C.M. Bishop's PRML: Chapter 10; Approximate Inference

Florence Forbes, Fabio Cuzzolin and Ramya Narasimha

20 March 2008

Chapter outline

Need for Approximate Inference

A central task in the application of probabilistic models is the evaluation of the posterior distribution and the evaluation of expectations computed with respect to this distribution

For many models the posterior distribution or expectations w.r.t this distribution may be infeasible

- Dimensionality is too high
- Posterior distribution has a complex form for which the expectations are not tractable
- For continuous variables the integrations may not have closed form analytical solutions or dimensionality may be too large for numerical integration
- For discrete variables summing over all possible configurations of hidden variables may be exponentially large

Two approximation schemes

- Stochastic Approximation: Ex Markov chain Monte Carlo
  - Given infinite computational resource can produce exact results
  - However, sampling methods can be computationally demanding
- Deterministic Approximation:
  - Based on analytical approximation of the posterior distribution; Ex it factorizes in a particular way or has parametric form
  - However, they can never generate exact results

Variational optimization

- Originates from calculus of variations
- Like $y = f(x)$ is a mapping from $x \overset{f}{\longmapsto} y$
- A functional maps a function to a value; For ex:

$$H[p] = \int p(x) \ln p(x) \mathrm{d}x$$

- Functional derivative: How the value of a functional changes w.r.t infinitesimal changes to the input function
- Many problems can be expressed as optimization problem of finding the function that maximizes/minimizes the functional
- Approximate solutions can be obtained by restricting the range of function over which the optimization is performed

Variational optimization as applied to inference

From our discussion on EM:

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

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$KL(q||p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

If we allow any possible $q(\mathbf{Z})$ the maximum occurs when KL divergence is zero or $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$

Assuming true posterior is intractable, consider a restricted family of distributions $q(\mathbf{Z})$ and seek a member of this family for which KL divergence is minimized

Factorized Distributions

- Let $\mathbf{Z}$ be partitioned into disjoint groups $\mathbf{Z}_i$ where $i = 1 \ldots M$
- Assume q factorizes as follows:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

- There is no restriction on the functional form of $q_i(\mathbf{Z}_i)$
- This factorized form of Variational inference corresponds to *Mean Field Theory* in physics
- We make free form variational optimization of $\mathcal{L}(q)$ w.r.t all $q_i(\mathbf{Z}_i)$

Factorized distributions(II)

$$\mathcal{L}(q) = \underbrace{\int q_i \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_i) - \int q_j \ln q_j \mathrm{d}\mathbf{Z}_i}_{\text{negative KL divergence}} + const$$

where

$$\tilde{p}(\mathbf{X}, \mathbf{Z}_i) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + const$$

and

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \mathrm{d}\mathbf{Z}_i$$

- Maximize $\mathcal{L}(q)$ by keeping $\{q_{i \neq j}\}$ fixed
- This is same as minimizing KL divergence between $\tilde{p}(\mathbf{X}, \mathbf{Z}_i)$ and $q_j(\mathbf{Z}_j)$

Optimal solution

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + const$$

or

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})])}{\int \exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})])\mathrm{d}\mathbf{Z}_j}$$

- The optimum depends on the expectations of $\{q_{i\neq j}\}$
- Initialize all factors appropriately
- Cycle through the factors and replace each with revised estimate evaluated using current estimates
- Convergence is guaranteed because bound is convex w.r.t each of the factors

Properties of factorized approximations

Approximate Gaussian Distribution with factorized Gaussian

Consider,

$$p(\mathbf{z}) = (N)(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

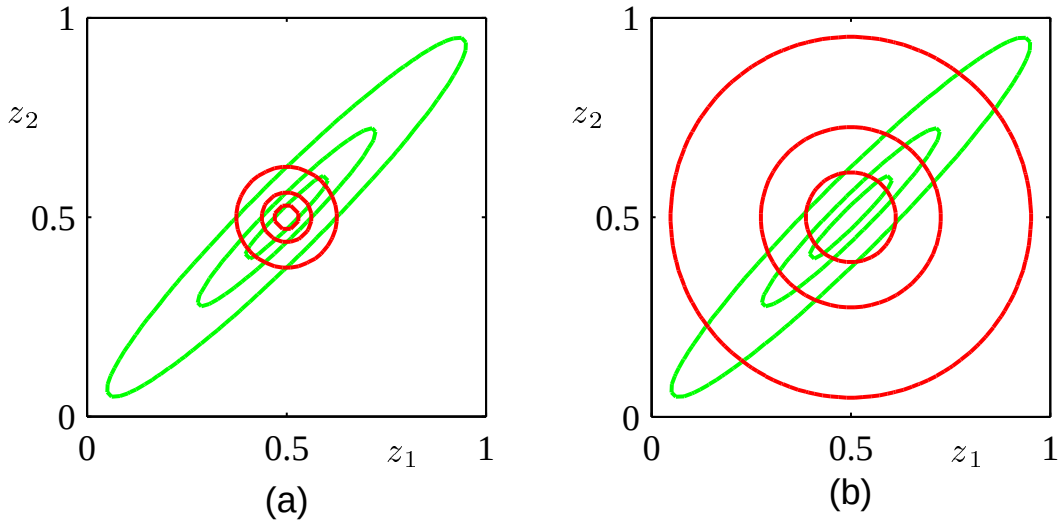
where $\mathbf{z} = (z_1, z_2)$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$

Approximate using $q(\mathbf{z}) = q_1(z_1)\, q_2(z_2)$ The optimal solution

$$q^\star(z_1) = \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1})$$

$$q^\star(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1})$$

where $m_1 = \mu_1 - \Lambda_{11}^{-1}\, \Lambda_{12}(\mathbb{E}[z_2] - \mu_2)$ and $m_2 = \mu_2 - \Lambda_{22}^{-1}\, \Lambda_{21}(\mathbb{E}[z_1] - \mu_1)$

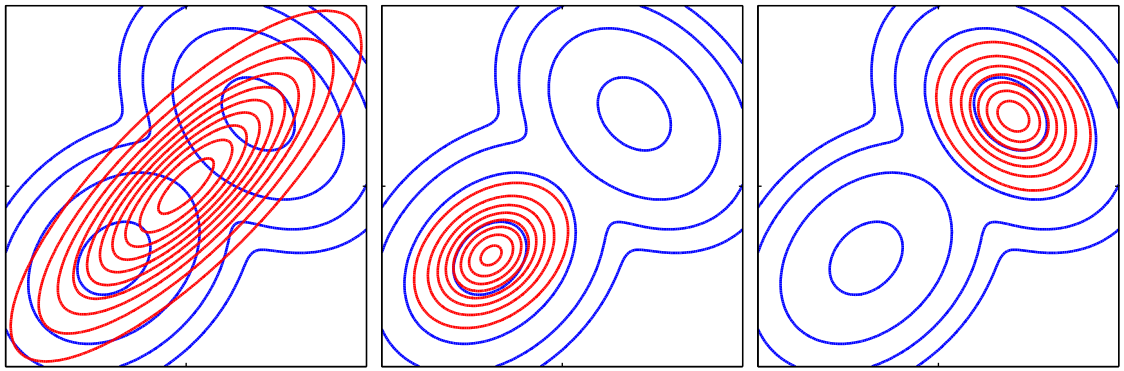


- The mean is captured correctly, but the variance is underestimated in the orthogonal direction
- Considering reverse KL divergence that is

$$KL(p||q) = -\int p(\mathbf{Z}) \Big[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \Big] \mathrm{d}\mathbf{Z} + const$$

- Optimal solution $q_j^\star(\mathbf{Z}_j) = p(\mathbf{Z}_j)$, that is the corresponding marginal distribution of $p(\mathbf{Z})$

More about divergence



- $KL(p||q)$ and $KL(q||p)$ belong to the alpha family of divergences

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2} \mathrm{d}x\right)$$

  where $-\infty < \alpha < \infty$
- if $\alpha \leq -1$, $q(x)$ will underestimate $p(x) \longrightarrow KL(q||p)$
- if $\alpha \geq -1$, $q(x)$ will overestimate $p(x) \longrightarrow KL(p||q)$
- if $\alpha = 0$ we obtain symmetric divergence measure related to *Hellinger distance*

Univariate Gaussian Example

- Goal: To infer posterior distribution for mean μ and precision τ given data set $\mathcal{D} = \{x_1, \ldots, x_N\}$
- Likelihood function

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{ -\frac{\tau}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

- Prior

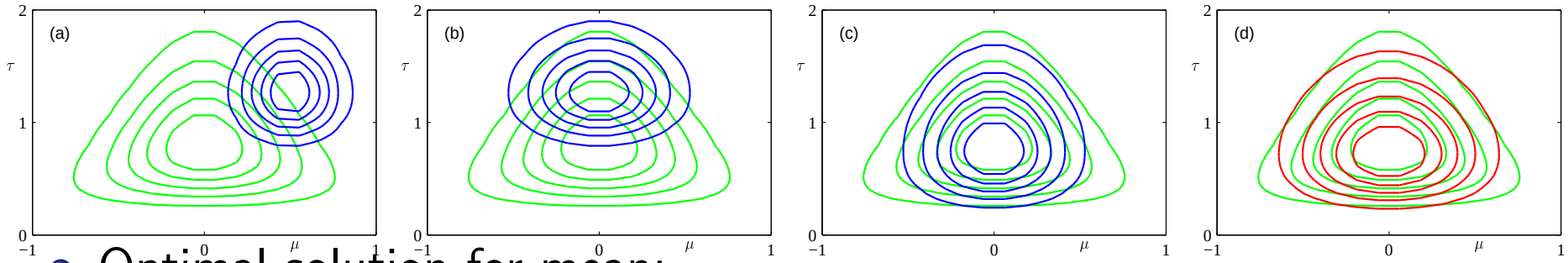
$$p(\mu|\tau) = \mathcal{N}\left(\mu|\mu_0, (\lambda_0 \tau)^{-1}\right)$$

and

$$p(\tau) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \tau^{a_0 - 1} e^{-b_0 \tau} \longrightarrow \mathsf{Gam}(\tau|a_0, b_0)$$

- Approximate

$$q(\mu, \tau) = q_\mu(\mu)\, q_\tau(\tau)$$



- Optimal solution for mean:

$$q_\mu^\star(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$$

where $\mu_N = \frac{\lambda_0\mu_0 + N\overline{x}}{\lambda_0 + N}$ and

$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau]$$

- Optimal solution for precision:

$$q_\tau^\star(\tau) = \mathsf{Gam}(\tau|a_N, b_N)$$

where $a_N = a_0 + \frac{N}{2}$ and

$$b_N = b_0 + (1/2)\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]$$

- The explicit solution can be found using simultaneous equations for the optimal factors q_μ and q_τ
- Non-informative priors $\mu_0 = a_0 = b_0 = \lambda_0 = 0$ and $\mathbb{E}[\tau] = a_N/b_N \longrightarrow$ mean of gamma distribution
- The first and second order moments of $q_\mu(\mu)$:

$$\mathbb{E}[\mu] = \overline{x}$$

and

$$\mathbb{E}[\mu^2] = \overline{x}^2 + \frac{1}{N\mathbb{E}[\tau]}$$

- Solving for $\mathbb{E}[\tau]$:

$$\frac{1}{\mathbb{E}[\tau]} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \overline{x})^2$$

Model Comparison

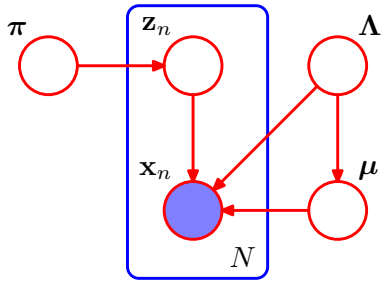
- Prior probabilities on the models be $p(m)$
- Goal: determine $p(m|\mathbf{X})$ where $\mathbf{X}$ is the observed data
- Approximate $q(\mathbf{Z}, m) = q(\mathbf{Z}|m)q(m)$

$$\ln p(\mathbf{X}) = \mathcal{L}_m - \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, m|\mathbf{X})}{q(\mathbf{Z}|m)q(m)} \right.$$

where $\mathcal{L}_m$ is the lower bound

- Maximizing $\mathcal{L}_m$ w.r.t $q(m)$ we get $q(m) \propto p(m) \exp\{\mathcal{L}\}$
- Maximizing w.r.t $q(\mathbf{Z}|m)$ we find solutions for different m are coupled due to the conditioning
- Thus, optimize each $q(\mathbf{Z}|m)$ individually and subsequently find $q(m)$
- normalized $q(m)$ can be then used for model selection

Variational mixture of Gaussian



Observation: $\mathbf{X}$,Latent variable: $\mathbf{Z}$ comprising a 1-of-K binary vector with elements z_{nk} for $k = 1, \ldots, K$

$$p(\mathbf{Z}|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi^{z_{nk}}$$

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$$

$$p(\pi) = \mathsf{Dir}(\pi|\alpha_0)$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|m_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0)$$

Variational distribution

$$p(\mathbf{X}, \mathbf{Z}, \pi, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z}|\pi) p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

Approximate

$$q(\mathbf{Z}, \pi, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z}) q(\pi, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

Optimal Solutions (I)

$$\begin{aligned}
\ln q^\star(\mathbf{Z}) &= \mathbb{E}_\pi[\ln p(\mathbf{Z}|\pi)] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\
\ln q^\star(\mathbf{Z}) &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + const \\
q^\star(\mathbf{Z}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}
\end{aligned}$$

r_{nk} are normalized ρ_{nk} values

Can also be seen as responsibilities as in case of EM

Optimal solution (II)

$$\ln q^\star(\pi) = Dir(\pi|\alpha)$$

where α has components $\alpha_k = \alpha_0 + N_k$

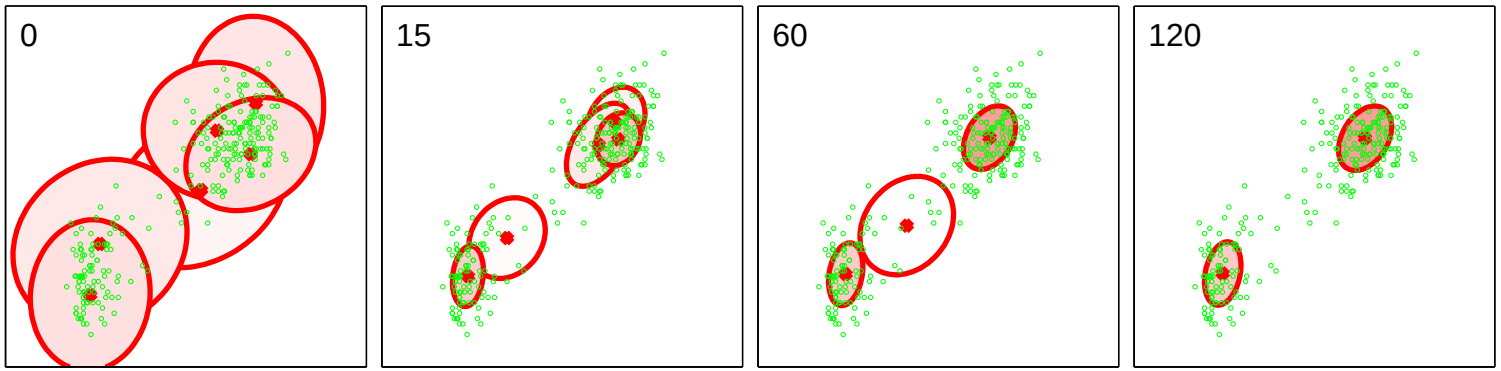
$$\ln q^\star(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}_k|m_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)$$

where $\beta_k, m_k, W_k^{-1}, \nu_k$ are updated using data and initial solutions

Remember that the responsibilities r_{nk} need the above parameters for updation, we can thus optimize the variational posterior distribution through cycling analogous to EM procedure

Variational equivalent of EM

- E step: Use the current distribution of parameters to evaluate the responsibilities
- M step: Fix responsibilities and use it to recompute the variational distribution over parameters



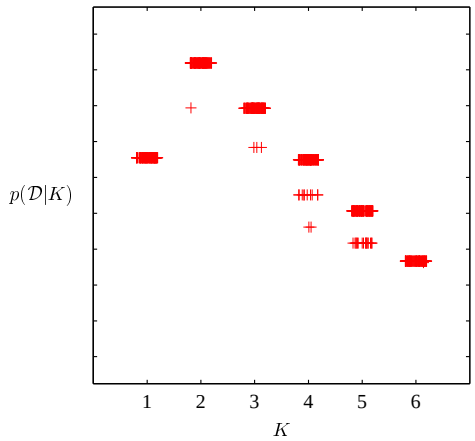
- As $N \rightarrow \infty$ Bayesian treatment converges to Maximum likelihood EM algorithm
- Singularities that arise in ML are absent in Bayesian treatment; removed by the introduction of prior
- No over-fitting; determines the number of components

More about variational approximation I

- Variational lower bound
  - Useful to test convergence
  - To check on the correctness of both mathematical expressions and implementation
- Predictive density $P(\hat{x}|X)$,for a new value $\hat{x}$ with corresponding latent variable $\hat{z}$
  - Depends on the posterior distribution of parameters
  - As the posterior distribution is intractable the variational approximation can be used to obtain an approximate predictive density

Number of components

- For a given mixture model of K components, each parameter setting is a member of a family of $K!$ equivalent settings
- This is not a problem when modeling given the number of components
- However, for model comparison an approximate solution is to add $\ln K!$

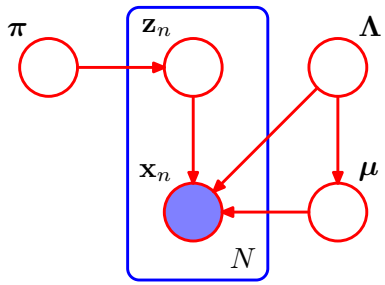


- The Bayesian inference makes automatic trade off between model complexity and fitting the data
- Starting with relative large value of K and components with insufficient contribution are pruned out : the mixing coefficient is driven to zero

Induced factorizations

Induced factorization arises from an interaction between the factorization assumption in variational posterior and the conditional independence properties of the true posterior

- For ex: Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ be disjoint groups of latent variables
- Factorization assumption $q(\mathbf{A}, \mathbf{B}, \mathbf{C}) = q(\mathbf{A}, \mathbf{B})q(\mathbf{C})$
- The optimal solution $\ln q^\star(\mathbf{A}, \mathbf{B}) = \mathbb{E}_{\mathbf{C}}[\ln p(\mathbf{A}, \mathbf{B}|\mathbf{X}, \mathbf{C})] + const$
- We need to determine if $q^\star(\mathbf{A}, \mathbf{B}) = q^\star(\mathbf{A})q^\star(\mathbf{B})$: This is possible iff $\mathbf{A} \perp \mathbf{B}|\mathbf{X}, \mathbf{C}$
- This can also determined from the graph using d-separation



Chap. 10: Approximate inference
Part 2

Variational logistic regression
Expectation Propagation

F. Forbes

Local variational models

- Global methods: approximation to the full posterior
- Local methods: approximation to individual or groups of variables

→ A practical example: Logistic regression

Variational Logistic Regression

Variational framework: 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

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

where $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$.

Variational lower bound on the logistic sigmoid function:

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

where

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

We can therefore write

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

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

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

→ a lower bound on the log of the joint distribution of $\mathbf{t}$ and $\mathbf{w}$ of the form

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

Hypothesis for the prior p(w): Gaussian with parameters m_0 and S_0 considered as fixed

The rhs of this inequality becomes

$$-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) + \sum_{n=1}^{N}\Big\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n(t_n - 1/2) - \lambda(\xi_n)\mathbf{w}^{\mathrm{T}}(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}})\mathbf{w}\Big\} + \text{const.} \quad (10.155)$$

Quantity of interest: exact **posterior distribution**, requires normalisation of the lhs in (10.152) usually intractable

Work instead with the rhs (10.155): a quadratic function of w which is a lower bound of $p(w; t)$

The corresponding variational approximation to the posterior is obtained by normalizing this lower bound which leads to a Gaussian distribution:

a Gaussian variational posterior of the form

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

where

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

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

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).

Optimizing the variational parameters

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

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

Two approaches:

7) View w as a latent variable and use the EM algorithm

8) Compute and maximize $\log p(\mathbf{w}; \boldsymbol{\xi})$ directly, using the fact that p(w) is Gaussian and $L(\xi)$ is a quadratic function of w.

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

1) and 2) lead to the same re-estimation equations

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

Illustration

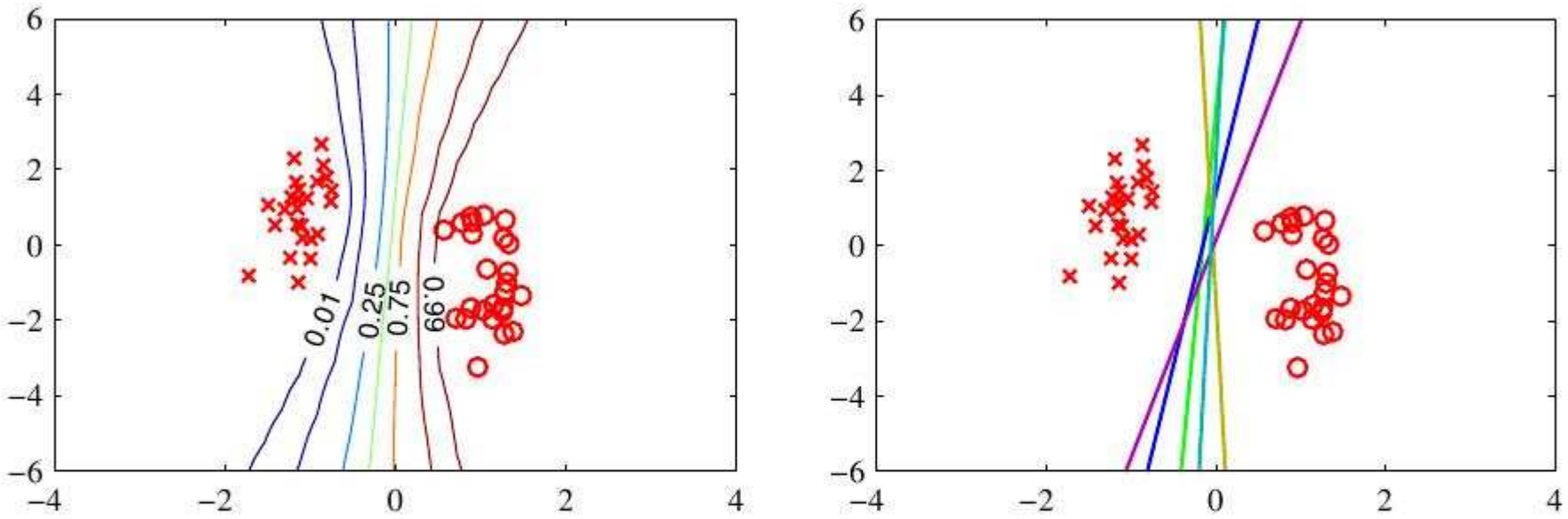


Figure 10.13 Illustration of the Bayesian approach to logistic regression for a simple linearly separable data set. The plot on the left shows the predictive distribution obtained using variational inference. We see that the decision boundary lies roughly mid way between the clusters of data points, 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 $p(\mathbf{w}|\mathbf{t})$.

Inference of hyperparameters

Allow the hyperparameters in the prior to be inferred from the data set

We consider A simple isotropic Gaussian prior of the form

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

we consider a conjugate hyperprior over α given by a gamma distribution

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

→ 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

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

An intractable integration over w and ® →combine local and global variational approaches

1) Global approach: consider a variational distribution $q(w;®)$ and apply the *standard* decomposition

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

2) But the lower bound $L(q)$ is intractable so apply the local approach as before to get a lower bound on $L(q)$ and on ln p(t):

$$\begin{aligned} \ln p(\mathbf{t}) &\geqslant \mathcal{L}(q) \geqslant \widetilde{\mathcal{L}}(q, \boldsymbol{\xi}) \\ &= \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w}|\alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \right\} \mathrm{d}\mathbf{w}\, \mathrm{d}\alpha. \end{aligned} \qquad (10.172)$$

3)Then assume that q factorizes and use the *standard* result

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

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

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

It follows (quadratic function of w):

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

where we have defined

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

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

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

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

$$q(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N) = \frac{1}{\Gamma(a_0)}a_0^{b_0}\alpha^{a_0-1}e^{-b_0\alpha} \tag{10.177}$$

where

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

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

The variational parameters are obtained by maximizing $\widetilde{\mathcal{L}}(q;\xi)$ similarly as before:

maximizing the lower bound $\widetilde{\mathcal{L}}(q, \boldsymbol{\xi})$. Omitting terms that are independent of $\boldsymbol{\xi}$, and integrating over α, we have

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

leading to re-estimation equations of the form

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

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

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

Expectation Propagation (EP)

An alternative form of deterministic approximate inference (Minka 2001) based on the *reverse* KL divergence KL(p ||q) (instead of KL(q||p)) where p is the *complexe* distribution

$$KL(q||p) = \int q(z) \ln \frac{q(z)}{p(z)} dz \qquad KL(p||q) = \int p(z) \ln \frac{p(z)}{q(z)} dz$$

When q is in the exponential family

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

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

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

It is minimized when

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

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

Sometimes called ***Moment matching***, eg. If q(z) is assumed Gaussian $\mathcal{N}(z|\mu, \Sigma)$

KL(p||q) is minimized by setting respectively μ and Σ to the mean and covariance of p.

Use this property for approximate inference. Assume the joint distribution of data and hidden variables and parameters is of the form:

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

Example: $f_n(\mu) = p(x_n|\mu)$ and $\tau_0(\mu) = p(\mu)$

Quantities of interest are $p(\mu|D)$ (prediction) and $p(D)$ (model comparison)

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

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

Expectation propagation starts by assuming that:

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

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

Where each factor $\widetilde{f}_i(\mu)$ comes from the exponential family

Ideally the factors are those minimizing KL(p||q): usually intractable

12) Minimize the KL divergences between each pair of factors $f(\mu); \widetilde{f}_i(\mu)$ *independently* but the product is usually a poor approximation

2) Expectation Propagation: optimize each factor *in turn* using the current values for the remaining factors

Advantages and limits of EP:

Out-performs in Logistic type models but bad for mixtures due to multi-modality

No guarantee of convergence but results in the exponential family case

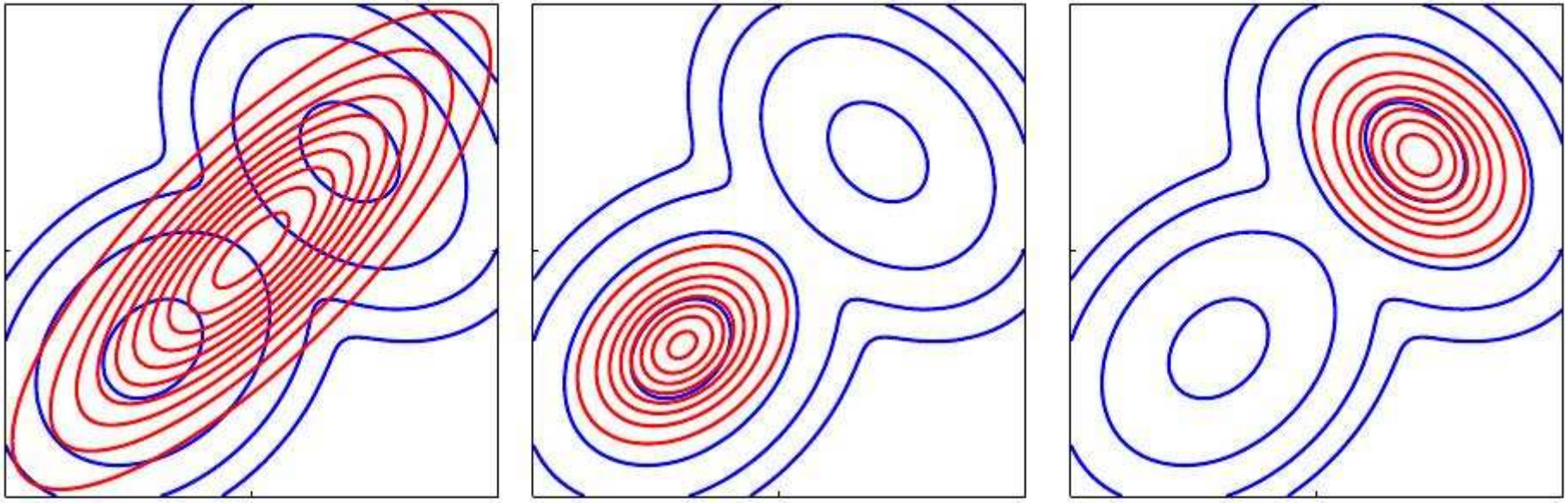


Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

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

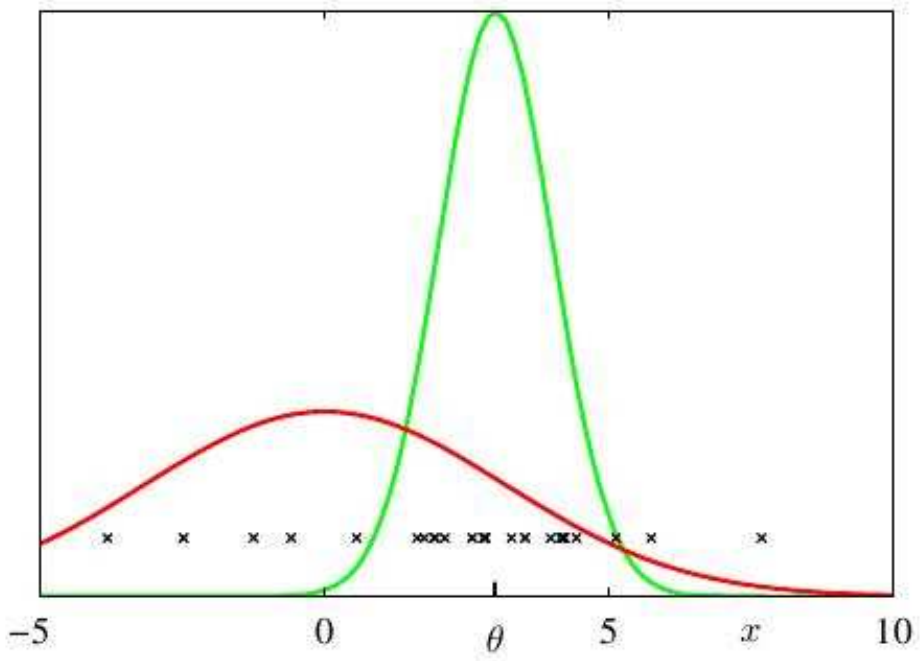
$$\widetilde{f}_j(\boldsymbol{\theta}) = Z_j \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\backslash j}(\boldsymbol{\theta})}. \tag{10.207}$$

4. Evaluate the approximation to the model evidence

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

Example: the clutter problem

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.



The clutter problem

The observed values come from a mixture of Gaussians 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

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

The joint distribution of N observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\boldsymbol{\theta}$ is given by

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

To apply EP, we first identify $f_0(\mu) = p(\mu)$ and $f_n(\mu) = p(x_n|\mu)$

We select an approximating distribution from the exponential family, here we choose:

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

It follows that the factor approximations take the form of exponentials of quadratic functions:

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

These factors are initialized to unity, ie. q is initialized to the prior. Then only the factors are iteratively refined one at a time.

Removing the current estimate $\tilde{f}_n(\mu)$ from $q(\mu)$ by division

to give $q^{\backslash n}(\boldsymbol{\theta})$, which has mean and inverse variance given by

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

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

Mean and variance designated the parameters in the factor definition, variance can be negative…

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

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

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

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

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

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

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

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

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

Finally, we use (10.208) to evaluate the approximation to the model evidence, given by

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

where

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

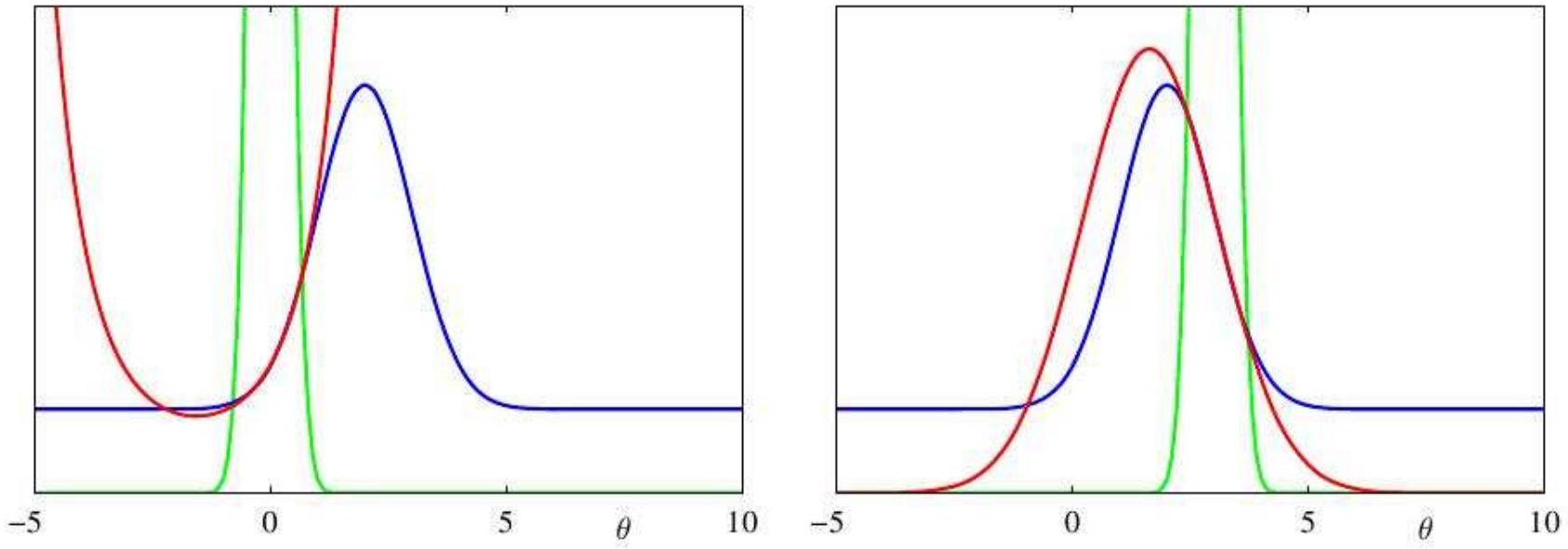
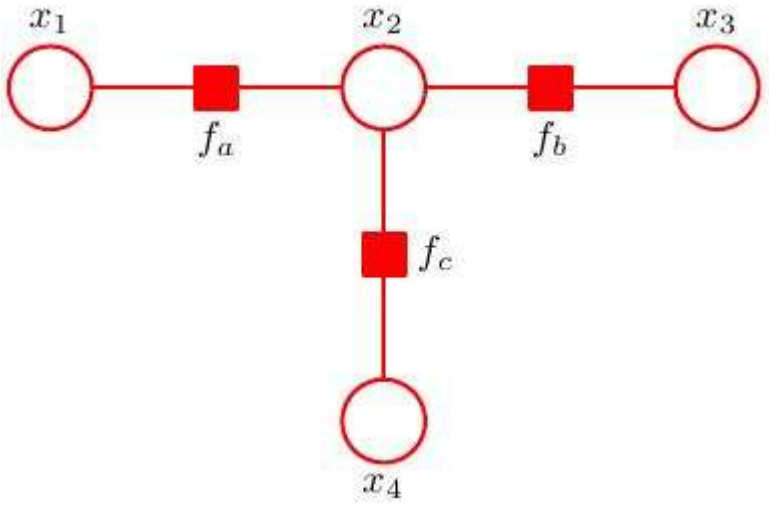


Figure 10.16 Examples of the approximation of specific factors for a one-dimensional version of the clutter problem, showing $f_n(\theta)$ in blue, $\widetilde{f}_n(\theta)$ in red, and $q^{\backslash n}(\theta)$ in green. Notice that the current form for $q^{\backslash n}(\theta)$ controls the range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

Expectation Propagation on graphs

The factors are not function of all variables. If the approximating distribution is fully factorized, EP reduces to Loopy Belief Propagation

A simple example



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

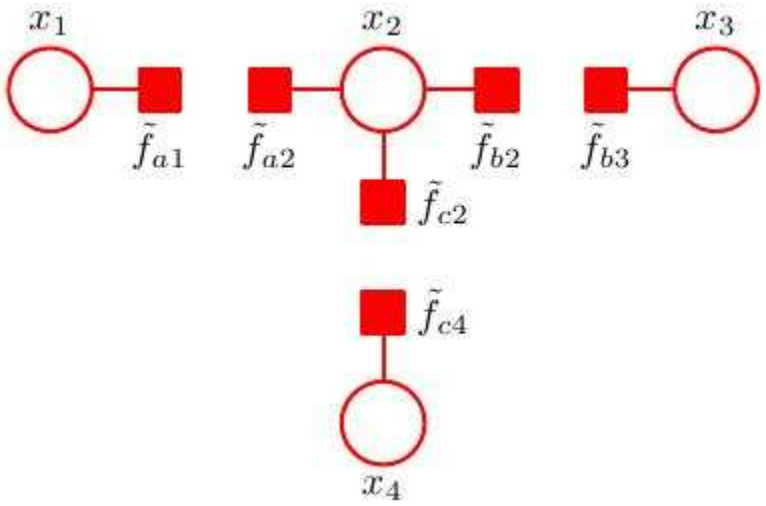
We seek an approximation $q(\mathbf{x})$ that has the same factorization, so that

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

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.



Suppose, all the factors are initialized and we chose to refine factor

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

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

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

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

We now find $q^{new}(x)$ by minimizing $KL(\hat{p}||q^{new})$ which leads to $q^{new}(x)$ as the product of the marginals of $\hat{p}$ which are given by

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

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

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

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

Recall that (10.17) minimizing the reverse KL when q factorizes, leads to an optimal solution q where the factors are the marginals of p

When updating $\tilde{f}_b(x_2, x_3)$ factors $\tilde{f}_{a_1}$ and $\tilde{f}_{c4}$ do not change. The ones that change are the ones that involve the variables in f_b

To obtain the refined factor $\hat{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2)\tilde{f}_{b3}(x_3)$ we simply divide $q^{\text{new}}(\mathbf{x})$ by $q^{\backslash b}(\mathbf{x})$, which gives

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

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

Standard belief propagation

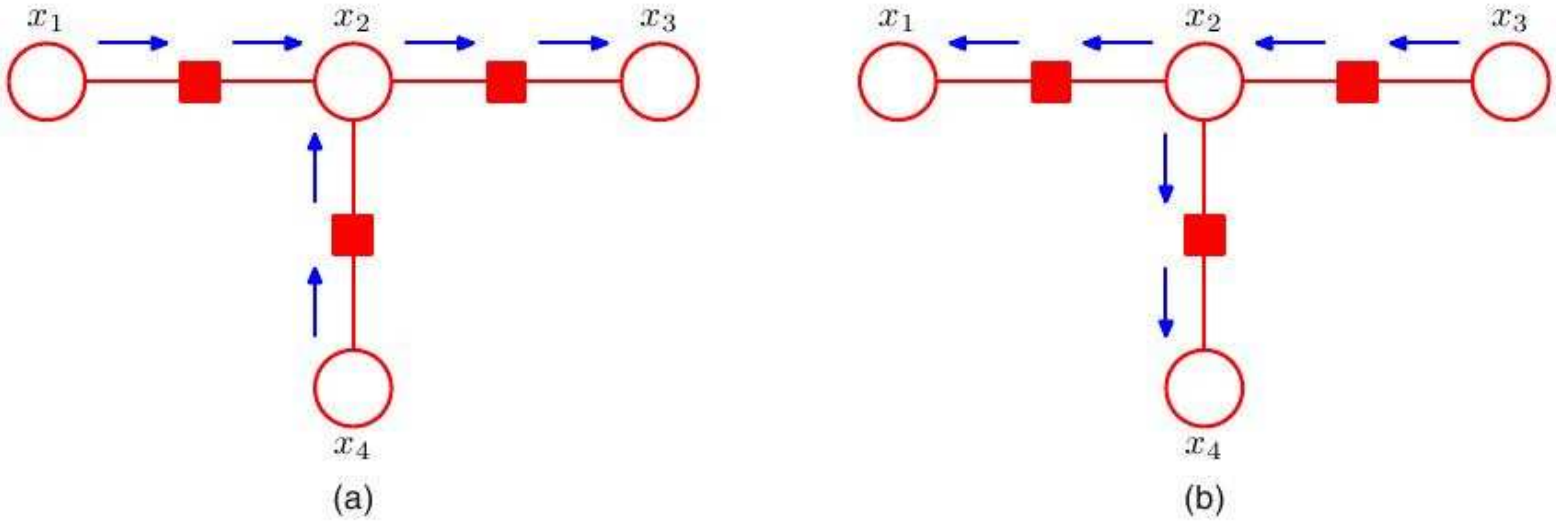


Figure 8.52 Flow of messages for the sum-product algorithm applied to the example graph in Figure 8.51. (a) From the leaf nodes x_1 and x_4 towards the root node x_3. (b) From the root node towards the leaf nodes.

$$\mu_{x_1 \to f_a}(x_1) = 1 \tag{8.74}$$

$$\mu_{f_a \to x_2}(x_2) = \sum_{x_1} f_a(x_1, x_2) \tag{8.75}$$

$$\mu_{x_4 \to f_c}(x_4) = 1 \tag{8.76}$$

$$\mu_{f_c \to x_2}(x_2) = \sum_{x_4} f_c(x_2, x_4) \tag{8.77}$$

$$\mu_{x_2 \to f_b}(x_2) = \mu_{f_a \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \tag{8.78}$$

$$\mu_{f_b \to x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3)\mu_{x_2 \to f_b}. \tag{8.79}$$

$$\mu_{x_3 \to f_b}(x_3) = 1 \tag{8.80}$$

$$\mu_{f_b \to x_2}(x_2) = \sum_{x_3} f_b(x_2, x_3) \tag{8.81}$$

$$\mu_{x_2 \to f_a}(x_2) = \mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \tag{8.82}$$

$$\mu_{f_a \to x_1}(x_1) = \sum_{x_2} f_a(x_1, x_2)\mu_{x_2 \to f_a}(x_2) \tag{8.83}$$

$$\mu_{x_2 \to f_c}(x_2) = \mu_{f_a \to x_2}(x_2)\mu_{f_b \to x_2}(x_2) \tag{8.84}$$

$$\mu_{f_c \to x_4}(x_4) = \sum_{x_2} f_c(x_2, x_4)\mu_{x_2 \to f_c}(x_2). \tag{8.85}$$

In particular,

$\tilde{f}_{b2}(x_2) \Rightarrow {}^{1}_{f_b! x_2}(x_2)$ (8.81)

$\tilde{f}_{b3}(x_3) \Rightarrow {}^{1}_{f_b! x_3}(x_3)$ (8.78) into (8.79)

$\tilde{f}_{a2}(x_2) \Rightarrow {}^{1}_{f_a! x_2}(x_2)$

$\tilde{f}_{c2}(x_2) \Rightarrow {}^{1}_{f_c! x_2}(x_2)$

This EP slightly differs from standard BP in that messages are passed in both direction at the same time → modified EP: update just one of the factor at a time

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

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

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 θ_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

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

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 θ_l, and so we simply find the corresponding marginal of

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

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

This is the sum-product rule seen in chapter 8

The **sum-product BP arises as a special case of EP** when a fully factorized approximating distributions is used.

EP can be seen as a way to generalized this: group factors and update them together, use partially disconnected graph.

Q: how to choose the best grouping and disconnection?

Summary: EP and Variational message passing correspond to the optimization of two different KL divergences

Minka 2005 gives **a more general point of view** using the family of **alpha-divergences** that includes both KL and reverse KL, but also other divergence like Hellinger distance, Chi2-distance…

He shows that by choosing to optimize one or the other of these divergences, you can **derive a broad range of message passing algorithms** including Variational message passing, Loopy BP, EP, Tree-Reweighted BP, Fractional BP, power EP.