.54), derive the result (10.59) for the optimum variational posterior distribution over $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ in the Bayesian mixture of Gaussians, and hence verify the expressions for the parameters of this distribution given by (10.60)–(10.63).

**10.14** (⋆⋆) Using the distribution (10.59), verify the result (10.64).

**10.15** (⋆) Using the result (B.17), show that the expected value of the mixing coefficients in the variational mixture of Gaussians is given by (10.69).

**10.16** (⋆⋆) **www** Verify the results (10.71) and (10.72) for the first two terms in the lower bound for the variational Gaussian mixture model given by (10.70).

**10.17** (⋆⋆⋆) Verify the results (10.73)–(10.77) for the remaining terms in the lower bound for the variational Gaussian mixture model given by (10.70).

**10.18** (⋆⋆⋆) In this exercise, we shall derive the variational re-estimation equations for the Gaussian mixture model by direct differentiation of the lower bound. To do this we assume that the variational distribution has the factorization defined by (10.42) and (10.55) with factors given by (10.48), (10.57), and (10.59). Substitute these into (10.70) and hence obtain the lower bound as a function of the parameters of the variational distribution. Then, by maximizing the bound with respect to these parameters, derive the re-estimation equations for the factors in the variational distribution, and show that these are the same as those obtained in Section 10.2.1.

**10.19** (⋆⋆) Derive the result (10.81) for the predictive distribution in the variational treatment of the Bayesian mixture of Gaussians model.

**10.20** (⋆⋆) **www** This exercise explores the variational Bayes solution for the mixture of Gaussians model when the size $N$ of the data set is large and shows that it reduces (as we would expect) to the maximum likelihood solution based on EM derived in Chapter 9. Note that results from Appendix B may be used to help answer this exercise. First show that the posterior distribution $q^\star(\boldsymbol{\Lambda}_k)$ of the precisions becomes sharply peaked around the maximum likelihood solution. Do the same for the posterior distribution of the means $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$. Next consider the posterior distribution $q^\star(\boldsymbol{\pi})$ for the mixing coefficients and show that this too becomes sharply peaked around the maximum likelihood solution. Similarly, show that the responsibilities become equal to the corresponding maximum likelihood values for large $N$, by making use of the following asymptotic result for the digamma function for large $x$

$$\psi(x) = \ln x + O\left(1/x\right). \tag{10.241}$$

Finally, by making use of (10.80), show that for large $N$, the predictive distribution becomes a mixture of Gaussians.

**10.21** (⋆) Show that the number of equivalent parameter settings due to interchange symmetries in a mixture model with $K$ components is $K!$.

解答表述得不严谨，结果也有误，详解另写。指参数之外的后验分布

**10.22** (⋆⋆) We have seen that each mode of the posterior distribution in a Gaussian mixture model is a member of a family of $K!$ equivalent modes. Suppose that the result of running the variational inference algorithm is an approximate posterior distribution $q$ that is localized in the neighbourhood of one of the modes. We can then approximate the full posterior distribution as a mixture of $K!$ such $q$ distributions, once centred on each mode and having equal mixing coefficients. Show that if we assume negligible overlap between the components of the $q$ mixture, the resulting lower bound differs from that for a single component $q$ distribution through the addition of an extra term $\ln K!$ / $K!$

不是指之为 1/K
而是指 K! 个 q distributions 混合成 而予均近似后验 的混合系数为 1/K! 。

**10.23** (⋆⋆) **www** Consider a variational Gaussian mixture model in which there is no prior distribution over mixing coefficients $\{\pi_k\}$. Instead, the mixing coefficients are treated as parameters, whose values are to be found by maximizing the variational lower bound on the log marginal likelihood. Show that maximizing this lower bound with respect to the mixing coefficients, using a Lagrange multiplier to enforce the constraint that the mixing coefficients sum to one, leads to the re-estimation result (10.83). Note that there is no need to consider all of the terms in the lower bound but only the dependence of the bound on the $\{\pi_k\}$.

**10.24** (⋆⋆) **www** We have seen in Section 10.2 that the singularities arising in the maximum likelihood treatment of Gaussian mixture models do not arise in a Bayesian treatment. Discuss whether such singularities would arise if the Bayesian model were solved using maximum posterior (MAP) estimation.

**10.25** (⋆⋆) The variational treatment of the Bayesian mixture of Gaussians, discussed in Section 10.2, made use of a factorized approximation (10.5) to the posterior distribution. As we saw in Figure 10.2, the factorized assumption causes the variance of the posterior distribution to be under-estimated for certain directions in parameter space. Discuss qualitatively the effect this will have on the variational approximation to the model evidence, and how this effect will vary with the number of components in the mixture. Hence explain whether the variational Gaussian mixture will tend to under-estimate or over-estimate the optimal number of components.

**10.26** (⋆⋆⋆) Extend the variational treatment of Bayesian linear regression to include a gamma hyperprior $\mathrm{Gam}(\beta|c_0, d_0)$ over $\beta$ and solve variationally, by assuming a factorized variational distribution of the form $q(\mathbf{w})q(\alpha)q(\beta)$. Derive the variational update equations for the three factors in the variational distribution and also obtain an expression for the lower bound and for the predictive distribution.

**10.27** (⋆⋆) By making use of the formulae given in Appendix B show that the variational lower bound for the linear basis function regression model~~, defined by (10.107),~~ can be written in the form (10.107) with the various terms defined by (10.108)–(10.112).

**10.28** (⋆⋆⋆) Rewrite the model for the Bayesian mixture of Gaussians, introduced in Section 10.2, as a conjugate model from the exponential family, as discussed in Section 10.4. Hence use the general results (10.115) and (10.119) to derive the specific results (10.48), (10.57), and (10.59).

**10.29** (⋆) **www** Show that the function $f(x) = \ln(x)$ is concave for $0 < x < \infty$ by computing its second derivative. Determine the form of the dual function $g(\lambda)$ defined by (10.133), and verify that minimization of $\lambda x - g(\lambda)$ with respect to $\lambda$ according to (10.132) indeed recovers the function $\ln(x)$.

**10.30** (⋆) By evaluating the second derivative, show that the log logistic function $f(x) = -\ln(1 + e^{-x})$ is concave. Derive the variational upper bound (10.137) directly by making a ~~second~~ first order Taylor expansion of the log logistic function around a point $x = \xi$.

**10.31** (⋆⋆) By finding the second derivative with respect to $x$, show that the function $f(x) = -\ln(e^{x/2} + e^{-x/2})$ is a concave function of $x$. Now consider the second derivatives with respect to the variable $x^2$ and hence show that it is a convex function of $x^2$. Plot graphs of $f(x)$ against $x$ and against $x^2$. Derive the lower bound (10.144) on the logistic sigmoid function directly by making a first order Taylor series expansion of the function $f(x)$ in the variable $x^2$ centred on the value $\xi^2$.

**10.32** (⋆⋆) **www** Consider the variational treatment of logistic regression with sequential learning in which data points are arriving one at a time and each must be processed and discarded before the next data point arrives. Show that a Gaussian approximation to the posterior distribution can be maintained through the use of the lower bound (10.151), in which the distribution is initialized using the prior, and as each data point is absorbed its corresponding variational parameter $\xi_n$ is optimized.

**10.33** (⋆) By differentiating the quantity $Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{\text{old}})$ defined by (10.161) with respect to the variational parameter $\xi_n$ show that the update equation for $\xi_n$ for the Bayesian logistic regression model is given by (10.163).

**10.34** (⋆⋆) In this exercise we derive re-estimation equations for the variational parameters $\boldsymbol{\xi}$ in the Bayesian logistic regression model of Section 4.5 by direct maximization of the lower bound given by (10.164). To do this set the derivative of $\mathcal{L}(\boldsymbol{\xi})$ with respect to $\xi_n$ equal to zero, making use of the result (3.117) for the derivative of the log of a determinant, together with the expressions (10.157) and (10.158) which define the mean and covariance of the variational posterior distribution $q(\mathbf{w})$.

**10.35** (⋆⋆) Derive the result (10.164) for the lower bound $\mathcal{L}(\boldsymbol{\xi})$ in the variational logistic regression model. This is most easily done by substituting the expressions for the Gaussian prior $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$, together with the lower bound $h(\mathbf{w}, \boldsymbol{\xi})$ on the likelihood function, into the integral (10.159) which defines $\mathcal{L}(\boldsymbol{\xi})$. Next gather together the terms which depend on $\mathbf{w}$ in the exponential and complete the square to give a Gaussian integral, which can then be evaluated by invoking the standard result for the normalization coefficient of a multivariate Gaussian. Finally take the logarithm to obtain (10.164).

**10.36** (⋆⋆) Consider the ADF approximation scheme discussed in Section 10.7, and show that inclusion of the factor $f_j(\boldsymbol{\theta})$ leads to an update of the model evidence of the form

$$p_j(\mathcal{D}) \simeq p_{j-1}(\mathcal{D}) Z_j \tag{10.242}$$

where $Z_j$ is the normalization constant defined by (10.197). By applying this result recursively, and initializing with $p_0(\mathcal{D}) = 1$, derive the result

$$p(\mathcal{D}) \simeq \prod_j Z_j. \tag{10.243}$$

**10.37** ($\star$) **www** Consider the expectation propagation algorithm from Section 10.7, and suppose that one of the factors $f_0(\boldsymbol{\theta})$ in the definition (10.188) has the same exponential family functional form as the approximating distribution $q(\boldsymbol{\theta})$. Show that if the factor $\widetilde{f}_0(\boldsymbol{\theta})$ is initialized to be $f_0(\boldsymbol{\theta})$, then an EP update to refine $\widetilde{f}_0(\boldsymbol{\theta})$ leaves $\widetilde{f}_0(\boldsymbol{\theta})$ unchanged. This situation typically arises when one of the factors is the prior $p(\boldsymbol{\theta})$, and so we see that the prior factor can be incorporated once exactly and does not need to be refined.

**10.38** ($\star\star\star$) In this exercise and the next, we shall verify the results (10.214)–(10.224) for the expectation propagation algorithm applied to the clutter problem. Begin by using the division formula (10.205) to derive the expressions (10.214) and (10.215) by completing the square inside the exponential to identify the mean and variance. Also, show that the normalization constant $Z_n$, defined by (10.206), is given for the clutter problem by (10.216). This can be done by making use of the general result (2.115).

**10.39** ($\star\star\star$) Show that the mean and variance of $q^{\text{new}}(\boldsymbol{\theta})$ for EP applied to the clutter problem are given by (10.217) and (10.218). To do this, first prove the following results for the expectations of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}$ under $q^{\text{new}}(\boldsymbol{\theta})$

$$\mathbb{E}[\boldsymbol{\theta}] = \mathbf{m}^{\backslash n} + v^{\backslash n}\nabla_{\mathbf{m}^{\backslash n}} \ln Z_n \tag{10.244}$$

$$\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}] = 2(v^{\backslash n})^2 \nabla_{v^{\backslash n}} \ln Z_n + 2\mathbb{E}[\boldsymbol{\theta}]^{\mathrm{T}}\mathbf{m}^{\backslash n} - \|\mathbf{m}^{\backslash n}\|^2 \tag{10.245}$$

and then make use of the result (10.216) for $Z_n$. Next, prove the results (10.220)–(10.222) by using (10.207) and completing the square in the exponential. Finally, use (10.208) to derive the result (10.223).