Chapter 2

Variational Bayesian Theory

2.1 Introduction

This chapter covers the majority of the theory for variational Bayesian learning that will be used in rest of this thesis. It is intended to give the reader a context for the use of variational methods as well as a insight into their general applicability and usefulness.

In a model selection task the role of a Bayesian is to calculate the posterior distribution over a set of models given some a priori knowledge and some new observations (data). The knowledge is represented in the form of a prior over model structures $p(m)$, and their parameters $p(\boldsymbol{\theta} \,|\, m)$ which define the probabilistic dependencies between the variables in the model. By Bayes' rule, the posterior over models m having seen data $\mathbf{y}$ is given by:

$$p(m \,|\, \mathbf{y}) = \frac{p(m)p(\mathbf{y} \,|\, m)}{p(\mathbf{y})} \; . \tag{2.1}$$

The second term in the numerator is the *marginal likelihood* or *evidence* for a model m, and is the key quantity for Bayesian model selection:

$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta} \; p(\boldsymbol{\theta} \,|\, m)p(\mathbf{y} \,|\, \boldsymbol{\theta}, m) \; . \tag{2.2}$$

For each model structure we can compute the posterior distribution over parameters:

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, m) = \frac{p(\boldsymbol{\theta} \,|\, m)p(\mathbf{y} \,|\, \boldsymbol{\theta}, m)}{p(\mathbf{y} \,|\, m)} \; . \tag{2.3}$$

We might also be interested in calculating other related quantities, such as the *predictive density* of a new datum $\mathbf{y}'$ given a data set $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$:

$$p(\mathbf{y}' \,|\, \mathbf{y}, m) = \int d\boldsymbol{\theta} \; p(\boldsymbol{\theta} \,|\, \mathbf{y}, m) \, p(\mathbf{y}' \,|\, \boldsymbol{\theta}, \mathbf{y}, m) \; , \tag{2.4}$$

which can be simplified into

$$p(\mathbf{y}' \,|\, \mathbf{y}, m) = \int d\boldsymbol{\theta} \; p(\boldsymbol{\theta} \,|\, \mathbf{y}, m) \, p(\mathbf{y}' \,|\, \boldsymbol{\theta}, m) \tag{2.5}$$

if $\mathbf{y}'$ is conditionally independent of $\mathbf{y}$ given $\boldsymbol{\theta}$. We also may be interested in calculating the posterior distribution of a hidden variable, $\mathbf{x}'$, associated with the new observation $\mathbf{y}'$

$$p(\mathbf{x}' \,|\, \mathbf{y}', \mathbf{y}, m) \propto \int d\boldsymbol{\theta} \; p(\boldsymbol{\theta} \,|\, \mathbf{y}, m) \, p(\mathbf{x}', \mathbf{y}' \,|\, \boldsymbol{\theta}, m) \; . \tag{2.6}$$

The simplest way to approximate the above integrals is to estimate the value of the integrand at a single point estimate of $\boldsymbol{\theta}$, such as the maximum likelihood (ML) or the maximum a posteriori (MAP) estimates, which aim to maximise respectively the second and both terms of the integrand in (2.2),

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \; p(\mathbf{y} \,|\, \boldsymbol{\theta}, m) \tag{2.7}$$

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m) \; . \tag{2.8}$$

ML and MAP examine only probability *density*, rather than *mass*, and so can neglect potentially large contributions to the integral. A more principled approach is to estimate the integral numerically by evaluating the integrand at many different $\boldsymbol{\theta}$ via Monte Carlo methods. In the limit of an infinite number of samples of $\boldsymbol{\theta}$ this produces an accurate result, but despite ingenious attempts to curb the curse of dimensionality in $\boldsymbol{\theta}$ using methods such as Markov chain Monte Carlo, these methods remain prohibitively computationally intensive in interesting models. These methods were reviewed in the last chapter, and the bulk of this chapter concentrates on a third way of approximating the integral, using *variational* methods. The key to the variational method is to approximate the integral with a simpler form that is tractable, forming a lower or upper *bound*. The integration then translates into the implementationally simpler problem of bound *optimisation*: making the bound as tight as possible to the true value.

We begin in section 2.2 by describing how variational methods can be used to derive the well-known expectation-maximisation (EM) algorithm for learning the maximum likelihood (ML) parameters of a model. In section 2.3 we concentrate on the Bayesian methodology, in which priors are placed on the parameters of the model, and their uncertainty integrated over to give the *marginal likelihood* (2.2). We then generalise the variational procedure to yield the *variational Bayesian EM* (VBEM) algorithm, which iteratively optimises a lower bound on this marginal likelihood. In analogy to the EM algorithm, the iterations consist of a variational Bayesian E (VBE) step in which the hidden variables are inferred using an *ensemble* of models according to their posterior probability, and a variational Bayesian M (VBM) step in which a posterior *distribution* over model parameters is inferred. In section 2.4 we specialise this algorithm to a large class of models which we call *conjugate-exponential* (CE): we present the variational Bayesian EM algorithm for CE models and discuss the implications for both directed graphs (Bayesian networks) and undirected graphs (Markov networks) in section 2.5. In particular we show that we can incorporate existing propagation algorithms into the variational Bayesian framework and that the complexity of inference for the variational Bayesian treatment is approximately the same as for the ML scenario. In section 2.6 we compare VB to the BIC and Cheeseman-Stutz criteria, and finally summarise in section 2.7.

2.2 Variational methods for ML / MAP learning

In this section we review the derivation of the EM algorithm for probabilistic models with hidden variables. The algorithm is derived using a variational approach, and has exact and approximate versions. We investigate themes on convexity, computational tractability, and the Kullback-Leibler divergence to give a deeper understanding of the EM algorithm. The majority of the section concentrates on maximum likelihood (ML) learning of the parameters; at the end we present the simple extension to maximum a posteriori (MAP) learning. The hope is that this section provides a good stepping-stone on to the variational Bayesian EM algorithm that is presented in the subsequent sections and used throughout the rest of this thesis.

2.2.1 The scenario for parameter learning

Consider a model with hidden variables $\mathbf{x}$ and observed variables $\mathbf{y}$. The parameters describing the (potentially) stochastic dependencies between variables are given by $\boldsymbol{\theta}$. In particular consider the generative model that produces a dataset $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ consisting of n independent and identically distributed (i.i.d.) items, generated using a set of hidden variables $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ such that the likelihood can be written as a function of $\boldsymbol{\theta}$ in the following way:

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \int d\mathbf{x}_i \; p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \; . \tag{2.9}$$

The integration over hidden variables $\mathbf{x}_i$ is required to form the likelihood of the parameters, as a function of just the observed data $\mathbf{y}_i$. We have assumed that the hidden variables are continuous as opposed to discrete (hence an integral rather than a summation), but we do so without loss of generality. As a point of nomenclature, note that we use $\mathbf{x}_i$ and $\mathbf{y}_i$ to denote collections of $|\mathbf{x}_i|$ hidden and $|\mathbf{y}_i|$ observed variables respectively: $\mathbf{x}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i|\mathbf{x}_i|}\}$, and

$\mathbf{y}_i = \{\mathbf{y}_{i1}, \ldots, \mathbf{y}_{i|\mathbf{y}_i|}\}$. We use $|\cdot|$ notation to denote the size of the collection of variables. ML learning seeks to find the parameter setting $\boldsymbol{\theta}_{\text{ML}}$ that maximises this likelihood, or equivalently the logarithm of this likelihood,

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \ln p(\mathbf{y}\,|\,\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{y}_i\,|\,\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \int d\mathbf{x}_i \; p(\mathbf{x}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}) \tag{2.10}$$

so defining

$$\boldsymbol{\theta}_{\text{ML}} \equiv \arg\max_{\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{\theta}) \,. \tag{2.11}$$

To keep the derivations clear, we write $\mathcal{L}$ as a function of $\boldsymbol{\theta}$ only; the dependence on $\mathbf{y}$ is implicit. In Bayesian networks without hidden variables and with independent parameters, the log-likelihood decomposes into local terms on each $\mathbf{y}_{ij}$, and so finding the setting of each parameter of the model that maximises the likelihood is straightforward. Unfortunately, if some of the variables are hidden this will in general induce dependencies between all the parameters of the model and so make maximising (2.10) difficult. Moreover, for models with many hidden variables, the integral (or sum) over $\mathbf{x}$ can be intractable.

We simplify the problem of maximising $\mathcal{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ by introducing an auxiliary distribution over the hidden variables. *Any* probability distribution $q_{\mathbf{x}}(\mathbf{x})$ over the hidden variables gives rise to a *lower bound* on $\mathcal{L}$. In fact, for each data point $\mathbf{y}_i$ we use a distinct distribution $q_{\mathbf{x}_i}(\mathbf{x}_i)$ over the hidden variables to obtain the lower bound:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_i \ln \int d\mathbf{x}_i \; p(\mathbf{x}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}) && (2.12)\\
&= \sum_i \ln \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \frac{p(\mathbf{x}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta})}{q_{\mathbf{x}_i}(\mathbf{x}_i)} && (2.13)\\
&\geq \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln \frac{p(\mathbf{x}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta})}{q_{\mathbf{x}_i}(\mathbf{x}_i)} && (2.14)\\
&= \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln p(\mathbf{x}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}) - \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln q_{\mathbf{x}_i}(\mathbf{x}_i) && (2.15)\\
&\equiv \mathcal{F}(q_{\mathbf{x}_1}(\mathbf{x}_1), \ldots, q_{\mathbf{x}_n}(\mathbf{x}_n), \boldsymbol{\theta}) && (2.16)
\end{aligned}$$

where we have made use of Jensen's inequality (Jensen, 1906) which follows from the fact that the log function is concave. $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ is a lower bound on $\mathcal{L}(\boldsymbol{\theta})$ and is a functional of the free distributions $q_{\mathbf{x}_i}(\mathbf{x}_i)$ and of $\boldsymbol{\theta}$ (the dependence on $\mathbf{y}$ is left implicit). Here we use $q_{\mathbf{x}}(\mathbf{x})$ to mean the set $\{q_{\mathbf{x}_i}(\mathbf{x}_i)\}_{i=1}^{n}$. Defining the *energy* of a global configuration $(\mathbf{x}, \mathbf{y})$ to be $-\ln p(\mathbf{x}, \mathbf{y}\,|\,\boldsymbol{\theta})$, the lower bound $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}) \leq \mathcal{L}(\boldsymbol{\theta})$ is the negative of a quantity known in statistical physics as the *free energy*: the expected energy under $q_{\mathbf{x}}(\mathbf{x})$ minus the entropy of $q_{\mathbf{x}}(\mathbf{x})$ (Feynman, 1972; Neal and Hinton, 1998).

2.2.2 EM for unconstrained (exact) optimisation

The Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977) alternates between an E step, which infers posterior distributions over hidden variables given a current parameter setting, and an M step, which maximises $\mathcal{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ given the statistics gathered from the E step. Such a set of updates can be derived using the lower bound: at each iteration, the E step maximises $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ with respect to each of the $q_{\mathbf{x}_i}(\mathbf{x}_i)$, and the M step does so with respect to $\boldsymbol{\theta}$. Mathematically speaking, using a superscript (t) to denote iteration number, starting from some initial parameters $\boldsymbol{\theta}^{(0)}$, the update equations would be:

$$\textbf{E step:} \qquad q_{\mathbf{x}_i}^{(t+1)} \leftarrow \arg\max_{q_{\mathbf{x}_i}} \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}^{(t)}) \; , \quad \forall \, i \in \{1, \ldots, n\} \; , \tag{2.17}$$

$$\textbf{M step:} \qquad \boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), \boldsymbol{\theta}) \; . \tag{2.18}$$

For the E step, it turns out that the maximum over $q_{\mathbf{x}_i}(\mathbf{x}_i)$ of the bound (2.14) is obtained by setting

$$q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) = p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \; , \quad \forall \, i \; , \tag{2.19}$$

at which point the bound becomes an equality. This can be proven by direct substitution of (2.19) into (2.14):

$$\begin{aligned}
\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), \boldsymbol{\theta}^{(t)}) &= \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) \ln \frac{p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)})}{q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i)} && (2.20)\\
&= \sum_i \int d\mathbf{x}_i \; p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \ln \frac{p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)})}{p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)})} && (2.21)\\
&= \sum_i \int d\mathbf{x}_i \; p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \ln \frac{p(\mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)}) \, p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)})}{p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)})} && (2.22)\\
&= \sum_i \int d\mathbf{x}_i \; p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \ln p(\mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)}) && (2.23)\\
&= \sum_i \ln p(\mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)}) \; , && (2.24)
\end{aligned}$$

where the last line follows as $\ln p(\mathbf{y}_i \,|\, \boldsymbol{\theta})$ is not a function of $\mathbf{x}_i$. After this E step the bound is tight. The same result can be obtained by functionally differentiating $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ with respect to $q_{\mathbf{x}_i}(\mathbf{x}_i)$, and setting to zero, subject to the normalisation constraints:

$$\int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) = 1 \; , \quad \forall \, i \; . \tag{2.25}$$

The constraints on each $q_{\mathbf{x}_i}(\mathbf{x}_i)$ can be implemented using Lagrange multipliers $\{\lambda_i\}_{i=1}^n$, forming the new functional:

$$\tilde{\mathcal{F}}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}) = \mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}) + \sum_i \lambda_i \left[\int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) - 1 \right] . \tag{2.26}$$

We then take the functional derivative of this expression with respect to each $q_{\mathbf{x}_i}(\mathbf{x}_i)$ and equate to zero, obtaining the following

$$\frac{\partial}{\partial q_{\mathbf{x}_i}(\mathbf{x}_i)} \tilde{\mathcal{F}}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}^{(t)}) = \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)}) - \ln q_{\mathbf{x}_i}(\mathbf{x}_i) - 1 + \lambda_i = 0 \tag{2.27}$$

$$\implies q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) = \exp\left(-1 + \lambda_i\right) p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)}) \tag{2.28}$$

$$= p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \; , \quad \forall \, i \; , \tag{2.29}$$

where each λ_i is related to the normalisation constant:

$$\lambda_i = 1 - \ln \int d\mathbf{x}_i \; p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}^{(t)}) \; , \quad \forall \, i \; . \tag{2.30}$$

In the remaining derivations in this thesis we always enforce normalisation constraints using Lagrange multiplier terms, although they may not always be explicitly written.

The M step is achieved by simply setting derivatives of (2.14) with respect to $\boldsymbol{\theta}$ to zero, which is the same as optimising the expected energy term in (2.15) since the entropy of the hidden state distribution $q_{\mathbf{x}}(\mathbf{x})$ is not a function of $\boldsymbol{\theta}$:

$$\textbf{M step:} \qquad \boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_i \int d\mathbf{x}_i \; p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \; . \tag{2.31}$$

Note that the optimisation is over the second $\boldsymbol{\theta}$ in the integrand, whilst holding $p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)})$ fixed. Since $\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), \boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)})$ at the beginning of each M step, and since the E step does not change the parameters, the likelihood is guaranteed not to decrease after each combined EM step. This is the well known lower bound interpretation of EM: $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ is an auxiliary function which lower bounds $\mathcal{L}(\boldsymbol{\theta})$ for any $q_{\mathbf{x}}(\mathbf{x})$, attaining equality after each E step. These steps are shown schematically in figure 2.1. Here we have expressed the E step as obtaining the full distribution over the hidden variables for each data point. However we note that, in general, the M step may require only a few statistics of the hidden variables, so only these need be computed in the E step.

2.2.3 EM with constrained (approximate) optimisation

Unfortunately, in many interesting models the data are explained by multiple interacting hidden variables which can result in intractable posterior distributions (Williams and Hinton, 1991;

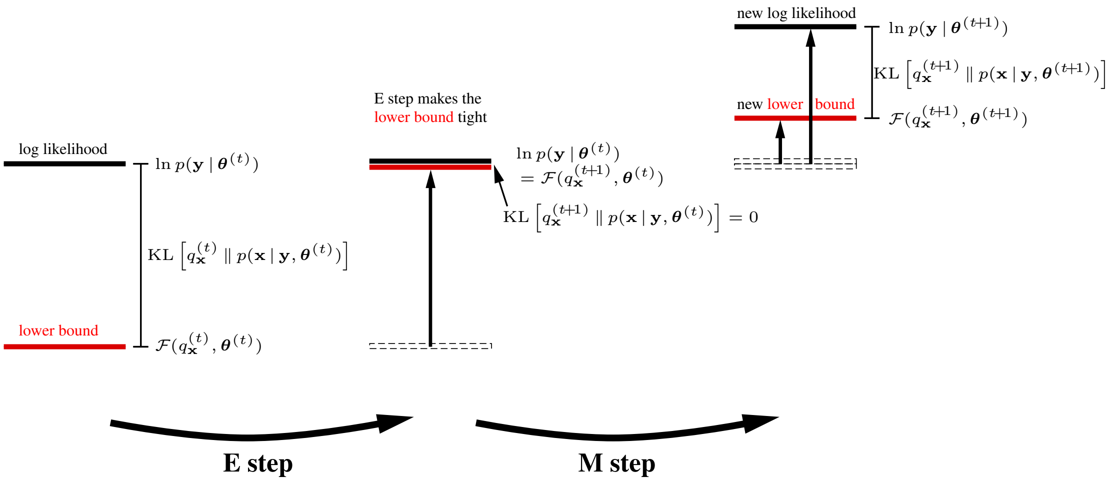

log likelihood
$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}^{(t)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t)} \,\|\, p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(t)})\right]$
lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t)}, \boldsymbol{\theta}^{(t)})$
E step makes the lower bound tight
$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}^{(t)}) = \mathcal{F}(q_{\mathbf{x}}^{(t+1)}, \boldsymbol{\theta}^{(t)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} \,\|\, p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(t)})\right] = 0$
new log likelihood
$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}^{(t+1)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} \,\|\, p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(t+1)})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$
E step
M step

Figure 2.1: The variational interpretation of EM for maximum likelihood learning. In the E step the hidden variable variational posterior is set to the exact posterior $p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(t)})$, making the bound tight. In the M step the parameters are set to maximise the lower bound $\mathcal{F}(q_{\mathbf{x}}^{(t+1)}, \boldsymbol{\theta})$ while holding the distribution over hidden variables $q_{\mathbf{x}}^{(t+1)}(\mathbf{x})$ fixed.

Neal, 1992; Hinton and Zemel, 1994; Ghahramani and Jordan, 1997; Ghahramani and Hinton, 2000). In the variational approach we can constrain the posterior distributions to be of a particular tractable form, for example factorised over the variable $\mathbf{x}_i = \{\mathbf{x}_{ij}\}_{j=1}^{|\mathbf{x}_i|}$. Using calculus of variations we can still optimise $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ as a functional of constrained distributions $q_{\mathbf{x}_i}(\mathbf{x}_i)$. The M step, which optimises $\boldsymbol{\theta}$, is conceptually identical to that described in the previous subsection, except that it is based on sufficient statistics calculated with respect to the constrained posterior $q_{\mathbf{x}_i}(\mathbf{x}_i)$ instead of the exact posterior.

We can write the lower bound $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ as

$$\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}) = \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln \frac{p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta})}{q_{\mathbf{x}_i}(\mathbf{x}_i)} \tag{2.32}$$

$$= \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) + \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln \frac{p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta})}{q_{\mathbf{x}_i}(\mathbf{x}_i)} \tag{2.33}$$

$$= \sum_i \ln p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) - \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln \frac{q_{\mathbf{x}_i}(\mathbf{x}_i)}{p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta})} \; . \tag{2.34}$$

Thus in the E step, maximising $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ with respect to $q_{\mathbf{x}_i}(\mathbf{x}_i)$ is equivalent to minimising the following quantity

$$\int d\mathbf{x}_i \; q_{\mathbf{x}_i}(\mathbf{x}_i) \ln \frac{q_{\mathbf{x}_i}(\mathbf{x}_i)}{p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta})} \equiv \mathrm{KL}\left[q_{\mathbf{x}_i}(\mathbf{x}_i) \,\|\, p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta})\right] \tag{2.35}$$

$$\geq 0 \; , \tag{2.36}$$

which is the Kullback-Leibler divergence between the variational distribution $q_{\mathbf{x}_i}(\mathbf{x}_i)$ and the exact hidden variable posterior $p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta})$. As is shown in figure 2.2, the E step does not

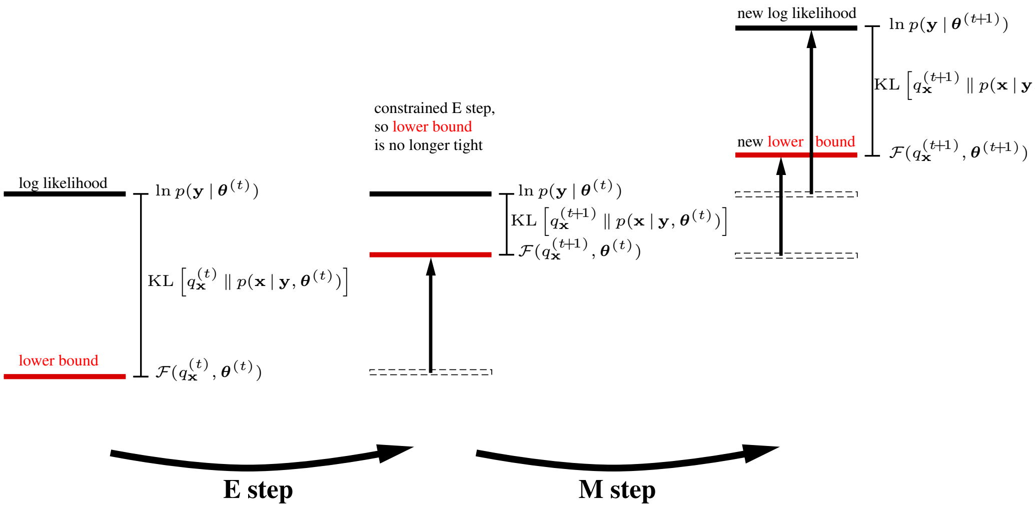

log likelihood
$\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t)} \,\|\, p(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\theta}^{(t)})\right]$
lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t)},\boldsymbol{\theta}^{(t)})$
constrained E step,
so lower bound
is no longer tight
$\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} \,\|\, p(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\theta}^{(t)})\right]$
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)},\boldsymbol{\theta}^{(t)})$
new log likelihood
$\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t+1)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} \,\|\, p(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\theta}^{(t+1)})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)},\boldsymbol{\theta}^{(t+1)})$
E step
M step

Figure 2.2: The variational interpretation of constrained EM for maximum likelihood learning. In the E step the hidden variable variational posterior is set to that which minimises $\mathrm{KL}\left[q_{\mathbf{x}}(\mathbf{x}) \,\|\, p(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\theta}^{(t)})\right]$, subject to $q_{\mathbf{x}}(\mathbf{x})$ lying in the family of constrained distributions. In the M step the parameters are set to maximise the lower bound $\mathcal{F}(q_{\mathbf{x}}^{(t+1)},\boldsymbol{\theta})$ given the current distribution over hidden variables.

generally result in the bound becoming an equality, unless of course the exact posterior lies in the family of constrained posteriors $q_{\mathbf{x}}(\mathbf{x})$.

The M step looks very similar to (2.31), but is based on the current variational posterior over hidden variables:

$$\textbf{M step:} \qquad \boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \, . \tag{2.37}$$

One can choose $q_{\mathbf{x}_i}(\mathbf{x}_i)$ to be in a particular parameterised family:

$$q_{\mathbf{x}_i}(\mathbf{x}_i) = q_{\mathbf{x}_i}(\mathbf{x}_i \,|\, \boldsymbol{\lambda}_i) \tag{2.38}$$

where $\boldsymbol{\lambda}_i = \{\boldsymbol{\lambda}_{i1}, \ldots, \boldsymbol{\lambda}_{ir}\}$ are r *variational parameters* for each datum. If we constrain each $q_{\mathbf{x}_i}(\mathbf{x}_i \,|\, \boldsymbol{\lambda}_i)$ to have easily computable moments (e.g. a Gaussian), and especially if $\ln p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta})$ is polynomial in $\mathbf{x}_i$, then we can compute the KL divergence up to a constant and, more importantly, can take its derivatives with respect to the set of variational parameters $\boldsymbol{\lambda}_i$ of each $q_{\mathbf{x}_i}(\mathbf{x}_i)$ distribution to perform the constrained E step. The E step of the *variational EM* algorithm therefore consists of a sub-loop in which each of the $q_{\mathbf{x}_i}(\mathbf{x}_i \,|\, \boldsymbol{\lambda}_i)$ is optimised by taking derivatives with respect to each $\boldsymbol{\lambda}_{is}$, for $s = 1, \ldots, r$.

The mean field approximation

The *mean field* approximation is the case in which each $q_{\mathbf{x}_i}(\mathbf{x}_i)$ is fully factorised over the hidden variables:

$$q_{\mathbf{x}_i}(\mathbf{x}_i) = \prod_{j=1}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \; . \tag{2.39}$$

In this case the expression for $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta})$ given by (2.32) becomes:

$$\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), \boldsymbol{\theta}) = \sum_i \int d\mathbf{x}_i \left[\prod_{j=1}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) - \prod_{j=1}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \ln \prod_{j=1}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \right] \tag{2.40}$$

$$= \sum_i \int d\mathbf{x}_i \left[\prod_{j=1}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) - \sum_{j=1}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \ln q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) \right] \; . \tag{2.41}$$

Using a Lagrange multiplier to enforce normalisation of the each of the approximate posteriors, we take the functional derivative of this form with respect to each $q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij})$ and equate to zero, obtaining:

$$q_{\mathbf{x}_{ij}}(\mathbf{x}_{ij}) = \frac{1}{\mathcal{Z}_{ij}} \exp \left[\int d\mathbf{x}_{i/j} \prod_{j'/j}^{|\mathbf{x}_i|} q_{\mathbf{x}_{ij'}}(\mathbf{x}_{ij'}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right] \; , \tag{2.42}$$

for each data point $i \in \{1, \ldots, n\}$, and each variational factorised component $j \in \{1, \ldots, |\mathbf{x}_i|\}$. We use the notation $d\mathbf{x}_{i/j}$ to denote the element of integration for all items in $\mathbf{x}_i$ except $\mathbf{x}_{ij}$, and the notation $\prod_{j'/j}$ to denote a product of all terms excluding j. For the ith datum, it is clear that the update equation (2.42) applied to each hidden variable j in turn represents a set of coupled equations for the approximate posterior over each hidden variable. These fixed point equations are called *mean-field equations* by analogy to such methods in statistical physics. Examples of these variational approximations can be found in the following: Ghahramani (1995); Saul et al. (1996); Jaakkola (1997); Ghahramani and Jordan (1997).

EM for maximum a posteriori learning

In MAP learning the parameter optimisation includes prior information about the parameters $p(\boldsymbol{\theta})$, and the M step seeks to find

$$\boldsymbol{\theta}_{\text{MAP}} \equiv \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta}) \; . \tag{2.43}$$

In the case of an exact E step, the M step is simply augmented to:

$$\textbf{M step:} \qquad \boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \left[\ln p(\boldsymbol{\theta}) + \sum_i \int d\mathbf{x}_i \; p(\mathbf{x}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right] \; . \tag{2.44}$$

In the case of a constrained approximate E step, the M step is given by

$$\textbf{M step:} \qquad \boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \left[\ln p(\boldsymbol{\theta}) + \sum_i \int d\mathbf{x}_i \; q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right] \; . \tag{2.45}$$

However, as mentioned in section 1.3.1, we reiterate that an undesirable feature of MAP estimation is that it is inherently basis-dependent: it is always possible to find a basis in which any particular $\boldsymbol{\theta}^*$ is the MAP solution, provided $\boldsymbol{\theta}^*$ has non-zero prior probability.

2.3 Variational methods for Bayesian learning

In this section we show how to extend the above treatment to use variational methods to approximate the integrals required for Bayesian learning. By treating the parameters as unknown quantities as well as the hidden variables, there are now correlations between the parameters and hidden variables in the posterior. The basic idea in the VB framework is to approximate the distribution over both hidden variables and parameters with a simpler distribution, usually one which assumes that the hidden states and parameters are independent given the data.

There are two main goals in Bayesian learning. The first is approximating the marginal likelihood $p(\mathbf{y} \,|\, m)$ in order to perform model comparison. The second is approximating the posterior distribution over the parameters of a model $p(\boldsymbol{\theta} \,|\, \mathbf{y}, m)$, which can then be used for prediction.

2.3.1 Deriving the learning rules

As before, let $\mathbf{y}$ denote the observed variables, $\mathbf{x}$ denote the hidden variables, and $\boldsymbol{\theta}$ denote the parameters. We assume a prior distribution over parameters $p(\boldsymbol{\theta} \,|\, m)$ conditional on the model m. The marginal likelihood of a model, $p(\mathbf{y} \,|\, m)$, can be lower bounded by introducing any distribution over both latent variables and parameters which has support where $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$ does, by appealing to Jensen's inequality once more:

$$\begin{aligned} \ln p(\mathbf{y} \,|\, m) &= \ln \int d\boldsymbol{\theta}\, d\mathbf{x}\; p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m) && (2.46) \\ &= \ln \int d\boldsymbol{\theta}\, d\mathbf{x}\; q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})} && (2.47) \\ &\geq \int d\boldsymbol{\theta}\, d\mathbf{x}\; q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})} \;. && (2.48) \end{aligned}$$

Maximising this lower bound with respect to the free distribution $q(\mathbf{x}, \boldsymbol{\theta})$ results in $q(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$ which when substituted above turns the inequality into an equality (in exact analogy with (2.19)). This does not simplify the problem since evaluating the exact posterior distribution $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$ requires knowing its normalising constant, the marginal likelihood. Instead we constrain the posterior to be a simpler, factorised (separable) approximation to $q(\mathbf{x}, \boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$\begin{aligned} \ln p(\mathbf{y} \,|\, m) &\geq \int d\boldsymbol{\theta}\, d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} && (2.49) \\ &= \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)}{q_{\mathbf{x}}(\mathbf{x})} + \ln \frac{p(\boldsymbol{\theta} \,|\, m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right] && (2.50) \\ &= \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) && (2.51) \\ &= \mathcal{F}_m(q_{\mathbf{x}_1}(\mathbf{x}_1), \ldots, q_{\mathbf{x}_n}(\mathbf{x}_n), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \;, && (2.52) \end{aligned}$$

where the last equality is a consequence of the data $\mathbf{y}$ arriving i.i.d. (this is shown in theorem 2.1 below). The quantity $\mathcal{F}_m$ is a functional of the free distributions, $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

The variational Bayesian algorithm iteratively maximises $\mathcal{F}_m$ in (2.51) with respect to the free distributions, $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, which is essentially coordinate ascent in the function space of variational distributions. The following very general theorem provides the update equations for variational Bayesian learning.

Theorem 2.1: Variational Bayesian EM (VBEM).
Let m be a model with parameters $\boldsymbol{\theta}$ giving rise to an i.i.d. data set $\mathbf{y} = \{\mathbf{y}_1, \ldots \mathbf{y}_n\}$ with corresponding hidden variables $\mathbf{x} = \{\mathbf{x}_1, \ldots \mathbf{x}_n\}$. A lower bound on the model log marginal likelihood is

$$\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta}\, d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \qquad (2.53)$$

and this can be iteratively optimised by performing the following updates, using superscript (t) to denote iteration number:

$$\textbf{VBE step:} \quad q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) = \frac{1}{\mathcal{Z}_{\mathbf{x}_i}} \exp \left[\int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \,\ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m) \right] \quad \forall\, i \qquad (2.54)$$

where

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) \ , \tag{2.55}$$

and

$$\textbf{VBM step:} \qquad q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} \, p(\boldsymbol{\theta} \,|\, m) \, \exp\left[\int d\mathbf{x} \; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \, \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)\right] \ . \tag{2.56}$$

Moreover, the update rules converge to a local maximum of $\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}))$.

Proof of $q_{\mathbf{x}_i}(\mathbf{x}_i)$ update: using variational calculus.

Take functional derivatives of $\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}))$ with respect to $q_{\mathbf{x}}(\mathbf{x})$, and equate to zero:

$$\begin{aligned}
\frac{\partial}{\partial q_{\mathbf{x}}(\mathbf{x})} \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) &= \int d\boldsymbol{\theta} \; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\frac{\partial}{\partial q_{\mathbf{x}}(\mathbf{x})} \int d\mathbf{x} \; q_{\mathbf{x}}(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)}{q_{\mathbf{x}}(\mathbf{x})}\right] && (2.57) \\
&= \int d\boldsymbol{\theta} \; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m) - \ln q_{\mathbf{x}}(\mathbf{x}) - 1\right] && (2.58) \\
&= 0 && (2.59)
\end{aligned}$$

which implies

$$\ln q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \int d\boldsymbol{\theta} \; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \, \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m) - \ln \mathcal{Z}_{\mathbf{x}}^{(t+1)} \ , \tag{2.60}$$

where $\mathcal{Z}_{\mathbf{x}}$ is a normalisation constant (from a Lagrange multiplier term enforcing normalisation of $q_{\mathbf{x}}(\mathbf{x})$, omitted for brevity). As a consequence of the i.i.d. assumption, this update can be broken down across the n data points

$$\ln q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \int d\boldsymbol{\theta} \; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \, \sum_{i=1}^{n} \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m) - \ln \mathcal{Z}_{\mathbf{x}}^{(t+1)} \ , \tag{2.61}$$

which implies that the optimal $q_{\mathbf{x}}^{(t+1)}(\mathbf{x})$ is factorised in the form $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i)$, with

$$\ln q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) = \int d\boldsymbol{\theta} \; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m) - \ln \mathcal{Z}_{\mathbf{x}_i}^{(t+1)} \quad \forall \, i \ , \tag{2.62}$$

$$\text{with} \qquad \mathcal{Z}_{\mathbf{x}} = \prod_{i=1}^{n} \mathcal{Z}_{\mathbf{x}_i} \ . \tag{2.63}$$

Thus for a given $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, there is a unique stationary point for each $q_{\mathbf{x}_i}(\mathbf{x}_i)$. □

Proof of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ update: using variational calculus.

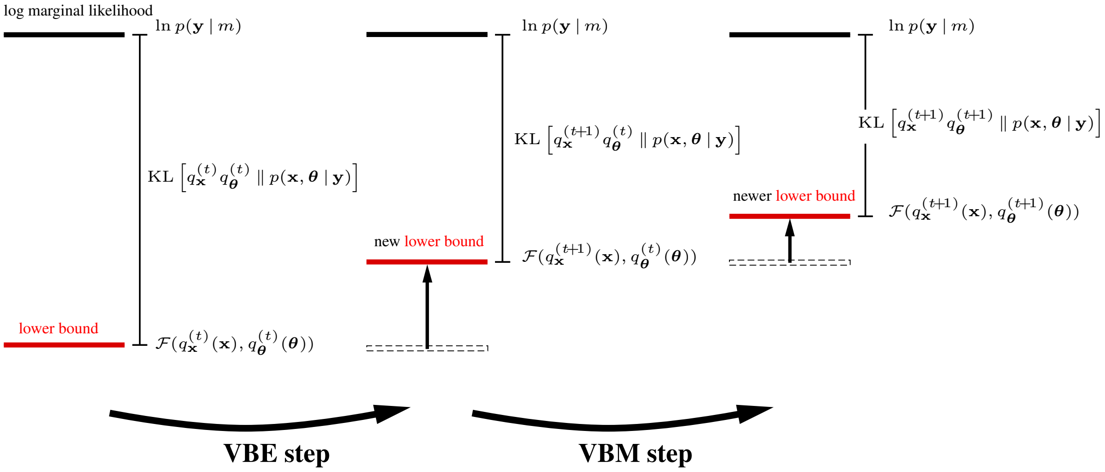

log marginal likelihood
$\ln p(\mathbf{y}\,|\,m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t)} q_{\boldsymbol{\theta}}^{(t)} \,\|\, p(\mathbf{x},\boldsymbol{\theta}\,|\,\mathbf{y})\right]$
lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}))$
$\ln p(\mathbf{y}\,|\,m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} q_{\boldsymbol{\theta}}^{(t)} \,\|\, p(\mathbf{x},\boldsymbol{\theta}\,|\,\mathbf{y})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}))$
$\ln p(\mathbf{y}\,|\,m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} q_{\boldsymbol{\theta}}^{(t+1)} \,\|\, p(\mathbf{x},\boldsymbol{\theta}\,|\,\mathbf{y})\right]$
newer lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}))$
VBE step
VBM step

Figure 2.3: The variational Bayesian EM (VBEM) algorithm. In the VBE step, the variational posterior over hidden variables $q_{\mathbf{x}}(\mathbf{x})$ is set according to (2.60). In the VBM step, the variational posterior over parameters is set according to (2.56). Each step is guaranteed to increase (or leave unchanged) the lower bound on the marginal likelihood. (Note that the exact log marginal likelihood is a *fixed* quantity, and does not change with VBE or VBM steps — it is only the lower bound which increases.)

Proceeding as above, take functional derivatives of $\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}))$ with respect to $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and equate to zero yielding:

$$\frac{\partial}{\partial q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \frac{\partial}{\partial q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x})\; \ln p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}, m) \right. \tag{2.64}$$

$$\left. + \ln \frac{p(\boldsymbol{\theta}\,|\,m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right] \tag{2.65}$$

$$= \int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x})\; \ln p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}\,|\,m) - \ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + c' \tag{2.66}$$

$$= 0\,, \tag{2.67}$$

which upon rearrangement produces

$$\ln q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta}\,|\,m) + \int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}) - \ln \mathcal{Z}_{\boldsymbol{\theta}}^{(t+1)}\,, \tag{2.68}$$

where $\mathcal{Z}_{\boldsymbol{\theta}}$ is the normalisation constant (related to the Lagrange multiplier which has again been omitted for succinctness). Thus for a given $q_{\mathbf{x}}(\mathbf{x})$, there is a unique stationary point for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

$\square$

At this point it is well worth noting the symmetry between the hidden variables and the parameters. The individual VBE steps can be written as one batch VBE step:

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \frac{1}{\mathcal{Z}_{\mathbf{x}}} \exp\left[\int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)\right] \tag{2.69}$$

$$\text{with} \qquad \mathcal{Z}_{\mathbf{x}} = \prod_{i=1}^{n} \mathcal{Z}_{\mathbf{x}_i} \; . \tag{2.70}$$

On the surface, it seems that the variational update rules (2.60) and (2.56) differ only in the prior term $p(\boldsymbol{\theta} \,|\, m)$ over the parameters. There actually also exists a prior term over the hidden variables as part of $p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)$, so this does not resolve the two. The distinguishing feature between hidden variables and parameters is that the number of hidden variables increases with data set size, whereas the number of parameters is assumed fixed.

Re-writing (2.53), it is easy to see that maximising $\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is simply equivalent to minimising the KL divergence between $q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and the joint posterior over hidden states and parameters $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$:

$$\ln p(\mathbf{y} \,|\, m) - \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta}\, d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x})\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{q_{\mathbf{x}}(\mathbf{x})\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)} \tag{2.71}$$

$$= \mathrm{KL}\left[q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \,\|\, p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)\right] \tag{2.72}$$

$$\geq 0 \; . \tag{2.73}$$

Note the similarity between expressions (2.35) and (2.72): while we minimise the former with respect to hidden variable distributions and the parameters, the latter we minimise with respect to the hidden variable distribution and a *distribution* over parameters.

The variational Bayesian EM algorithm reduces to the ordinary EM algorithm for ML estimation if we restrict the parameter distribution to a point estimate, i.e. a Dirac delta function, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, in which case the M step simply involves re-estimating $\boldsymbol{\theta}^*$. Note that the same cannot be said in the case of MAP estimation, which is inherently basis dependent, unlike both VB and ML algorithms. By construction, the VBEM algorithm is guaranteed to monotonically increase an objective function $\mathcal{F}$, as a function of a distribution over parameters and hidden variables. Since we integrate over model parameters there is a naturally incorporated model complexity penalty. It turns out that for a large class of models (see section 2.4) the VBE step has approximately the same computational complexity as the standard E step in the ML framework, which makes it viable as a Bayesian replacement for the EM algorithm.

2.3.2 Discussion

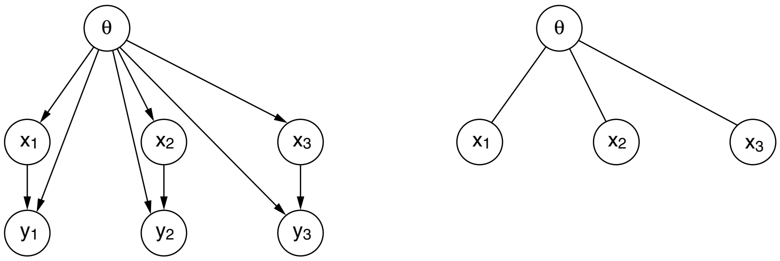
The impact of the $q(\mathbf{x}, \boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ factorisation

Unless we make the assumption that the posterior over parameters and hidden variables factorises, we will not generally obtain the further hidden variable factorisation over n that we have in equation (2.55). In that case, the distributions of $\mathbf{x}_i$ and $\mathbf{x}_j$ will be coupled for all cases $\{i, j\}$ in the data set, greatly increasing the overall computational complexity of inference. This further factorisation is depicted in figure 2.4 for the case of $n = 3$, where we see: (a) the original directed graphical model, where $\boldsymbol{\theta}$ is the collection of parameters governing prior distributions over the hidden variables $\mathbf{x}_i$ and the conditional probability $p(\mathbf{y}_i \,|\, \mathbf{x}_i, \boldsymbol{\theta})$; (b) the moralised graph given the data $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$, which shows that the hidden variables are now dependent in the posterior through the uncertain parameters; (c) the effective graph after the factorisation assumption, which not only removes arcs between the parameters and hidden variables, but also removes the dependencies between the hidden variables. This latter independence falls out from the optimisation as a result of the i.i.d. nature of the data, and is not a further approximation.

Whilst this factorisation of the posterior distribution over hidden variables and parameters may seem drastic, one can think of it as replacing *stochastic* dependencies between $\mathbf{x}$ and $\boldsymbol{\theta}$ with *deterministic* dependencies between relevant moments of the two sets of variables. The advantage of ignoring how fluctuations in $\mathbf{x}$ induce fluctuations in $\boldsymbol{\theta}$ (and vice-versa) is that we can obtain analytical approximations to the log marginal likelihood. It is these same ideas that underlie mean-field approximations from statistical physics, from where these lower-bounding variational approximations were inspired (Feynman, 1972; Parisi, 1988). In later chapters the consequences of the factorisation for particular models are studied in some detail; in particular we will use sampling methods to estimate by how much the variational bound falls short of the marginal likelihood.

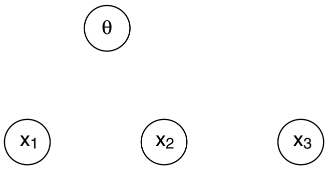
What forms for $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$?

One might need to approximate the posterior further than simply the hidden-variable / parameter factorisation. A common reason for this is that the parameter posterior may still be intractable despite the hidden-variable / parameter factorisation. The free-form extremisation of $\mathcal{F}$ normally provides us with a functional form for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, but this may be unwieldy; we therefore need to assume some simpler space of parameter posteriors. The most commonly used distributions are those with just a few sufficient statistics, such as the Gaussian or Dirichlet distributions. Taking a Gaussian example, $\mathcal{F}$ is then explicitly extremised with respect to a set of variational parameters $\boldsymbol{\zeta}_{\boldsymbol{\theta}} = (\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\boldsymbol{\theta}})$ which parameterise the Gaussian $q_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\zeta}_{\boldsymbol{\theta}})$. We will see examples of this approach in later chapters. There may also exist intractabilities in the hidden variable



(a) The generative graphical model.

(b) Graph representing the exact posterior.



(c) Posterior graph after the variational approximation.

Figure 2.4: Graphical depiction of the hidden-variable / parameter factorisation. **(a)** The original generative model for $n = 3$. **(b)** The exact posterior graph given the data. Note that for all case pairs $\{i, j\}$, $\mathbf{x}_i$ and $\mathbf{x}_j$ are not directly coupled, but interact through $\boldsymbol{\theta}$. That is to say all the hidden variables are conditionally independent of one another, but only given the parameters. **(c)** the posterior graph after the variational approximation between parameters and hidden variables, which removes arcs between parameters and hidden variables. Note that, on assuming this factorisation, as a consequence of the i.i.d. assumption the hidden variables become independent.

posterior, for which further approximations need be made (some examples are mentioned below).

There is something of a dark art in discovering a factorisation amongst the hidden variables and parameters such that the approximation remains faithful at an 'acceptable' level. Of course it does not make sense to use a posterior form which holds fewer conditional independencies than those implied by the *moral* graph (see section 1.1). The key to a good variational approximation is then to remove as few arcs as possible from the moral graph such that inference becomes tractable. In many cases the goal is to find tractable substructures (*structured* approximations) such as trees or mixtures of trees, which capture as many of the arcs as possible. Some arcs may capture crucial dependencies between nodes and so need be kept, whereas other arcs might induce a weak local correlation at the expense of a long-range correlation which to first order can be ignored; removing such an arc can have dramatic effects on the tractability.

The advantage of the variational Bayesian procedure is that *any* factorisation of the posterior yields a lower bound on the marginal likelihood. Thus in practice it may pay to approximately evaluate the computational cost of several candidate factorisations, and implement those which can return a completed optimisation of $\mathcal{F}$ within a certain amount of computer time. One would expect the more complex factorisations to take more computer time but also yield progressively tighter lower bounds on average, the consequence being that the marginal likelihood estimate improves over time. An interesting avenue of research in this vein would be to use the variational posterior resulting from a simpler factorisation as the initialisation for a slightly more complicated factorisation, and move in a chain from simple to complicated factorisations to help avoid local free energy minima in the optimisation. Having proposed this, it remains to be seen if it is possible to form a coherent closely-spaced chain of distributions that are of any use, as compared to starting from the fullest posterior approximation from the start.

Using the lower bound for model selection and averaging

The log ratio of posterior probabilities of two competing models m and m' is given by

$$\ln \frac{p(m \,|\, \mathbf{y})}{p(m' \,|\, \mathbf{y})} = + \ln p(m) + p(\mathbf{y} \,|\, m) - \ln p(m') - \ln p(\mathbf{y} \,|\, m') \tag{2.74}$$

$$\begin{aligned} &= + \ln p(m) \; + \mathcal{F}(q_{\mathbf{x},\boldsymbol{\theta}}) \; + \mathrm{KL}\left[q(\mathbf{x},\boldsymbol{\theta}) \,\|\, p(\mathbf{x},\boldsymbol{\theta} \,|\, \mathbf{y}, m)\right] \\ &\quad - \ln p(m') - \mathcal{F}'(q'_{\mathbf{x},\boldsymbol{\theta}}) - \mathrm{KL}\left[q'(\mathbf{x},\boldsymbol{\theta}) \,\|\, p(\mathbf{x},\boldsymbol{\theta} \,|\, \mathbf{y}, m')\right] \end{aligned} \tag{2.75}$$

where we have used the form in (2.72), which is exact regardless of the quality of the bound used, or how tightly that bound has been optimised. The lower bounds for the two models, $\mathcal{F}$ and $\mathcal{F}'$, are calculated from VBEM optimisations, providing us for each model with an approximation to the posterior over the hidden variables and parameters of that model, $q_{\mathbf{x},\boldsymbol{\theta}}$ and $q'_{\mathbf{x},\boldsymbol{\theta}}$; these may in general be functionally very different (we leave aside for the moment local maxima problems in the optimisation process which can be overcome to an extent by using several differently initialised optimisations or in some models by employing heuristics tailored to exploit the model structure). When we perform model selection by comparing the lower bounds, $\mathcal{F}$ and $\mathcal{F}'$, we are assuming that the KL divergences in the two approximations are the same, so that we can use just these lower bounds as guide. Unfortunately it is non-trivial to predict how tight in theory any particular bound can be — if this were possible we could more accurately estimate the marginal likelihood from the start.

Taking an example, we would like to know whether the bound for a model with S mixture components is similar to that for $S+1$ components, and if not then how badly this inconsistency affects the posterior over this set of models. Roughly speaking, let us assume that every component in our model contributes a (constant) KL divergence penalty of KL_s. For clarity we use the notation $\mathcal{L}(S)$ and $\mathcal{F}(S)$ to denote the exact log marginal likelihood and lower bounds, respectively, for a model with S components. The difference in log marginal likelihoods, $\mathcal{L}(S+1) - \mathcal{L}(S)$, is the quantity we wish to estimate, but if we base this on the lower bounds the difference becomes

$$\begin{aligned}\mathcal{L}(S+1) - \mathcal{L}(S) &= [\mathcal{F}(S+1) + (S+1)\,\mathrm{KL}_s] - [\mathcal{F}(S) + S\,\mathrm{KL}_s] && (2.76)\\ &= \mathcal{F}(S+1) - \mathcal{F}(S) + \mathrm{KL}_s && (2.77)\\ &\neq \mathcal{F}(S+1) - \mathcal{F}(S)\ , && (2.78)\end{aligned}$$

where the last line is the result we would have basing the difference on lower bounds. Therefore there exists a systematic error when comparing models if each component contributes independently to the KL divergence term. Since the KL divergence is strictly positive, and we are basing our model selection on (2.78) rather than (2.77), this analysis suggests that there is a systematic bias towards simpler models. We will in fact see this in chapter 4, where we find an importance sampling estimate of the KL divergence showing this behaviour.

Optimising the prior distributions

Usually the parameter priors are functions of hyperparameters, $\mathbf{a}$, so we can write $p(\boldsymbol{\theta} \,|\, \mathbf{a}, m)$. In the variational Bayesian framework the lower bound can be made higher by maximising $\mathcal{F}_m$ with respect to these hyperparameters:

$$\mathbf{a}^{(t+1)} = \arg\max_{\mathbf{a}}\ \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}, \mathbf{a})\ . \qquad (2.79)$$

A simple depiction of this optimisation is given in figure 2.5. Unlike earlier in section 2.3.1, the marginal likelihood of model m can now be increased with hyperparameter optimisation. As we will see in later chapters, there are examples where these hyperparameters themselves have governing hyperpriors, such that they can be integrated over as well. The result being that

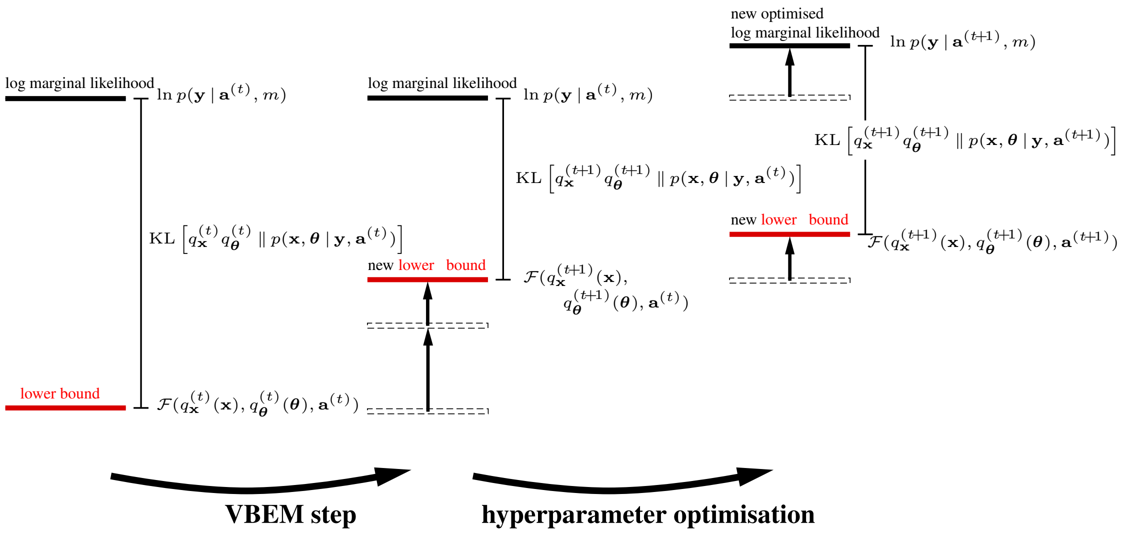

log marginal likelihood
$\ln p(\mathbf{y} \mid \mathbf{a}^{(t)}, m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t)} q_{\boldsymbol{\theta}}^{(t)} \,\|\, p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{a}^{(t)})\right]$
lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}), \mathbf{a}^{(t)})$
log marginal likelihood
$\ln p(\mathbf{y} \mid \mathbf{a}^{(t)}, m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} q_{\boldsymbol{\theta}}^{(t+1)} \,\|\, p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{a}^{(t)})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}), \mathbf{a}^{(t)})$
new optimised
log marginal likelihood
$\ln p(\mathbf{y} \mid \mathbf{a}^{(t+1)}, m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} q_{\boldsymbol{\theta}}^{(t+1)} \,\|\, p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{a}^{(t+1)})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}), \mathbf{a}^{(t+1)})$
VBEM step
hyperparameter optimisation

Figure 2.5: The variational Bayesian EM algorithm with hyperparameter optimisation. The VBEM step consists of VBE and VBM steps, as shown in figure 2.3. The hyperparameter optimisation increases the lower bound and also improves the marginal likelihood.

we can infer distributions over these as well, just as for parameters. The reason for abstracting from the parameters this far is that we would like to integrate out all variables whose cardinality increases with model complexity; this standpoint will be made clearer in the following chapters.

Previous work, and general applicability of VBEM

The variational approach for lower bounding the marginal likelihood (and similar quantities) has been explored by several researchers in the past decade, and has received a lot of attention recently in the machine learning community. It was first proposed for one-hidden layer neural networks (which have no hidden variables) by Hinton and van Camp (1993) where $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ was restricted to be Gaussian with diagonal covariance. This work was later extended to show that tractable approximations were also possible with a full covariance Gaussian (Barber and Bishop, 1998) (which in general will have the mode of the posterior at a different location than in the diagonal case). Neal and Hinton (1998) presented a generalisation of EM which made use of Jensen's inequality to allow partial E-steps; in this paper the term *ensemble learning* was used to describe the method since it fits an ensemble of models, each with its own parameters. Jaakkola (1997) and Jordan et al. (1999) review variational methods in a general context (i.e. non-Bayesian). Variational Bayesian methods have been applied to various models with hidden variables and no restrictions on $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{x}_i}(\mathbf{x}_i)$ other than the assumption that they factorise in some way (Waterhouse et al., 1996; Bishop, 1999; Ghahramani and Beal, 2000; Attias, 2000). Of particular note is the variational Bayesian HMM of MacKay (1997), in which free-form optimisations are explicitly undertaken (see chapter 3); this work was the inspiration for the examination of Conjugate-Exponential (CE) models, discussed in the next section. An example of a constrained optimisation for a logistic regression model can be found in Jaakkola and Jordan (2000).

Several researchers have investigated using mixture distributions for the approximate posterior, which allows for more flexibility whilst maintaining a degree of tractability (Lawrence et al., 1998; Bishop et al., 1998; Lawrence and Azzouzi, 1999). The lower bound in these models is a sum of a two terms: a first term which is a convex combination of bounds from each mixture component, and a second term which is the mutual information between the mixture labels and the hidden variables of the model. The first term offers no improvement over a naive combination of bounds, but the second (which is non-negative) has to improve on the simple bounds. Unfortunately this term contains an expectation over all configurations of the hidden states and so has to be itself bounded with a further use of Jensen's inequality in the form of a convex bound on the log function ($\ln(x) \leq \lambda x - \ln(\lambda) - 1$) (Jaakkola and Jordan, 1998). Despite this approximation drawback, empirical results in a handful of models have shown that the approximation does improve the simple mean field bound and improves monotonically with the number of mixture components.

A related method for approximating the integrand for Bayesian learning is based on an idea known as *assumed density filtering* (ADF) (Bernardo and Giron, 1988; Stephens, 1997; Boyen and Koller, 1998; Barber and Sollich, 2000; Frey et al., 2001), and is called the Expectation Propagation (EP) algorithm (Minka, 2001a). This algorithm approximates the integrand of interest with a set of *terms*, and through a process of repeated deletion-inclusion of term expressions, the integrand is iteratively refined to resemble the true integrand as closely as possible. Therefore the key to the method is to use terms which can be tractably integrated. This has the same flavour as the variational Bayesian method described here, where we iteratively update the approximate posterior over a hidden state $q_{\mathbf{x}_i}(\mathbf{x}_i)$ or over the parameters $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. The key difference between EP and VB is that in the update process (i.e. deletion-inclusion) EP seeks to minimise the KL divergence which averages according to the true distribution, $\mathrm{KL}\left[p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \,\|\, q(\mathbf{x}, \boldsymbol{\theta})\right]$ (which is simply a moment-matching operation for exponential family models), whereas VB seeks to minimise the KL divergence according to the approximate distribution, $\mathrm{KL}\left[q(\mathbf{x}, \boldsymbol{\theta}) \,\|\, p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y})\right]$. Therefore, EP is at least attempting to average according to the correct distribution, whereas VB has the wrong cost function at heart. However, in general the KL divergence in EP can only be minimised separately one term at a time, while the KL divergence in VB is minimised globally over all terms in the approximation. The result is that EP may still not result in representative posterior distributions (for example, see Minka, 2001a, figure 3.6, p. 6). Having said that, it may be that more generalised deletion-inclusion steps can be derived for EP, for example removing two or more terms at a time from the integrand, and this may alleviate some of the 'local' restrictions of the EP algorithm. As in VB, EP is constrained to use particular parametric families with a small number of moments for tractability. An example of EP used with an assumed Dirichlet density for the term expressions can be found in Minka and Lafferty (2002).

In the next section we take a closer look at the variational Bayesian EM equations, (2.54) and (2.56), and ask the following questions:

- To which models can we apply VBEM? i.e. which forms of data distributions $p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})$ and priors $p(\boldsymbol{\theta} \,|\, m)$ result in tractable VBEM updates?
- How does this relate formally to conventional EM?
- When can we utilise existing belief propagation algorithms in the VB framework?

2.4 Conjugate-Exponential models

2.4.1 Definition

We consider a particular class of graphical models with latent variables, which we call *conjugate-exponential* (CE) models. In this section we explicitly apply the variational Bayesian method to these parametric families, deriving a simple general form of VBEM for the class.

Conjugate-exponential models satisfy two conditions:

Condition (1). *The complete-data likelihood is in the exponential family:*

$$p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) = g(\boldsymbol{\theta})\, f(\mathbf{x}_i, \mathbf{y}_i)\, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)} \;, \tag{2.80}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of natural parameters, $\mathbf{u}$ and f are the functions that define the exponential family, and g is a normalisation constant:

$$g(\boldsymbol{\theta})^{-1} = \int d\mathbf{x}_i\, d\mathbf{y}_i\; f(\mathbf{x}_i, \mathbf{y}_i)\, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)} \;. \tag{2.81}$$

The natural parameters for an exponential family model $\boldsymbol{\phi}$ are those that interact linearly with the sufficient statistics of the data $\mathbf{u}$. For example, for a univariate Gaussian in x with mean μ and standard deviation σ, the necessary quantities are obtained from:

$$p(x \,|\, \mu, \sigma) = \exp\left\{ -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \right\} \tag{2.82}$$

$$\boldsymbol{\theta} = \big(\sigma^2, \mu\big) \tag{2.83}$$

and are:

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right) \tag{2.84}$$

$$\mathbf{u}(x) = \left(-\frac{x^2}{2}, x\right) \tag{2.85}$$

$$f(x) = 1 \tag{2.86}$$

$$g(\boldsymbol{\theta}) = \exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} . \tag{2.87}$$

Note that whilst the parameterisation for $\boldsymbol{\theta}$ is arbitrary, e.g. we could have let $\boldsymbol{\theta} = (\sigma, \mu)$, the natural parameters $\boldsymbol{\phi}$ are unique up to a multiplicative constant.

Condition (2). *The parameter prior is conjugate to the complete-data likelihood:*

$$p(\boldsymbol{\theta} \,|\, \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\, g(\boldsymbol{\theta})^{\eta}\, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^{\top}\boldsymbol{\nu}} , \tag{2.88}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior, and h is a normalisation constant:

$$h(\eta, \boldsymbol{\nu})^{-1} = \int d\boldsymbol{\theta}\; g(\boldsymbol{\theta})^{\eta}\, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^{\top}\boldsymbol{\nu}} . \tag{2.89}$$

Condition 1 (2.80) in fact usually implies the existence of a conjugate prior which satisfies condition 2 (2.88). The prior $p(\boldsymbol{\theta} \,|\, \eta, \boldsymbol{\nu})$ is said to be conjugate to the likelihood $p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta})$ if and only if the posterior

$$p(\boldsymbol{\theta} \,|\, \eta', \boldsymbol{\nu}') \propto p(\boldsymbol{\theta} \,|\, \eta, \boldsymbol{\nu}) p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) \tag{2.90}$$

is of the same parametric form as the prior. In general the exponential families are the only classes of distributions that have natural conjugate prior distributions because they are the only distributions with a fixed number of sufficient statistics apart from some irregular cases (see Gelman et al., 1995, p. 38). From the definition of conjugacy, we see that the hyperparameters of a conjugate prior can be interpreted as the number (η) and values ($\boldsymbol{\nu}$) of pseudo-observations under the corresponding likelihood.

We call models that satisfy conditions 1 (2.80) and 2 (2.88) *conjugate-exponential*.

The list of latent-variable models of practical interest with complete-data likelihoods in the exponential family is very long, for example: Gaussian mixtures, factor analysis, principal components analysis, hidden Markov models and extensions, switching state-space models, discrete-variable belief networks. Of course there are also many as yet undreamt-of models combining Gaussian, gamma, Poisson, Dirichlet, Wishart, multinomial, and other distributions in the exponential family.

However there are some notable outcasts which do not satisfy the conditions for membership of the CE family, namely: Boltzmann machines (Ackley et al., 1985), logistic regression and sigmoid belief networks (Bishop, 1995), and independent components analysis (ICA) (as presented in Comon, 1994; Bell and Sejnowski, 1995), all of which are widely used in the machine learning community. As an example let us see why logistic regression is not in the conjugate-exponential family: for $y_i \in \{-1, 1\}$, the likelihood under a logistic regression model is

$$p(y_i \,|\, \mathbf{x}_i, \boldsymbol{\theta}) = \frac{e^{y_i \boldsymbol{\theta}^\top \mathbf{x}_i}}{e^{\boldsymbol{\theta}^\top \mathbf{x}_i} + e^{-\boldsymbol{\theta}^\top \mathbf{x}_i}} \,, \tag{2.91}$$

where $\mathbf{x}_i$ is the regressor for data point i and $\boldsymbol{\theta}$ is a vector of weights, potentially including a bias. This can be rewritten as

$$p(y_i \,|\, \mathbf{x}_i, \boldsymbol{\theta}) = e^{y_i \boldsymbol{\theta}^\top \mathbf{x}_i - f(\boldsymbol{\theta}, \mathbf{x}_i)} \,, \tag{2.92}$$

where $f(\boldsymbol{\theta}, \mathbf{x}_i)$ is a normalisation constant. To belong in the exponential family the normalising constant must split into functions of only $\boldsymbol{\theta}$ and only $(\mathbf{x}_i, \mathbf{y}_i)$. Expanding $f(\boldsymbol{\theta}, \mathbf{x}_i)$ yields a series of powers of $\boldsymbol{\theta}^\top \mathbf{x}_i$, which *could* be assimilated into the $\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)$ term by augmenting the natural parameter and sufficient statistics vectors, if it were not for the fact that the series is infinite meaning that there would need to be an infinity of natural parameters. This means we cannot represent the likelihood with a finite number of sufficient statistics.

Models whose complete-data likelihood is not in the exponential family can often be approximated by models which are in the exponential family and have been given additional hidden variables. A very good example is the Independent Factor Analysis (IFA) model of Attias (1999a). In conventional ICA, one can think of the model as using non-Gaussian sources, or using Gaussian sources passed through a non-linearity to make them non-Gaussian. For most non-linearities commonly used (such as the logistic), the complete-data likelihood becomes non-CE. Attias recasts the model as a mixture of Gaussian sources being fed into a linear mixing matrix. This model is in the CE family and so can be tackled with the VB treatment. It is an open area of research to investigate how best to bring models into the CE family, such that inferences in the modified model resemble the original as closely as possible.

2.4.2 Variational Bayesian EM for CE models

In Bayesian inference we want to determine the posterior over parameters and hidden variables $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, \eta, \boldsymbol{\nu})$. In general this posterior is *neither* conjugate nor in the exponential family. In this subsection we see how the properties of the CE family make it especially amenable to the VB approximation, and derive the VBEM algorithm for CE models.

Theorem 2.2: Variational Bayesian EM for Conjugate-Exponential Models.
Given an i.i.d. data set $\mathbf{y} = \{\mathbf{y}_1, \dots \mathbf{y}_n\}$*, if the model satisfies conditions (1) and (2), then the following (a), (b) and (c) hold:*

(a) *the VBE step yields:*

$$q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}(\mathbf{x}_i) \;, \tag{2.93}$$

and $q_{\mathbf{x}_i}(\mathbf{x}_i)$ *is in the exponential family:*

$$q_{\mathbf{x}_i}(\mathbf{x}_i) \propto f(\mathbf{x}_i, \mathbf{y}_i)\, e^{\overline{\boldsymbol{\phi}}^{\top} \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)} = p(\mathbf{x}_i \,|\, \mathbf{y}_i, \overline{\boldsymbol{\phi}}) \;, \tag{2.94}$$

with a natural parameter vector

$$\overline{\boldsymbol{\phi}} = \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \boldsymbol{\phi}(\boldsymbol{\theta}) \equiv \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{2.95}$$

obtained by taking the expectation of $\boldsymbol{\phi}(\boldsymbol{\theta})$ *under* $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *(denoted using angle-brackets* $\langle \cdot \rangle$*). For invertible* $\boldsymbol{\phi}$*, defining* $\tilde{\boldsymbol{\theta}}$ *such that* $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}$*, we can rewrite the approximate posterior as*

$$q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i \,|\, \mathbf{y}_i, \tilde{\boldsymbol{\theta}}) \;. \tag{2.96}$$

(b) *the VBM step yields that* $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *is conjugate and of the form:*

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}})\, g(\boldsymbol{\theta})^{\tilde{\eta}}\, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^{\top} \tilde{\boldsymbol{\nu}}} \;, \tag{2.97}$$

where

$$\tilde{\eta} = \eta + n \;, \tag{2.98}$$

$$\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^{n} \overline{\mathbf{u}}(\mathbf{y}_i) \;, \tag{2.99}$$

and

$$\overline{\mathbf{u}}(\mathbf{y}_i) = \langle \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i) \rangle_{q_{\mathbf{x}_i}(\mathbf{x}_i)} \tag{2.100}$$

is the expectation of the sufficient statistic $\mathbf{u}$*. We have used* $\langle \cdot \rangle_{q_{\mathbf{x}_i}(\mathbf{x}_i)}$ *to denote expectation under the variational posterior over the latent variable(s) associated with the* i*th datum.*

(c) *parts (a) and (b) hold for every iteration of variational Bayesian EM.*

Proof of (a): by direct substitution.

Starting from the variational extrema solution (2.60) for the VBE step:

$$q_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\mathcal{Z}_{\mathbf{x}}} \, e^{\langle \ln p(\mathbf{x},\mathbf{y} \,|\, \boldsymbol{\theta},m) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}} \ , \tag{2.101}$$

substitute the parametric form for $p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m)$ in condition 1 (2.80), which yields (omitting iteration superscripts):

$$\begin{aligned} q_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{\mathcal{Z}_{\mathbf{x}}} \, e^{\sum_{i=1}^n \left\langle \ln g(\boldsymbol{\theta}) + \ln f(\mathbf{x}_i,\mathbf{y}_i) + \boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i,\mathbf{y}_i) \right\rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}} &(2.102)\\ &= \frac{1}{\mathcal{Z}_{\mathbf{x}}} \left[\prod_{i=1}^n f(\mathbf{x}_i, \mathbf{y}_i) \right] e^{\sum_{i=1}^n \overline{\boldsymbol{\phi}}^\top u(\mathbf{x}_i,\mathbf{y}_i)} \ , &(2.103) \end{aligned}$$

where $\mathcal{Z}_{\mathbf{x}}$ has absorbed constants independent of $\mathbf{x}$, and we have defined without loss of generality:

$$\overline{\boldsymbol{\phi}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \ . \tag{2.104}$$

If $\boldsymbol{\phi}$ is invertible, then there exists a $\tilde{\boldsymbol{\theta}}$ such that $\overline{\boldsymbol{\phi}} = \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})$, and we can rewrite (2.103) as:

$$\begin{aligned} q_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{\mathcal{Z}_{\mathbf{x}}} \left[\prod_{i=1}^n f(\mathbf{x}_i, \mathbf{y}_i) e^{\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})^\top \mathbf{u}(\mathbf{x}_i,\mathbf{y}_i)} \right] &(2.105)\\ &\propto \prod_{i=1}^n p(\mathbf{x}_i, \mathbf{y}_i \,|\, \tilde{\boldsymbol{\theta}}, m) &(2.106)\\ &= \prod_{i=1}^n q_{\mathbf{x}_i}(\mathbf{x}_i) &(2.107)\\ &= p(\mathbf{x}, \mathbf{y} \,|\, \tilde{\boldsymbol{\theta}}, m) \ . &(2.108) \end{aligned}$$

Thus the result of the approximate VBE step, which averages over the ensemble of models $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, is exactly the same as an exact E step, calculated at the *variational Bayes point* estimate $\tilde{\boldsymbol{\theta}}$. □

Proof of (b): by direct substitution.

Starting from the variational extrema solution (2.56) for the VBM step:

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} \, p(\boldsymbol{\theta} \,|\, m) \, e^{\langle \ln p(\mathbf{x},\mathbf{y} \,|\, \boldsymbol{\theta},m) \rangle_{q_{\mathbf{x}}(\mathbf{x})}} \ , \tag{2.109}$$

substitute the parametric forms for $p(\boldsymbol{\theta}\,|\,m)$ and $p(\mathbf{x}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}, m)$ as specified in conditions 2 (2.88) and 1 (2.80) respectively, which yields (omitting iteration superscripts):

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} h(\eta, \boldsymbol{\nu}) g(\boldsymbol{\theta})^{\eta} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}} \, e^{\left\langle \sum_{i=1}^n \ln g(\boldsymbol{\theta}) + \ln f(\mathbf{x}_i, \mathbf{y}_i) + \boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i) \right\rangle_{q_{\mathbf{x}}(\mathbf{x})}} \tag{2.110}$$

$$= \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} h(\eta, \boldsymbol{\nu}) g(\boldsymbol{\theta})^{\eta+n} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \left[\boldsymbol{\nu} + \sum_{i=1}^n \overline{\mathbf{u}}(\mathbf{y}_i)\right]} \underbrace{e^{\sum_{i=1}^n \langle \ln f(\mathbf{x}_i, \mathbf{y}_i) \rangle_{q_{\mathbf{x}}(\mathbf{x})}}}_{\text{has no } \boldsymbol{\theta} \text{ dependence}} \tag{2.111}$$

$$= h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}} \, , \tag{2.112}$$

where

$$h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} e^{\sum_{i=1}^n \langle \ln f(\mathbf{x}_i, \mathbf{y}_i) \rangle_{q_{\mathbf{x}}(\mathbf{x})}} \, . \tag{2.113}$$

Therefore the variational posterior $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ in (2.112) is of conjugate form, according to condition 2 (2.88). □

Proof of (c): by induction.

Assume conditions 1 (2.80) and 2 (2.88) are met (i.e. the model is in the CE family). From part (a), the VBE step produces a posterior distribution $q_{\mathbf{x}}(\mathbf{x})$ in the exponential family, preserving condition 1 (2.80); the parameter distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ remains unaltered, preserving condition 2 (2.88). From part (b), the VBM step produces a parameter posterior $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ that is of conjugate form, preserving condition 2 (2.88); $q_{\mathbf{x}}(\mathbf{x})$ remains unaltered from the VBE step, preserving condition 1 (2.80). Thus under both the VBE and VBM steps, conjugate-exponentiality is preserved, which makes the theorem applicable at every iteration of VBEM. □

As before, since $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{x}_i}(\mathbf{x}_i)$ are coupled, (2.97) and (2.94) do not provide an analytic solution to the minimisation problem, so the optimisation problem is solved numerically by iterating between the fixed point equations given by these equations. To summarise briefly:

> **VBE Step**: *Compute the expected sufficient statistics* $\{\overline{\mathbf{u}}(\mathbf{y}_i)\}_{i=1}^n$ *under the hidden variable distributions* $q_{\mathbf{x}_i}(\mathbf{x}_i)$*, for all* i*.*
>
> **VBM Step**: *Compute the expected natural parameters* $\overline{\boldsymbol{\phi}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ *under the parameter distribution given by* $\tilde{\eta}$ *and* $\tilde{\boldsymbol{\nu}}$*.*

2.4.3 Implications

In order to really understand what the conjugate-exponential formalism buys us, let us reiterate the main points of theorem 2.2 above. The first result is that in the VBM step the analytical form of the variational posterior $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ does not change during iterations of VBEM — e.g. if the posterior is Gaussian at iteration $t = 1$, then only a Gaussian need be represented at future iterations. If it were able to change, which is the case in general (theorem 2.1), the

EM for MAP estimation	**Variational Bayesian EM**
Goal: maximise $p(\boldsymbol{\theta} \mid \mathbf{y}, m)$ w.r.t. $\boldsymbol{\theta}$	**Goal:** lower bound $p(\mathbf{y} \mid m)$
E Step: compute $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)})$	**VBE Step:** compute $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} \mid \mathbf{y}, \overline{\boldsymbol{\phi}}^{(t)})$
M Step: $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \int d\mathbf{x}\, q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$	**VBM Step:** $q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp \int d\mathbf{x}\, q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$

Table 2.1: Comparison of EM for ML/MAP estimation against variational Bayesian EM for CE models.

posterior could quickly become unmanageable, and (further) approximations would be required to prevent the algorithm becoming too complicated. The second result is that the posterior over hidden variables calculated in the VBE step is exactly the posterior that would be calculated had we been performing an ML/MAP E step. That is, the inferences using an ensemble of models $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ can be represented by the effect of a point parameter, $\tilde{\boldsymbol{\theta}}$. The task of performing many inferences, each of which corresponds to a different parameter setting, can be replaced with a single inference step — it is possible to infer the hidden states in a conjugate exponential model tractably while integrating over an ensemble of model parameters.

Comparison to EM for ML/MAP parameter estimation

We can draw a tight parallel between the EM algorithm for ML/MAP estimation, and our VBEM algorithm applied specifically to conjugate-exponential models. These are summarised in table 2.1. This general result of VBEM for CE models was reported in Ghahramani and Beal (2001), and generalises the well known EM algorithm for ML estimation (Dempster et al., 1977). It is a special case of the variational Bayesian algorithm (theorem 2.1) used in Ghahramani and Beal (2000) and in Attias (2000), yet encompasses many of the models that have been so far subjected to the variational treatment. Its particular usefulness is as a guide for the design of models, to make them amenable to efficient approximate Bayesian inference.

The VBE step has about the same time complexity as the E step, and is in all ways identical except that it is re-written in terms of the expected natural parameters. In particular, we can make use of all relevant propagation algorithms such as junction tree, Kalman smoothing, or belief propagation. The VBM step computes a *distribution* over parameters (in the conjugate family) rather than a point estimate. Both ML/MAP EM and VBEM algorithms monotonically increase an objective function, but the latter also incorporates a model complexity penalty by integrating over parameters so embodying an Occam's razor effect. Several examples will be presented in the following chapters of this thesis.

Natural parameter inversions

Unfortunately, even though the algorithmic complexity is the same, the implementations may be hampered since the propagation algorithms need to be re-derived in terms of the natural parameters (this is essentially the difference between the forms in (2.94) and (2.96)). For some models, such as HMMs (see chapter 3, and MacKay, 1997), this is very straightforward, whereas the LDS model (see chapter 5) quickly becomes quite involved. Automated algorithm derivation programs are currently being written to alleviate this complication, specifically for the case of variational Bayesian EM operations (Bishop et al., 2003), and also for generic algorithm derivation (Buntine, 2002; Gray et al., 2003); both these projects build on results in Ghahramani and Beal (2001).

The difficulty is quite subtle and lies in the natural parameter inversion problem, which we now briefly explain. In theorem 2.2 we *conjectured* the existence of a $\tilde{\boldsymbol{\theta}}$ such that $\overline{\boldsymbol{\phi}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \overset{?}{=} \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})$, which was a point of convenience. But, the operation $\boldsymbol{\phi}^{-1}\left[\langle \boldsymbol{\phi} \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\right]$ may not be well defined if the dimensionality of $\boldsymbol{\phi}$ is greater than that of $\boldsymbol{\theta}$. Whilst not undermining the theorem's result, this does mean that representationally speaking the resulting algorithm may look different having had to be cast in terms of the natural parameters.

Online and continuous variants

The VBEM algorithm for CE models very readily lends itself to online learning scenarios in which data arrives incrementally. I briefly present here an online version of the VBEM algorithm above (but see also Ghahramani and Attias, 2000; Sato, 2001). In the standard VBM step (2.97) the variational posterior hyperparameter $\tilde{\eta}$ is updated according to the size of the dataset n (2.98), and $\tilde{\boldsymbol{\nu}}$ is updated with a simple sum of contributions from each datum $\overline{\mathbf{u}}(\mathbf{y}_i)$, (2.99).

For the online scenario, we can take the posterior over parameters described by $\tilde{\eta}$ and $\tilde{\boldsymbol{\nu}}$ to be the *prior* for subsequent inferences. Let the data be split in to batches indexed by k, each of size $n^{(k)}$, which are presented one by one to the model. Thus if the kth batch of data consists of the

$n^{(k)}$ i.i.d. points $\{\mathbf{y}_i\}_{i=j^{(k)}}^{j^{(k)}+n^{(k)}-1}$, then the online VBM step replaces equations (2.98) and (2.99) with

$$\tilde{\eta} = \eta^{(k-1)} + n^{(k)} \,, \tag{2.114}$$

$$\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu}^{(k-1)} + \sum_{i=j^{(k)}}^{j^{(k)}+n^{(k)}-1} \overline{\mathbf{u}}(\mathbf{y}_i) \,. \tag{2.115}$$

In the online VBE step only the hidden variables $\{\mathbf{x}_i\}_{i=j^{(k)}}^{j^{(k)}+n^{(k)}-1}$ need be inferred to calculate the required $\overline{\mathbf{u}}$ statistics. The online VBM and VBE steps are then iterated until convergence, which may be fast if the size of the batch $n^{(k)}$ is small compared to the amount of data previously seen $\sum_{k'=1}^{k-1} n^{(k')}$. After convergence, the prior for the next batch is set to the current posterior, according to

$$\eta^{(k)} \leftarrow \tilde{\eta} \,, \tag{2.116}$$

$$\boldsymbol{\nu}^{(k)} \leftarrow \tilde{\boldsymbol{\nu}} \,. \tag{2.117}$$

The online VBEM algorithm has several benefits. First and foremost, the update equations give us a very transparent picture of how the algorithm incorporates evidence from a new batch of data (or single data point). The way in which it does this makes it possible to discard data from earlier batches: the hyperparameters $\tilde{\eta}$ and $\tilde{\boldsymbol{\nu}}$ represent *all* information gathered from previous batches, and the process of incorporating new information is not a function of the previous batches' statistics $\{\overline{\mathbf{u}}(\mathbf{y}_i)\}_{i=j^{(1)}}^{j^{(k-1)}+n^{(k-1)}-1}$, nor previous hyperparameter settings $\{\eta^{(l)}, \boldsymbol{\nu}^{(l)}\}_{l=1}^{k-2}$, nor the previous batch sizes $\{n^{(l)}\}_{l=1}^{k-1}$, nor the previous data $\{\mathbf{y}_i\}_{i=j^{(1)}}^{j^{(k-1)}+n^{(k-1)}-1}$. Implementationally this offers a large memory saving. Since we hold a distribution over the parameters of the model, which is updated in a consistent way using Bayesian inference, we should hope that the online model makes a flexible and measured response to data as it arrives. However it has been observed (personal communication, Z. Ghahramani) that serious underfitting occurs in this type of online algorithm; this is due to excessive self-pruning of the parameters by the VB algorithm.

From the VBM step (2.97) we can straightforwardly propose an annealing variant of the VBEM algorithm. This would make use of an inverse temperature parameter $\beta \in [0, 1]$ and adopt the following updates for the VBM step:

$$\tilde{\eta} = \eta + \beta n \,, \tag{2.118}$$

$$\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \beta \sum_{i=1}^{n} \overline{\mathbf{u}}(\mathbf{y}_i) \,, \tag{2.119}$$

which is similar to the online algorithm but "introduces" the data continuously with a schedule of β from $0 \rightarrow 1$. Whilst this is a tempting avenue for research, it is not clear that in this setting we should expect any better results than if we were to present the algorithm with all the data (i.e. $\beta = 1$) from the start — after all, the procedure of Bayesian inference should produce the same inferences whether presented with the data incrementally, continuously or all at once. The advantage of an annealed model, however, is that we are giving the algorithm a better chance of escaping the local minima in the free energy that plague EM-type algorithms, so that the Bayesian inference procedure can be given a better chance of reaching the proper conclusions, whilst at every iteration receiving information (albeit β-muted) about all the data at every iteration.

2.5 Directed and undirected graphs

In this section we present several important results which build on theorems 2.1 and 2.2 by specifying the *form* of the joint density $p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$. A convenient way to do this is to use the formalism and expressive power of graphical models. We derive variational Bayesian learning algorithms for two important classes of these models: directed graphs (Bayesian networks) and undirected graphs (Markov networks), and also give results pertaining to CE families for these classes. The corollaries refer to propagation algorithms material which is covered in section 1.1.2; for a tutorial on belief networks and Markov networks the reader is referred to Pearl (1988). In the theorems and corollaries, VBEM and CE are abbreviations for *variational Bayesian Expectation-Maximisation* and *conjugate-exponential*.

2.5.1 Implications for directed networks

Corollary 2.1: (theorem 2.1) VBEM for Directed Graphs (Bayesian Networks).
Let m be a model with parameters $\boldsymbol{\theta}$ and hidden and visible variables $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^n = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ that satisfy a belief network factorisation. That is, each variable $\mathbf{z}_{ij}$ has parents $\mathbf{z}_{i\mathbf{pa}(j)}$ such that the complete-data joint density can be written as a product of conditional distributions,

$$p(\mathbf{z} \,|\, \boldsymbol{\theta}) = \prod_i \prod_j p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \ . \tag{2.120}$$

Then the approximating joint distribution for m satisfies the same belief network factorisation:

$$q_{\mathbf{z}}(\mathbf{z}) = \prod_i q_{\mathbf{z}_i}(\mathbf{z}_i) \ , \qquad q_{\mathbf{z}_i}(\mathbf{z}_i) = \prod_j \overline{q}_j(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}) \ , \tag{2.121}$$

where

$$\overline{q}_j(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}) = \frac{1}{\mathcal{Z}_{\overline{q}_j}} e^{\left\langle \ln p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \right\rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}} \quad \forall \, \{i, j\} \tag{2.122}$$

are new conditional distributions *obtained by averaging over* $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$*, and* $\mathcal{Z}_{\overline{q}_j}$ *are normalising constants.*

This corollary is interesting in that it states that a Bayesian network's posterior distribution can be factored into the same terms as the original belief network factorisation (2.120). This means that the inference for a particular variable depends only on those other variables in its *Markov blanket*; this result is trivial for the point parameter case, but definitely non-trivial in the Bayesian framework in which all the parameters and hidden variables are potentially coupled.

Corollary 2.2: (theorem 2.2) VBEM for CE Directed Graphs (CE Bayesian Networks). *Furthermore, if* m *is a conjugate-exponential model, then the conditional distributions of the approximate posterior joint have exactly the same form as those in the complete-data likelihood in the original model:*

$$\overline{q}_j(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}) = p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \tilde{\boldsymbol{\theta}}) \;, \tag{2.123}$$

but with natural parameters $\phi(\tilde{\boldsymbol{\theta}}) = \overline{\phi}$*. Moreover, with the modified parameters* $\tilde{\boldsymbol{\theta}}$*, the expectations under the approximating posterior* $q_{\mathbf{x}}(\mathbf{x}) \propto q_{\mathbf{z}}(\mathbf{z})$ *required for the VBE step can be obtained by applying the belief propagation algorithm if the network is singly connected and the junction tree algorithm if the network is multiply-connected.*

This result generalises the derivation of variational learning for HMMs (MacKay, 1997), which uses the forward-backward algorithm as a subroutine. We investigate the variational Bayesian HMM in more detail in chapter 3. Another example is *dynamic trees* (Williams and Adams, 1999; Storkey, 2000; Adams et al., 2000) in which belief propagation is executed on a single tree which represents an ensemble of singly-connected structures. Again there exists the natural parameter inversion issue, but this is merely an implementational inconvenience.

2.5.2 Implications for undirected networks

Corollary 2.3: (theorem 2.1) VBEM for Undirected Graphs (Markov Networks). *Let* m *be a model with hidden and visible variables* $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^n = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ *that satisfy a Markov network factorisation. That is, the joint density can be written as a product of clique-potentials* $\{\psi_j\}_{j=1}^J$*,*

$$p(\mathbf{z} \,|\, \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}} \prod_i \prod_j \psi_j(C_j(\mathbf{z}_i), \boldsymbol{\theta}) \;, \tag{2.124}$$

where each clique C_j *is a (fixed) subset of the variables in* $\mathbf{z}_i$*, such that* $\{C_1(\mathbf{z}_i) \cup \cdots \cup C_J(\mathbf{z}_i)\} = \mathbf{z}_i$*. Then the approximating joint distribution for* m *satisfies the same Markov network factorisation:*

$$q_{\mathbf{z}}(\mathbf{z}) = \prod_i q_{\mathbf{z}_i}(\mathbf{z}_i) \;, \qquad q_{\mathbf{z}_i}(\mathbf{z}_i) = \frac{1}{\mathcal{Z}_q} \prod_j \overline{\psi}_j(C_j(\mathbf{z}_i)) \;, \tag{2.125}$$

where

$$\overline{\psi}_j(C_j(\mathbf{z}_i)) = e^{\langle \ln \psi_j(C_j(\mathbf{z}_i), \boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}} \quad \forall \, \{i, j\} \tag{2.126}$$

are new clique potentials obtained by averaging over $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, and $\mathcal{Z}_q$ is a normalisation constant.

Corollary 2.4: (theorem 2.2) VBEM for CE Undirected Graphs (CE Markov Networks). *Furthermore, if m is a conjugate-exponential model, then the approximating clique potentials have exactly the same form as those in the original model:*

$$\overline{\psi}_j(C_j(\mathbf{z}_i)) \propto \psi_j(C_j(\mathbf{z}_i), \tilde{\boldsymbol{\theta}}) \; , \tag{2.127}$$

but with natural parameters $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}$. Moreover, the expectations under the approximating posterior $q_{\mathbf{x}}(\mathbf{x}) \propto q_{\mathbf{z}}(\mathbf{z})$ required for the VBE Step can be obtained by applying the junction tree algorithm.

For conjugate-exponential models in which belief propagation and the junction tree algorithm over hidden variables are intractable, further applications of Jensen's inequality can yield tractable factorisations (Jaakkola, 1997; Jordan et al., 1999).

2.6 Comparisons of VB to other criteria

2.6.1 BIC is recovered from VB in the limit of large data

We show here informally how the Bayesian Information Criterion (BIC, see section 1.3.4) is recovered in the large data limit of the variational Bayesian lower bound (Attias, 1999b). $\mathcal{F}$ can be written as a sum of two terms:

$$\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \underbrace{-\mathrm{KL}\left[q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \,|\, m)\right]}_{\mathcal{F}_{m,pen}} + \underbrace{\left\langle \ln \frac{p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)}{q_{\mathbf{x}}(\mathbf{x})} \right\rangle_{q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}}_{\mathcal{D}_m} \; . \tag{2.128}$$

Let us consider separately the limiting forms of these two terms, constraining ourselves to the cases in which the model m is in the CE family. In such cases, theorem 2.2 states that $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is of conjugate form (2.97) with parameters given by (2.98) and (2.99). It can be shown that under mild conditions exponential family distributions of this form exhibit asymptotic normality (see, for example, the proof given in Bernardo and Smith, 1994, pp. 293–4). Therefore, the entropy of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ appearing in $\mathcal{F}_{m,pen}$ can be calculated assuming a Gaussian form (see appendix A), and the limit becomes

$$\lim_{n\to\infty} \mathcal{F}_{m,pen} = \lim_{n\to\infty} \left[\langle \ln p(\boldsymbol{\theta}\,|\,m) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|H| \right] \tag{2.129}$$

$$= -\frac{d}{2}\ln n + \mathcal{O}(1) \,, \tag{2.130}$$

where H is the Hessian (matrix of second derivatives of the parameter posterior evaluated at the mode), and we have used similar arguments to those taken in the derivation of BIC (section 1.3.4). The second term, $\mathcal{D}_m$, can be analysed by appealing to the fact that the term inside the expectation is equal to $\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}, m)$ if and only if $q_{\mathbf{x}}(\mathbf{x}) = p(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\theta}, m)$. Theorem 2.1 states that the form of the variational posterior over hidden states $q_{\mathbf{x}}(\mathbf{x})$ is given by

$$\ln q_{\mathbf{x}}(\mathbf{x}) = \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln p(\mathbf{x}, \mathbf{y}\,|\,\boldsymbol{\theta}, m) - \ln \mathcal{Z}_{\mathbf{x}} \tag{2.131}$$

(which does not depend on CE family membership conditions). Therefore as $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ becomes concentrated about $\boldsymbol{\theta}_{\text{MAP}}$, this results in $q_{\mathbf{x}}(\mathbf{x}) = p(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\theta}_{\text{MAP}}, m)$. Then $\mathcal{D}_m$ asymptotically becomes $\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}_{\text{MAP}}, m)$. Combining this with the limiting form for $\mathcal{F}_{m,pen}$ given by (2.130) results in:

$$\lim_{n\to\infty} \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = -\frac{d}{2}\ln n + \ln p(\mathbf{y}\,|\,\boldsymbol{\theta}_{\text{MAP}}, m) + \mathcal{O}(1) \,, \tag{2.132}$$

which is the BIC approximation given by (1.49). For the case of a non-CE model, we would have to prove asymptotic normality for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ outside of the exponential family, which may become complicated or indeed impossible. We note that this derivation of the limiting form of VB is heuristic in the sense that we have neglected concerns on precise regularity conditions and identifiability.

2.6.2 Comparison to Cheeseman-Stutz (CS) approximation

In this section we present results regarding the approximation of Cheeseman and Stutz (1996), covered in section 1.3.5. We briefly review the CS criterion, as used to approximate the marginal likelihood of finite mixture models, and then show that it is in fact a strict lower bound on the marginal likelihood. We conclude the section by presenting a construction that proves that VB can be used to obtain a bound that is *always* tighter than CS.

Let m be a directed acyclic graph with parameters $\boldsymbol{\theta}$ giving rise to an i.i.d. data set denoted by $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with corresponding discrete hidden variables $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ each of cardinality k. Let $\hat{\boldsymbol{\theta}}$ be a result of an EM algorithm which has converged to a local maximum in the likelihood $p(\mathbf{y}\,|\,\boldsymbol{\theta})$, and let $\hat{\mathbf{s}} = \{\hat{\mathbf{s}}_i\}_{i=1}^n$ be a completion of the hidden variables, chosen according to the posterior distribution over hidden variables given the data and $\hat{\boldsymbol{\theta}}$, such that $\hat{\mathbf{s}}_{ij} = p(\mathbf{s}_{ij} = j \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) \; \forall \; i = 1, \ldots, n$.

Since we are completing the hidden variables with real, as opposed to discrete values, this complete data set does not in general correspond to a realisable data set under the generative model. This point raises the question of how its marginal probability $p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m)$ is defined. We will see in the following theorem and proof (theorem 2.3) that both the completion required of the hidden variables and the completed data marginal probability are well-defined, and follow from equations 2.141 and 2.142 below.

The CS approximation is given by

$$p(\mathbf{y} \,|\, m) \approx p(\mathbf{y} \,|\, m)_{\text{CS}} = p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) \frac{p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}})}{p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}})} \; . \tag{2.133}$$

The CS approximation exploits the fact that, for many models of interest, the first term on the right-hand side, the complete-data marginal likelihood, is tractable to compute (this is the case for discrete-variable directed acyclic graphs with Dirichlet priors, see chapter 6 for details). The term in the numerator of the second term on the right-hand side is simply the likelihood of the data, which is an output of the EM algorithm (as is the parameter estimate $\hat{\boldsymbol{\theta}}$), and the denominator is a straightforward calculation that involves no summations over hidden variables or integrations over parameters.

Theorem 2.3: Cheeseman-Stutz approximation is a lower bound on the marginal likelihood.
Let $\hat{\boldsymbol{\theta}}$ be the result of the M step of EM, and let $\{p(\mathbf{s}_i \,|\, \mathbf{y}_i, \hat{\boldsymbol{\theta}})\}_{i=1}^{n}$ be the set of posterior distributions over the hidden variables obtained in the next E step of EM. Furthermore, let $\hat{\mathbf{s}} = \{\hat{\mathbf{s}}_i\}_{i=1}^{n}$ be a completion of the hidden variables, such that $\hat{\mathbf{s}}_{ij} = p(\mathbf{s}_{ij} = j \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) \; \forall \; i = 1, \ldots, n$. Then the CS approximation is a lower bound on the marginal likelihood:

$$p(\mathbf{y} \,|\, m)_{CS} = p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) \frac{p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}})}{p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}})} \leq p(\mathbf{y} \,|\, m) \; . \tag{2.134}$$

This observation should be attributed to Minka (2001b), where it was noted that (in the context of mixture models with unknown mixing proportions and component parameters) whilst the CS approximation has been reported to obtain good performance in the literature (Cheeseman and Stutz, 1996; Chickering and Heckerman, 1997), it was not known to be a bound on the marginal likelihood. Here we provide a proof of this statement that is generally applicable to any model.

Proof of theorem 2.3: via marginal likelihood bounds using approximations over the posterior distribution of only the hidden variables. The marginal likelihood can be lower bounded by introducing a distribution over the settings of each data point's hidden variables $q_{\mathbf{s}_i}(s_i)$:

$$p(\mathbf{y}\,|\,m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{y}_i\,|\,\boldsymbol{\theta}) \tag{2.135}$$

$$\geq \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} . \tag{2.136}$$

We return to this quantity shortly, but presently place a similar lower bound over the likelihood of the data:

$$p(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} p(\mathbf{y}_i\,|\,\hat{\boldsymbol{\theta}}) \geq \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i\,|\,\hat{\boldsymbol{\theta}})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} \tag{2.137}$$

which can be made an equality if, for each data point, $q(\mathbf{s}_i)$ is set to the exact posterior distribution given the parameter setting $\boldsymbol{\theta}$ (for example see equation (2.19) and the proof following it),

$$p(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} p(\mathbf{y}_i\,|\,\hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i\,|\,\hat{\boldsymbol{\theta}})}{\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} , \tag{2.138}$$

where

$$\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \equiv p(\mathbf{s}_i\,|\,\mathbf{y}, \hat{\boldsymbol{\theta}}) , \tag{2.139}$$

which is the result obtained from an exact E step with the parameters set to $\hat{\boldsymbol{\theta}}$. Now rewrite the marginal likelihood bound (2.136), using this same choice of $\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)$, separate those terms that depend on $\boldsymbol{\theta}$ from those that do not, and substitute in the form from equation (2.138) to obtain:

$$p(\mathbf{y}\,|\,m) \geq \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} \cdot \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}) \right\} \tag{2.140}$$

$$= \frac{p(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i\,|\,\hat{\boldsymbol{\theta}}) \right\}} \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}) \right\} \tag{2.141}$$

$$= \frac{p(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} p(\hat{\mathbf{s}}_i, \mathbf{y}_i\,|\,\hat{\boldsymbol{\theta}})} \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\hat{\mathbf{s}}_i, \mathbf{y}_i\,|\,\boldsymbol{\theta}) , \tag{2.142}$$

where $\hat{\mathbf{s}}_i$ are defined such that they satisfy:

$$\hat{\mathbf{s}}_i \quad \text{defined such that:} \qquad \ln p(\hat{\mathbf{s}}_i, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) = \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \tag{2.143}$$

$$= \sum_{\mathbf{s}_i} p(\mathbf{s}_i \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \tag{2.144}$$

where the second line comes from the requirement of bound equality in (2.139). The existence of such a completion follows from the fact that, in discrete-variable directed acyclic graphs of the sort considered in Chickering and Heckerman (1997), the hidden variables appear only linearly in logarithm of the joint probability $p(\mathbf{s}, \mathbf{y} \,|\, \boldsymbol{\theta})$. Equation (2.142) is the Cheeseman-Stutz criterion, and is also a lower bound on the marginal likelihood. □

It is possible to derive CS-like approximations for types of graphical model other than discrete-variables DAGs. In the above proof no constraints were placed on the forms of the joint distributions over hidden and observed variables, other than in the simplifying step in equation (2.142). So, similar results to corollaries 2.2 and 2.4 can be derived straightforwardly to extend theorem 2.3 to incorporate CE models.

The following corollary shows that variational Bayes can always obtain a tighter bound than the Cheeseman-Stutz approximation.

Corollary 2.5: (theorem 2.3) VB is at least as tight as CS.
That is to say, it is always possible to find distributions $q_{\mathbf{s}}(\mathbf{s})$ *and* $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *such that*

$$\ln p(\mathbf{y} \,|\, m)_{CS} \leq \mathcal{F}_m(q_{\mathbf{s}}(\mathbf{s}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \leq \ln p(\mathbf{y} \,|\, m) \; . \tag{2.145}$$

Proof of corollary 2.5. Consider the following forms for $q_{\mathbf{s}}(\mathbf{s})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$q_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} q_{\mathbf{s}_i}(\mathbf{s}_i) \; , \quad \text{with} \quad q_{\mathbf{s}_i}(\mathbf{s}_i) = p(\mathbf{s}_i \,|\, \mathbf{y}_i, \hat{\boldsymbol{\theta}}) \; , \tag{2.146}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \langle \ln p(\boldsymbol{\theta}) p(\mathbf{s}, \mathbf{y} \,|\, \boldsymbol{\theta}) \rangle_{q_{\mathbf{s}}(\mathbf{s})} \; . \tag{2.147}$$

We write the form for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ explicitly:

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\}}{\int d\boldsymbol{\theta}' \; p(\boldsymbol{\theta}') \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}') \right\}} \; , \tag{2.148}$$

and note that this is exactly the result of a VBM step. We substitute this and the form for $q_{\mathbf{s}}(\mathbf{s})$ directly into the VB lower bound stated in equation (2.53) of theorem 2.1, obtaining:

$$\mathcal{F}(q_{\mathbf{s}}(\mathbf{s}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta})}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{2.149}$$

$$= \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \int d\boldsymbol{\theta}'\; p(\boldsymbol{\theta}') \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}') \right\} \tag{2.150}$$

$$= \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \ln \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\} , \tag{2.151}$$

which is exactly the logarithm of equation (2.140). And so with this choice of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{s}}(\mathbf{s})$ we achieve equality between the CS and VB approximations in (2.145).

We complete the proof of corollary 2.5 by noting that any further VB optimisation is guaranteed to increase or leave unchanged the lower bound, and hence surpass the CS lower bound. We would expect the VB lower bound starting from the CS solution to improve upon the CS lower bound in *all* cases, except in the very special case when the MAP parameter $\hat{\boldsymbol{\theta}}$ is exactly the *variational Bayes point*, defined as $\boldsymbol{\theta}_{\mathrm{BP}} \equiv \boldsymbol{\phi}^{-1}(\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})})$ (see proof of theorem 2.2(a)). Therefore, since VB is a lower bound on the marginal likelihood, the entire statement of (2.145) is proven. □

2.7 Summary

In this chapter we have shown how a variational bound can be used to derive the EM algorithm for ML/MAP parameter estimation, for both unconstrained and constrained representations of the hidden variable posterior. We then moved to the Bayesian framework, and presented the *variational Bayesian EM* algorithm which iteratively optimises a lower bound on the marginal likelihood of the model. The marginal likelihood, which integrates over model parameters, is the key component to Bayesian model selection. The VBE and VBM steps are obtained by taking functional derivatives with respect to variational distributions over hidden variables and parameters respectively.

We gained a deeper understanding of the VBEM algorithm by examining the specific case of *conjugate-exponential* models and showed that, for this large class of models, the posterior distributions $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ have intuitive and analytically stable forms. We have also presented

VB learning algorithms for both directed and undirected graphs (Bayesian networks and Markov networks).

We have explored the Cheeseman-Stutz model selection criterion as a lower bound of the marginal likelihood of the data, and have explained how it is a very specific case of variational Bayes. Moreover, using this intuition, we have shown that any CS approximation can be improved upon by building a VB approximation over it. It is tempting to derive conjugate-exponential versions of the CS criterion, but in my opinion this is not necessary since any implementations based on these results can be made only more accurate by using conjugate-exponential VB instead, which is at least as general in every case. In chapter 6 we present a comprehensive comparison of VB to a variety of approximation methods, including CS, for a model selection task involving discrete-variable DAGs.

The rest of this thesis applies the VB lower bound to several commonly used statistical models, with a view to performing model selection, learning from both real and synthetic data sets. Throughout we compare the variational Bayesian framework to competitor approximations, such as those reviewed in section 1.3, and also critically analyse the quality of the lower bound using advanced sampling methods.