Bayesian Inference for Latent Variable Models

Ulrich Paquet

Wolfson College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

March 2007

Declaration

I hereby declare that my thesis entitled *Bayesian Inference for Latent Variable Models* is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University, and does not exceed 60,000 words.

Ulrich Paquet

Abstract

Bayes' theorem is the cornerstone of statistical inference. It provides the tools for dealing with knowledge in an uncertain world, allowing us to explain observed phenomena through the refinement of belief in model parameters. At the heart of this elegant framework lie intractable integrals, whether in computing an average over some posterior distribution, or in determining the normalizing constant of a distribution. This thesis examines both deterministic and stochastic methods in which these integrals can be treated. Of particular interest shall be parametric models where the parameter space can be extended with additional latent variables to get distributions that are easier to handle algorithmically.

Deterministic methods approximate the posterior distribution with a simpler distribution over which the required integrals become tractable. We derive and examine a new generic α-divergence message passing scheme for a multivariate mixture of Gaussians, a particular modeling problem requiring latent variables. This algorithm minimizes local α-divergences over a chosen posterior factorization, and includes variational Bayes and expectation propagation as special cases.

Stochastic (or Monte Carlo) methods rely on a sample from the posterior to simplify the integration tasks, giving exact estimates in the limit of an infinite sample. Parallel tempering and thermodynamic integration are introduced as 'gold standard' methods to sample from multimodal posterior distributions and determine normalizing constants. A parallel tempered approach to sampling from a mixture of Gaussians posterior through Gibbs sampling is derived, and novel methods are introduced to improve the numerical stability of thermodynamic integration. A full comparison with parallel tempering and thermodynamic integration shows variational Bayes, expectation propagation, and message passing with the Hellinger distance $\alpha = \frac{1}{2}$ to be perfectly suitable for model selection, and for approximating the predictive distribution with high accuracy.

Variational and stochastic methods are combined in a novel way to design Markov chain Monte Carlo (MCMC) transition densities, giving a variational transition kernel, which lower bounds an exact transition kernel. We highlight the general need to mix variational methods with other MCMC moves, by proving that the variational kernel does not necessarily give a geometrically ergodic chain.

Acknowledgments

My time in Cambridge was generously supported by the Association of Commonwealth Universities through a Commonwealth Scholarship, for which I am deeply thankful.

Numerous people have been instrumental in the way I think about inference, and I thank Ole Winther for great collaboration and support, and for hosting me for a week at the Technical University of Denmark, Tom Minka and Martin Szummer for organizing the reading group at Microsoft Research, and David MacKay and the Inference group at the Cavendish Laboratory for providing a platform for many stimulating talks. At the Computer Laboratory I thank Sean Holden for allowing me the freedom to satisfy my curiosity, and Andrew Naish-Guzman for being a fantastic labmate, conference partner and friend. I had the best of times working on some odd projects: with Joseph Stevick on Gaussian Processes to correct MRI images, and with Blaise Thomson on collaborative filtering.

At College I thank a group of great friends, especially Joseph Stevick, Blaise Thomson, Muzoora Bishanga, Glenton and Christine Jelbert, Heather Harrison, Werner Bäumker, and Gerhard Hancke. Thanks also to my teammates at Cambridge University Hare & Hounds.

Further from Cambridge I thank my family at home, who have given me tremendous encouragement throughout my graduate years.

Ulrich Paquet, Cambridge, March 2007

Contents

Chapter 1

Introduction

1.1 Learning from data

Bayesian theory provides a general and consistent framework for dealing with uncertainty. In everyday life uncertainty often permeates our choices, and when choices need to be made, past experience frequently proves a helpful aid.

This very same principle is applicable when machines are faced with the task of learning and dealing with uncertainty. Learning from past experience may take many guises, of which classification, regression and density estimation are but a few. As a practical example, we may care about the automatic classification of handwritten digits. When given an image of a written digit, we wish to predict whether it is a number from zero to nine. This task is simpler if we actually know how typical examples are classified, and a helpful aid in this case is a set of example classifications, or a data set of labeled images of handwritten digits. Uncertainty can then be dealt with in a crisp manner: what is the *probability* that an image corresponds to a nine, say, *given* that we know how a few other images should be classified?

It is impractical to enumerate and store every possible variation of a written digit. Therefore the approach forwarded by *machine learning* is to assume that some parametric model is responsible for generating the labels for written digits. This model can be used for prediction of previously unseen digits by tuning its parameters to predict the observed examples well, and we effectively learn a functional mapping (or model) from an input to an output space. At the core, we hope to make good predictions in the future by fitting a model to known predictions. As an aside, nothing confines us to use a single model, as the rules of probability advocate an averaging of predictions over a set of plausible models or possible parameter settings.

In the above example, and also in regression, we are concerned with the probability distribution of an output variable; given some input variable, the output is treated as a random variable. In the same manner the input variable can be treated with uncertainty. In density estimation, we are interested in the unknown distribution from which some data points have been generated. Continuing the same example, we may be presented with an unlabeled set of images of written characters, and asked to infer the probability density of an image of a character, given the observations. Again, we would assume some underlying model with tunable parameters to describe the density well.

1.2 Bayes' theorem

The problem of learning from data can be cast into a formal Bayesian framework. Say we observe data $\mathbf{x} = \{\mathbf{x}_n\}_{n=1}^N$, or equally say that some observations from a random variable have been made. To 'learn' from the observed data, or use it for inference, it is necessary to assume that it was generated by some model $\mathcal{M}$, possibly with parameters $\boldsymbol{\theta}$. A common assumption is that the data are independent and identically distributed and drawn from some likelihood $p(\mathbf{x}_n|\boldsymbol{\theta}, \mathcal{M})$. This sets the scene for parametric inference. It is not always necessary to explicitly work with our model parameters; non-parametric methods can provide for an equally elegant example of Bayesian inference. Although the methods discussed here are general, all examples in this thesis come from the parametric camp.

Bayes' theorem forces us to make our model assumptions $\mathcal{M}$ explicit; in other words, we are asked to specify the model that we believe in. This opens the door for sensibly comparing models, which will be explored later. From Bayes' theorem the posterior distribution over the parameters is equal to the likelihood of observing that data, given a particular parameter setting, multiplied by our prior belief about the parameter values. This is scaled by a normalizing constant that is known as the evidence or marginal likelihood,

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M}) = \frac{p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{M})}{\int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{M})\, d\boldsymbol{\theta}} = \frac{p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathbf{x}|\mathcal{M})} \ . \tag{1.1}$$

Under the assumption of independent and identically distributed data, the likelihood is a product over individual example likelihoods, $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}) = \prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\theta}, \mathcal{M})$.

Three common tasks of interest are: 1) the prediction of unseen data conditioned on the observed data, or more generally determining *expectations* over the posterior distribution; 2) integrating away parameters we are not interested in, also called *marginalization*; 3) the evaluation of the validity of our assumed model, which includes the task of computing the *normalizing constant* in Bayes' theorem.

1.2.1 Prediction

The first question is that of prediction—determining the distribution of a new data point given the observed data—and is answered by averaging over the posterior distribution,

$$p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}) = \int p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})\, d\boldsymbol{\theta} \ . \tag{1.2}$$

This is often a difficult and analytically intractable integration problem, as the posterior may have a very convoluted form, often being high-dimensional with many modes. Even more generally we may want to average functions over the posterior with

$$\Phi = \langle \phi(\boldsymbol{\theta}) \rangle = \int \phi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})\, d\boldsymbol{\theta} \ , \tag{1.3}$$

which may include determining the posterior mean, with $\phi(\boldsymbol{\theta}) = \boldsymbol{\theta}$, for example.

1.2.2 Marginalization

If we have a joint distribution over variables $\boldsymbol{\theta}$ and $\mathbf{z}$, we may only be interested in the marginal distribution over $\boldsymbol{\theta}$, and average over the other variables

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M}) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x}, \mathcal{M})\, d\mathbf{z} \ . \tag{1.4}$$

1.2.3 Model selection

The task of estimating the normalizing constant in Bayes' theorem is related to another question that we may ask, namely how well our assumed model supports the data. There is no guarantee that a specific model $\mathcal{M}$ provides a preferable description of the data, and the road of inference diverges into two paths—

1. We make the assumption that each model in a set of models $\{\mathcal{M}_i\}$ has some possibility of generating the data, and make predictions by *averaging* over the respective posterior distributions of each $\mathcal{M}_i$.
2. We prefer one model from $\{\mathcal{M}_i\}$, and base our choice on the marginal likelihood as a natural embodiment of *Ockham's razor*.

Both these paths are discussed below.

Averaging over a set of models

Prediction may rely on higher levels of inference, where we average the predictive distribution of equation (1.2) over the posterior distribution of a set of plausible models $\{\mathcal{M}_i\}$, with

$$p(\mathbf{x}_{\text{new}}|\mathbf{x}) = \sum_{\mathcal{M}_i} p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}_i)p(\mathcal{M}_i|\mathbf{x}) \ . \tag{1.5}$$

In each case $p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}_i)$ will involve integration over a set of parameters specific to $\mathcal{M}_i$.

For model averaging to be possible, we have to define a prior distribution $p(\mathcal{M}_i)$ over the set of models, and again rely on Bayes' theorem for the posterior,

$$p(\mathcal{M}_i|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathbf{x})} \ . \tag{1.6}$$

The term 'marginal likelihood'—the normalizer or evidence from equation (1.1)—is the *likelihood* term in equation (1.6). The likelihood is *marginal*, as the model parameters are integrated (or marginalized) out.

Ockham's razor

In the context of Bayes' theorem, the question of how well our assumed model supports the data is answered by the marginal likelihood. This is the question of model selection: we may want to know how many clusters would be sufficient to model the data well, what the intrinsic dimensionality of the data is, whether an input is relevant to predicting an output, and so forth.

When comparing models $\mathcal{M}_1$ and $\mathcal{M}_2$ on seeing data $\mathbf{x}$, we consider the probability ratio between the posterior probabilities of the two models. From equation (1.6) we have

$$\frac{p(\mathcal{M}_1|\mathbf{x})}{p(\mathcal{M}_2|\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{M}_1)}{p(\mathbf{x}|\mathcal{M}_2)}\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \ . \tag{1.7}$$

When determining the posterior ratio, two ratios are taken into account. A prior ratio $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ encodes how much our initial beliefs favor one model over the other. The ratio of marginal likelihoods $p(\mathbf{x}|\mathcal{M}_1)/p(\mathbf{x}|\mathcal{M}_2)$ gives an indication of how much better one model is in explaining the data, compared to the other. We may let the prior ratio prefer a simpler model to a more complex one, but the beautiful consequence of dealing with uncertainty using Bayesian theory is that Ockham's razor is *automatically* expressed (MacKay, 1992).

The English Franciscan friar William of Ockham is known in the scientific community by his famous razor,

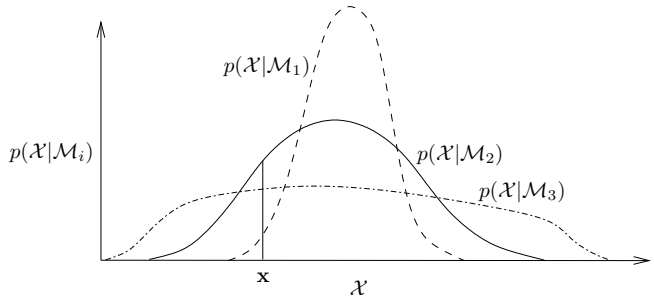


FIGURE 1.1: A schematic illustration of Ockham's razor. We imagine that all data sets, which we call $\mathcal{X}$, are projected onto the one-dimensional horizontal axis. $\mathcal{X}$ is a random variable, and therefore $p(\mathcal{X}|\mathcal{M}_i)$ should integrate to one for each model. As a complex model $\mathcal{M}_3$ can explain many data sets, it should also spread its probability mass over a large 'area' of data sets. Consequently, if a *specific* data set $\mathbf{x}$ is observed, the most probable model is the model with largest marginal likelihood, i.e. a model that is neither too simple nor too complex *for* $\mathbf{x}$. This figure is adapted from (MacKay, 1995).

entia non sunt multiplicanda praeter necessitatem,

which translates to "entities should not be multiplied beyond necessity". It advocates the simplest possible explanation for the data that we have observed, but no simpler explanation than that. Figure 1.1 gives a cartoon, illustrating how a simple model $\mathcal{M}_2$ may give a reasonable explanation to a few data sets, while a complex model $\mathcal{M}_3$ with more parameters may explain a wider variety of data sets.[1] The probability mass $p(\mathcal{X}|\mathcal{M}_3)$ of the a more complex $\mathcal{M}_3$ should be spread over a larger 'area' of data sets $\mathcal{X}$. Hence, when a particular data set $\mathbf{x}$ can be explained well by both models, we observe a higher marginal likelihood for the simpler model. The higher the marginal likelihood, the better the model supports the data. If we follow the same argument for figure 1.1, it is clear why a 'too simple' model $\mathcal{M}_1$ would also not be preferred. Bayes' theorem provides a natural way of penalizing models with superfluous power of explanation through the marginal likelihood.

In both the case of model averaging and model selection through Ockham's razor, we need to determine the marginal likelihood. As in the case of prediction, this is often a difficult problem, as evaluation of the integral $\int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{M})\, d\boldsymbol{\theta}$ can be analytically intractable.

1.3 Practical approaches

Many problems in Bayesian inference therefore leave us with intractable questions: we cannot simply write down the answer in a closed form solution. The posterior or joint distribution that we are interested in is often of high dimensionality, and in cases like mixture models can exhibit an exponentially increasing number of modes. We are faced with resorting to either deterministic or stochastic (Monte Carlo) methods to perform inference. Deterministic methods aim to simplify the problem to an analytically tractable one by finding approximations to the joint or posterior distribution. Monte Carlo methods, on the other hand, rely on large samples from the distribution in question to provide asymptotically correct answers.

[1]This does in no way imply that we can equate the predictive power of a model with its number of parameters.

1.3.1 Deterministic methods

Many high dimensional integration problems in machine learning can be simplified through an analytically tractable *approximation* to the joint distribution,

$$p(\boldsymbol{\theta}, \mathbf{x}|\mathcal{M}) = p(\mathbf{x}|\mathcal{M})p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M}) \approx sq(\boldsymbol{\theta}) \ . \tag{1.8}$$

The problem of prediction and model selection becomes greatly simplified when we have an approximation that summarizes the important features of the joint distribution. The joint distribution $p(\boldsymbol{\theta}, \mathbf{x})$ is approximated by an 'easier' normalized distribution $q(\boldsymbol{\theta})$, appropriately scaled by s. The posterior will then be approximated by q, allowing us to use it as a surrogate to the posterior to make predictions (as integrating over q in (1.2) should be a simpler task). The scale s gives an approximation to the marginal likelihood, needed for model comparison and selection.

We have effectively replaced an integration problem by an *optimization* problem: how to best fit $sq(\boldsymbol{\theta})$ to the joint distribution. We are left with a few unanswered questions, namely how to choose a parameterized q, and how to measure the goodness of fit.

Maximum a posteriori

At the very simplest level, we can replace the posterior distribution with a point mass at its maximum, so that $q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MP}})$. Here $\delta(\cdot)$ denotes the Dirac delta function, which is infinite when its argument is zero, and zero otherwise, and is defined to have unit mass . The maximum a posteriori (MAP) parameter estimate $\boldsymbol{\theta}_{\mathrm{MP}}$ would be the mode,

$$\boldsymbol{\theta}_{\mathrm{MP}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}) \ . \tag{1.9}$$

This often gives over-confident predictions, as areas of *mass* of the posterior, critical in evaluating integrals like equation (1.2), are not taken into account. In this case the task of prediction simplifies as $p(\mathbf{x}_{\mathrm{new}}|\mathbf{x}, \mathcal{M}) \approx p(\mathbf{x}_{\mathrm{new}}|\boldsymbol{\theta}_{\mathrm{MP}}, \mathcal{M})$. The MAP estimate can be seen as a penalized version of the maximum likelihood (ML) estimate, $\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})$, where the 'penalty' for big, finely-tuned parameter values is given by the prior. A common interpretation and link with learning theory views the log prior as a *regularizer* on a set of functions (the log likelihood).

Laplace's method

A common way of including probability mass in a MAP estimate is the method of Laplace. The approximation relies on the *curvature* of the joint distribution at $\boldsymbol{\theta}_{\mathrm{MP}}$. By taking the Taylor series of the log joint distribution around its mode, truncating it after the quadratic term and exponentiating, we obtain a Gaussian approximation to the posterior, and scale to approximate the marginal likelihood. Let the negative log joint distribution, as a function of the model parameters, be

$$M(\boldsymbol{\theta}) = -\ln p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}) - \ln p(\boldsymbol{\theta}|\mathcal{M}) = -\sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\theta}, \mathcal{M}) - \ln p(\boldsymbol{\theta}|\mathcal{M}) \ , \tag{1.10}$$

so that $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{M}) = e^{-M(\boldsymbol{\theta})}$. Were we interested a single parameter setting for classification or regression, $M(\boldsymbol{\theta})$ could be viewed as an error function that is minimized (to find $\boldsymbol{\theta}_{\mathrm{MP}}$). The

error of a single prediction would then have been given by $-\ln p(\mathbf{x}_n|\boldsymbol{\theta},\mathcal{M})$, while $-\ln p(\boldsymbol{\theta}|\mathcal{M})$ would be used for 'weight decay', or act as a regularizer.

However, we are rather interested in posterior mass, and for that purpose Taylor-expand $M(\boldsymbol{\theta})$ around its most probable parameter value,

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}_{\mathrm{MP}}) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MP}})^\top \mathbf{A}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MP}}) + \cdots \ . \tag{1.11}$$

The first derivative term is excluded from the expansion, as $\partial M(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ evaluates to zero at $\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}$. Matrix $\mathbf{A}$ is the Hessian, the matrix of second derivatives,

$$\mathbf{A} = -\frac{\partial^2 \ln p(\mathbf{x},\boldsymbol{\theta}|\mathcal{M})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}} = \frac{\partial^2 M(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}} \ . \tag{1.12}$$

Finding an approximation $sq(\boldsymbol{\theta})$ to the joint distribution then simply involves re-exponentiating the truncated Taylor approximation around the mode, i.e.

$$\begin{aligned} sq(\boldsymbol{\theta}) &= e^{-M(\boldsymbol{\theta}_{\mathrm{MP}})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MP}})^\top\mathbf{A}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MP}})} \\ &= e^{-M(\boldsymbol{\theta}_{\mathrm{MP}})}(2\pi)^{d/2}|\mathbf{A}|^{-1/2}\mathcal{N}(\boldsymbol{\theta}\mid\boldsymbol{\theta}_{\mathrm{MP}},\mathbf{A}^{-1}) \ . \end{aligned} \tag{1.13}$$

The result is a Gaussian approximation $q(\boldsymbol{\theta})$ to the posterior distribution, with covariance matrix given by the inverse Hessian. The log marginal likelihood will be approximated with $\ln s = \ln p(\mathbf{x}|\boldsymbol{\theta}_{\mathrm{MP}},\mathcal{M}) + \ln p(\boldsymbol{\theta}_{\mathrm{MP}}|\mathcal{M}) + \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|$. The dimensionality of $\boldsymbol{\theta}$, or size of $\mathbf{A}$, is indicated by d.

The approximation may suffer from a few drawbacks, most notably when the log joint is not approximately quadratic. This may for example occur in the case of smaller datasets, where the advantage of a closer approximation to the probability mass becomes more evident. The Gaussian approximation should work well in the large data limit. In the case where parameters are constrained to be positive, for example, we have to rely on a change of basis to make the approximation work (MacKay, 1998). When the posterior is multimodal with well separated modes, the approximation will be local to a particular mode.

Methods relying on divergence measures

A sensible way to find a suitably scaled approximation to the joint distribution is to precisely define the 'distance' between them. We can them aim to minimize this measure of divergence to the best of our abilities: in some cases an exact minimization may be possible, and in others we have to be content with minimizing some surrogate to the chosen measure. Two popular methods to achieve this goal are variational Bayes (Hinton & van Camp, 1993) and expectation propagation (Minka, 2001c). Both minimize, or approximately minimize, some form of α-divergence (Amari, 1985), indexed by a continuous parameter $\alpha \in \mathbb{R}$:

$$D_\alpha\big(p(\mathbf{x},\boldsymbol{\theta}|\mathcal{M}) \,\big\|\, sq(\boldsymbol{\theta})\big) = \frac{\int \alpha p(\mathbf{x},\boldsymbol{\theta}|\mathcal{M}) + (1-\alpha)sq(\boldsymbol{\theta}) - p(\mathbf{x},\boldsymbol{\theta}|\mathcal{M})^\alpha[sq(\boldsymbol{\theta})]^{1-\alpha}d\boldsymbol{\theta}}{\alpha(1-\alpha)} \ . \tag{1.14}$$

This approach holds a number of advantages. Variational Bayes always gives a scale s that is a lower bound to the marginal likelihood, allowing informed choices about model selection to be made. If we write the joint distribution as a product of factors, expectation propagation performs termwise moment-matching, and in solving these smaller subproblems aims to match the scale and moments of the full joint distribution.

Chapter 2 presents a detailed introductory account of these methods through a toy example, and we shall not delve into the same level of detail here.

1.3.2 Stochastic (Monte Carlo) methods

Another approach to estimating integrals like equation (1.2) is to use Monte Carlo methods to draw a sample from the posterior distribution, and rely on the law of large numbers to estimate these integrals using the sample (Robert & Casella, 2004).

If we have some random sample $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$ from a distribution of interest at our disposal—say it is the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})$—we can estimate expectations under this distribution. The expectation of some scalar functional $\phi(\boldsymbol{\theta})$ under the posterior distribution, $\Phi = \langle\phi(\boldsymbol{\theta})\rangle$ from equation (1.3), can be empirically estimated with the ergodic average

$$\hat{\Phi}_T = \frac{1}{T}\sum_{t=1}^{T}\phi(\boldsymbol{\theta}^{(t)}) \ . \tag{1.15}$$

The estimate $\hat{\Phi}_T$ is unbiased and will almost surely converge to Φ, as $T \to \infty$, by the strong law of large numbers. The distribution of interest is typically referred to as the *target* distribution, which we assume can be evaluated anywhere up to a normalizing constant. Therefore let $p^*(\boldsymbol{\theta}) \equiv p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ be the unnormalized posterior distribution (or joint distribution).

Monte Carlo methods come in many guises, but for our purposes we shall restrict ourselves to two methods, Importance Sampling and Markov chain Monte Carlo (MCMC) (MacKay, 2003; Neal, 1993; Robert & Casella, 2004).

Importance sampling

When $p^*(\boldsymbol{\theta})$ is sufficiently complex so that we cannot sample from it directly, we may opt for sampling from a distribution $q(\boldsymbol{\theta})$ from which we *can* generate samples. It may also only be needed to evaluate q up to a normalizing constant, such that $q(\boldsymbol{\theta}) = q^*(\boldsymbol{\theta})/\int q^*(\boldsymbol{\theta})d\boldsymbol{\theta}$. We generate T samples from q^*, and can determine the estimator given in equation (1.15) if we appropriately *reweigh* the samples. Samples where $p(\boldsymbol{\theta})$ is greater than $q(\boldsymbol{\theta})$ are under-represented, and need to have a greater influence in the estimator; the reverse applies for where samples are over-represented. *Importance* weights

$$w^{(t)} = \frac{p^*(\boldsymbol{\theta}^{(t)})}{q^*(\boldsymbol{\theta}^{(t)})} \tag{1.16}$$

are then used to compensate for sampling from the wrong distribution, and the empirical expectation (1.15) becomes

$$\hat{\Phi}_T = \frac{\sum_{t=1}^{T} w^{(t)}\phi(\boldsymbol{\theta}^{(t)})}{\sum_{t=1}^{T} w^{(t)}} \ . \tag{1.17}$$

This estimator is consistent and biased when q is unnormalized as well. Although simple to implement, importance sampling typically becomes impractical when higher-dimensional distributions are involved. It is possible to show that even for simple cases the variance of the importance weights can be infinite (MacKay, 2003). Some of these subtleties are highlighted in chapter 5, where importance sampling and its cousin, the 'independent Metropolis-Hastings sampler', are reviewed in relation to approximate distributions $q(\boldsymbol{\theta})$ of the sort found in section 1.3.1.

Markov chain Monte Carlo

An indirect method of sampling from $p(\boldsymbol{\theta}|\mathbf{x})$ (with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$) is to construct a Markov chain with state space $\boldsymbol{\Theta}$ and $p(\boldsymbol{\theta}|\mathbf{x})$ as stationary or invariant distribution. If this chain is then run for long

enough, the simulated values can be treated as coming from the required target distribution, and again used in obtaining empirical estimates.

A Markov chain is generated by sampling for a new state of the chain based on the present state of the chain, independent of other past states. If the current state is $\boldsymbol{\theta}^{(t)}$, a new state is generated from a transition density that is only dependent upon $\boldsymbol{\theta}^{(t)}$,[2]

$$\boldsymbol{\theta}^{(t+1)} \sim \mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \ . \tag{1.18}$$

Density $\mathcal{K}$ may also be referred to as the transition kernel for the chain, and uniquely describes the dynamics of the chain.

In general we shall be interested in Markov chains over continuous state spaces. Under certain conditions that shall be expanded in chapter 5 (the Markov chain must be both periodic and irreducible), convergence of the chain will be to its stationary distribution,

$$\mathbb{P}(\boldsymbol{\theta}^{(t)} \in A) \to \int_A p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \quad \forall A \in \boldsymbol{\Theta}, \text{ as } t \to \infty \ . \tag{1.19}$$

The stationary distribution is unique when the entire state space can reasonably be explored; formally if any set of states can be reached from any other set of states within a finite number of transitions. Such a chain is irreducible, and if it has a stationary distribution $p(\boldsymbol{\theta}|\mathbf{x})$, we can assert the *ergodic theorem*, which states that the ergodic average from equation (1.15) will converge to the true expectation.

The stationary distribution is known, and MCMC methods require the construction of an appropriate transition kernel. A possible way to find such a kernel is to construct one that satisfies detailed balance,

$$p^*(\boldsymbol{\theta}^{(t)})\mathcal{K}(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = p^*(\boldsymbol{\theta}^{(t+1)})\mathcal{K}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)}) \tag{1.20}$$

Under the stationary distribution, we want the probability of 'moving forward from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$' to match the probability of 'moving back again'. Two methods for construcing Markov chains are described here.

Metropolis-Hastings. The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) samples from a Markov chain with $p(\mathbf{x}|\boldsymbol{\theta})$ as invariant distribution by making use of a proposal density $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ that depends on $\boldsymbol{\theta}^{(t)}$, the current state of the chain. A possible new state is generated from the proposal density, i.e. $\boldsymbol{\theta}^{\text{new}} \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. To decide whether to accept the new state, we determine a ratio of importance weights, and accept the new state and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{\text{new}}$ with probability

$$\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{\text{new}}) = \min\left(1, \frac{p^*(\boldsymbol{\theta}^{\text{new}})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\text{new}})}{p^*(\boldsymbol{\theta}^{(t)})q(\boldsymbol{\theta}^{\text{new}}|\boldsymbol{\theta}^{(t)})}\right) \tag{1.21}$$

and reject it (i.e. keep the current state with $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$) otherwise. When the support of q includes $\boldsymbol{\Theta}$, the resulting transition kernel

$$\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + [1 - \mathsf{acc}(\boldsymbol{\theta}^{(t)})]\delta(\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}) \tag{1.22}$$

satisfies the detailed balance condition (1.20) with p^*, and p^* (or rather its normalized version $p(\boldsymbol{\theta}|\mathbf{x})$) is a stationary distribution of the chain. The transition kernel consists

[2]For the sake of clarity in the spirit of Bayesian theory, we choose this 'conditional' notation for $\mathcal{K}$, rather than the more usual $\mathcal{K}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta})$.

of two terms, the first is the probability of generating a new point multiplied by the probability of accepting it, and the second is the probability of repeating the previous sample $\boldsymbol{\theta}^{(t)}$. Notation $\mathsf{acc}(\boldsymbol{\theta}^{(t)}) = \int \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) d\boldsymbol{\theta}$ indicates the average probability of accepting a new point, while $\delta(\cdot = \boldsymbol{\theta}^{(t)})$ indicates the Dirac delta mass at $\boldsymbol{\theta}^{(t)}$.

Gibbs sampling. Gibbs sampling is a powerful tool when we cannot sample directly from the joint distribution, but when sampling from the conditional distributions of each variable, or set of variables, is possible (Geman & Geman, 1984). If our parameters of interest are of multiple dimensions and can be divided as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, the Gibbs sampler uses the conditional distributions $p(\boldsymbol{\theta}_k|\{\boldsymbol{\theta}_j\}_{j\neq k})$ (if they can be sampled from directly) to draw a sample from the target distribution. Gibbs sampling may be interpreted as a Metropolis method with a sequence of always-accept proposal densities, all defined in terms of the conditional distributions of the target. Given $\boldsymbol{\theta}^{(t)}$, an iteration

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t+1)} &\sim p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_K^{(t)}) \\
\boldsymbol{\theta}_2^{(t+1)} &\sim p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_K^{(t)}) \\
\boldsymbol{\theta}_3^{(t+1)} &\sim p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_K^{(t)}), \quad \text{etc.}
\end{aligned} \tag{1.23}$$

samples a new state $\boldsymbol{\theta}^{(t+1)}$ in the chain.

A very large body of knowledge exits around Monte Carlo methods, and an introduction to its application to Machine Learning is given by Andrieu et al. (2003). The basis of the MH algorithm has been adapted into many variants. One algorithm that is particularly relevant to model averaging is the *Reversible Jump MCMC*—it is an extension to standard MH method to average over parameter spaces of different sizes, so that the Markov chain is run over different models $\mathcal{M}$ (Green, 1995).

A number of methods extend the parameter space so that a sample is taken from a joint distribution $p^*(\boldsymbol{\theta}, \mathbf{u})$, where the parameter space is extended with some additional auxiliary variables $\mathbf{u}$. Marginal samples $\boldsymbol{\theta}^{(t)}$ can then be obtained from sampling over $(\boldsymbol{\theta}^{(t)}, \mathbf{u}^{(t)})$, and ignoring the additional samples $\mathbf{u}^{(t)}$. Gibbs sampling for latent variable models, discussed in greater detail in section 1.4.2, also falls in this class of methods. The *hybrid Monte Carlo* algorithm, or *Hamiltonian Monte Carlo* algorithm, tries to avoid random walk behaviour by incorporating information about the gradient of the target distribution into the proposals through the auxiliary or "momentum" variables (Duane et al., 1987; MacKay, 2003). *Slice sampling* (Neal, 2003) uses auxiliary variables to draw uniform samples from the *volume* under $p^*(\boldsymbol{\theta})$, such that the pair $(\boldsymbol{\theta}^{(t)}, u^{(t)})$ defines a parameter sample and a height $0 < u^{(t)} < p^*(\boldsymbol{\theta}^{(t)})$.

The methods discussed above all relate to drawing samples from the posterior distribution, from which we can determine the expectations given in (1.3). When faced with estimating the marginal likelihood, we can simulate parallel chains at different temperatures, and use thermodynamic integration to estimate the log marginal likelihood. Chapter 4 discusses *parallel tempering* and thermodynamic integration, as well as a related method called *annealed importance sampling*, in greater depth.

1.4 Latent variable models

A key property of some complex posterior distributions over visible parameters $\boldsymbol{\theta}$ is that the addition of some hidden or latent parameters $\mathbf{z}$ can turn the distribution into an analytically

tractable form. The joint distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$ is first decomposed into the marginal distribution of the latent variables $p(\mathbf{z})$ and the conditional distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$. The latent variable marginal does not depend on the observed data, and can be referred to as some prior over the latent variables. (For the sake of clarity the dependence on the model assumptions $\mathcal{M}$ is dropped in the notation.) With this expansion of the parameter space to a joint distribution of visible and latent parameters, the corresponding distribution over the visible parameters can again be obtained by marginalization. The required marginal distribution—or parameter posterior—is then determined with

$$p(\boldsymbol{\theta}|\mathbf{x}) = \int p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) p(\mathbf{z}) \, d\mathbf{z} \ . \tag{1.24}$$

Except for very specific forms of the distributions $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$ and $p(\mathbf{z})$, this *marginalization* is in general not analytically tractable, as it may involve, for example, an exponential number of terms.

The goal of latent variables in this thesis is to extend the parameter space to allow for intractable distributions to be tractably treated. This is by no means their only use; dimensionality reduction, where the latent variables capture some underlying smaller-dimensional manifold, relies on similar methods, albeit with continuous latent variables (Bishop, 1999).

1.4.1 Mixtures of distributions

A mixture of distributions is a general framework for density modeling. It removes the restriction of fitting only unimodal distributions to data by allowing an arbitrary number of densities (or mixture components) to be scattered across the observed data, such that properties like the clustering of the data can be described well. Mixture models allow for a typical use of latent variables, where the latent variables capture the discrete component labels.

Let $\mathbf{x}_n$ come from a density model of the form

$$p(\mathbf{x}_n|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j) \ , \tag{1.25}$$

which is a mixture of J simpler parametric distributions. Our model choice $\mathcal{M}$ can specify the number of component distributions, their parametric form, etc. Parameters $\boldsymbol{\theta}$ can encompass all unknowns in the model: the parameters $\boldsymbol{\theta}_j$ of each of the component distributions, and possibly even the mixing coefficients π_j. The mixing coefficients are nonnegative and sum to one, $\sum_{j=1}^{J} \pi_j = 1$. Hence $p(\mathbf{x}_n|\boldsymbol{\theta})$ is nonnegative and integrates to unity if each of the individual components does.

The likelihood, which considers all possible partitions of the sample $\mathbf{x}$ into the J components and consequently expands exponentially into J^N terms, is

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \Big[\sum_{j=1}^{J} \pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j) \Big] \ . \tag{1.26}$$

Hidden latent variables $\mathbf{z} = \{z_{nj}\}$, where $z_{nj} = 1$ if component j was responsible for generating data point $\mathbf{x}_n$, and zero otherwise, naturally augment the data. This gives a much more manageable *complete-data* likelihood,

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{j=1}^{J} \Big[\pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j) \Big]^{z_{nj}} \ . \tag{1.27}$$

1.4.2 Gibbs sampling and latent variable models

The complete-data likelihood allows the conditional distributions $p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})$ to both be analytically tractable, giving rise to a classic two-stage Gibbs sampler that draws a sample $\{\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)}\}$ from $p(\boldsymbol{\theta},\mathbf{z}|\mathbf{x})$. Given $(\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)})$, an iteration

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &\sim p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z}^{(t)}),\\ \mathbf{z}^{(t+1)} &\sim p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}^{(t+1)})\end{aligned} \tag{1.28}$$

samples a new state $(\boldsymbol{\theta}^{(t+1)},\mathbf{z}^{(t+1)})$ in the chain.

The idea of sampling with *data augmentation* was originally introduced by Tanner & Wong (1987). For mixtures of Gaussian distributions, this was extended by Diebolt & Robert (1994) and others.

1.4.3 Variational Bayes and latent variable models

Variational inference—also called *ensemble learning*—is an alternative deterministic approximation scheme when exact inference for the posterior is intractable. As was discussed in section 1.3.1, the method arises from a particular case of α-divergence, by taking the limit $\alpha \to 0$ in (1.14). The method relies on a choice of a tractable family of distributions that are sufficiently flexible to give a good approximation to the posterior distribution; this approximation is achieved by minimizing the Kullback-Leibler (KL) divergence between the true and approximate posterior (Waterhouse et al., 1996). Although not confined to latent variable models, we shall restrict this section to the latent variable case, for which variational methods through an expectation maximization (EM) algorithm have proved to be extremely popular (see). A broader introduction to variational methods in graphical models is given by Jordan et al. (1999).

As the joint distribution is completed with latent variables, we restrict the approximations to $p(\mathbf{x},\mathbf{z},\boldsymbol{\theta})$ to be of factorized form, $sq(\boldsymbol{\theta})q(\mathbf{z})$. In the spirit of minimizing a divergence measure between a scaled approximating distribution and a joint distribution, the standard variational Bayesian framework will be approached from a slightly different angle. Following section 1.3.1, the KL divergence between the approximation and joint can be written as

$$\begin{aligned}\mathsf{KL}\big(sq(\boldsymbol{\theta})q(\mathbf{z}) \,\big\|\, p(\mathbf{x},\mathbf{z},\boldsymbol{\theta})\big) &= \int sq(\boldsymbol{\theta})q(\mathbf{z}) \ln \frac{sq(\boldsymbol{\theta})q(\mathbf{z})}{p(\boldsymbol{\theta},\mathbf{z}|\mathbf{x})p(\mathbf{x})}\, d\boldsymbol{\theta}d\mathbf{z} + \int p(\mathbf{x},\mathbf{z},\boldsymbol{\theta})\, d\boldsymbol{\theta}d\mathbf{z} - s\\ &= s\int q(\boldsymbol{\theta})q(\mathbf{z}) \ln \frac{q(\boldsymbol{\theta})q(\mathbf{z})}{p(\boldsymbol{\theta},\mathbf{z}|\mathbf{x})}\, d\boldsymbol{\theta}d\mathbf{z} + s\ln s - s\ln p(\mathbf{x}) + p(\mathbf{x}) - s\ .\end{aligned} \tag{1.29}$$

If we now set the partial derivative of the divergence with respect to scale s to zero and rearrange, we arrive at the usual *free energy* formulation (Feynman, 1972),

$$\ln s = -\mathsf{KL}\big(q(\boldsymbol{\theta})q(\mathbf{z}) \,\big\|\, p(\boldsymbol{\theta},\mathbf{z}|\mathbf{x})\big) + \ln p(\mathbf{x})\ . \tag{1.30}$$

From the nonnegativity of the KL divergence, $\ln s$ *lower bounds* the true evidence. In statistical physics, the negative $-\ln s$ would be equivalent to the *variational* free energy of a system that would be minimized, and $-\ln p(\mathbf{x})$ would be equivalent to the true free energy of the system. In fact, as we care about minimizing the free energy (or equivalently maximizing a *lower bound* on the evidence) we can write $\ln s$ as a function of $q(\boldsymbol{\theta})$ and $q(\mathbf{z})$. The objective function therefore measures the relative entropy between the approximating ensemble and the true distribution.

A popular algorithm for minimizing the free energy is the "variational Bayesian EM algorithm" (VBEM), which can be traced back to Hinton & van Camp (1993) and Neal & Hinton (1998)'s observation that EM algorithms can be viewed as variational free energy minimization methods. We can perform a free-form optimization over the two distributions $q(\boldsymbol{\theta})$ and $q(\mathbf{z})$ to give an *expectation* and *maximization* step, which is iteratively repeated until convergence,

$$q^{(t+1)}(\mathbf{z}) \propto \exp\left\{ \int q^{(t)}(\boldsymbol{\theta}) \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \right\} \tag{1.31}$$

$$q^{(t+1)}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \exp\left\{ \int q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \, d\mathbf{z} \right\} . \tag{1.32}$$

The algorithm follows from using calculus of variations to take the functional derivatives of $\ln s$ with respect to $q(\boldsymbol{\theta})$ and $q(\mathbf{z})$, while holding the other distribution fixed. The exact details of such a derivation follows in sections 2.5 and 2.9.3 in chapter 2.

Lower-bounding an integrand

A complementary interpretation of variational inference is to lower bound the integrand with a function that depends on some additional variational parameters (Saul et al., 1996; Jordan et al., 1999; Minka, 2001b). In other words, if we are interested in evaluating an intractable integral of the form $p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}) \, d\boldsymbol{\theta}$, an approximate solution can be found by lower-bounding the integrand $p(\boldsymbol{\theta}, \mathbf{x})$ with some function $g(\boldsymbol{\theta}, \boldsymbol{\phi})$, i.e.

$$g(\boldsymbol{\theta}, \boldsymbol{\phi}) \leq p(\boldsymbol{\theta}, \mathbf{x}) \quad \text{for all } \boldsymbol{\phi} \ , \tag{1.33}$$

where $\boldsymbol{\phi}$ are additional parameters chosen to make the integral $G = \int g(\boldsymbol{\theta}, \boldsymbol{\phi}) \, d\boldsymbol{\theta}$ tractable. In the process of making integral G—which is a lower bound on $p(\mathbf{x})$, the quantity of interest—as big as possible, a difficult integration problem has been turned into an optimization problem over parameters $\boldsymbol{\phi}$. We shall now use a *distribution* $q(\mathbf{z})$ instead of merely some variational parameters $\boldsymbol{\phi}$. Start by writing the integrand as a function of some latent variables, in this case $p(\boldsymbol{\theta}, \mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) \, d\mathbf{z}$. From Jensen's inequality we therefore have

$$p(\boldsymbol{\theta}, \mathbf{x}) = \exp\left\{ \ln \int q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{q(\mathbf{z})} \, d\mathbf{z} \right\} \geq \exp\left\{ \int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{q(\mathbf{z})} \, d\mathbf{z} \right\} \equiv g[\boldsymbol{\theta}, q(\mathbf{z})] \ . \tag{1.34}$$

The *function* that is integrated to find $p(\mathbf{x})$ is being bounded. The biggest lower bound $G \leq p(\mathbf{x})$ can be found by choosing some distribution $q(\mathbf{z})$ such that the integral $G = \int g[\boldsymbol{\theta}, q(\mathbf{z})] \, d\boldsymbol{\theta}$ is maximized. By writing

$$q(\boldsymbol{\theta}) \equiv \frac{g[\boldsymbol{\theta}, q(\mathbf{z})]}{\int g[\boldsymbol{\theta}, q(\mathbf{z})] \, d\boldsymbol{\theta}} = \frac{g[\boldsymbol{\theta}, q(\mathbf{z})]}{G} \ , \tag{1.35}$$

and substituting this distribution into the right hand side of (1.30), we find that $\ln G$ gives the usual free energy,

$$\ln s = \ln G \ , \tag{1.36}$$

and the EM algorithm given by (1.31) and (1.32) is again applicable.

1.5 Conclusion and summary of the remaining chapters

Problems of inference can be elegantly addressed with Bayes' theorem, but it typically requires the evaluation of large sums (often with an exponential number of terms) or intractable integrals.

There are various ways to practically address these difficulties, which include approximating the distribution in question with a simpler one, or using a MCMC sample to estimate unknown quantities.

This thesis investigates both these approaches in a latent variable setting. We give here a short summary of the rest of the thesis, with emphasis on new contributions made to the field of machine learning:

Chapter 2 introduces methods of approximate inference that rely on divergence measures. A simple mixture of Gaussians with unknown means is taken as a running toy example to fully illustrate Minka (2005)'s generic message passing algorithm with α-divergences over a factor graph. Both EP and VB can be seen as specific cases of this generic algorithm. The treatment of the illustrative example with VB or EP is well known, but

- we add the treatment of the illustrative example with α-divergences to the pool of knowledge, allowing us to interpolate between VB and EP and beyond. This particular algorithm is presented in sections 2.6 and 2.7.

In chapter 3 we use this as a base from which to compare deterministic and MCMC approaches to inference on real world problems.

- We proceed to give some new intuition on the effect of the *width* of the prior distribution to model pruning and local minima in VB (section 2.8) and why EP is not prone to the same behaviour.

In chapter 3 we extrapolate from these model-pruning results to increase the robustness of the message passing algorithm for VB.

We proceed to give a review of the various objective functions that are minimized for various choices of α, and discuss EP in terms of the *expectation consistent* framework for inference in section 2.9.

- Section 2.9.3 presents new analysis on VB message passing schemes over a factor graph, where updates are over separate factors, and not an entire distribution. We show that the algorithm behaves like the standard VBEM algorithm, where a lower bound on the marginal likelihood is always increased, *only* when the factors all obey a certain proportionality ratio.

Chapter 3 takes the ideas from chapter 2, and expands the toy example into a higher-dimensional mixture of Gaussians.

- This chapter contributes two new approaches to inference for a mixture of Gaussians, namely EP and the more general α-divergence message passing scheme. These algorithms are derived and combined in sections 3.3 to 3.6 into a single framework that is governed by a choice of $\alpha \geq 0$. The well known VB algorithm and the new mixture of Gaussians EP algorithm are both special cases at $\alpha = 0$ and $\alpha = 1$.

To investigate the merits of these approximate methods for prediction and model selection, experimental results are presented on a number of real life data sets, showing the approximate predictive distributions and log marginal likelihoods. As a benchmark, a comparison is also done with the results obtained with parallel tempering, a state of the art MCMC method presented in chapter 4. With EP we generally find closer log marginal

likelihood estimates than VB (which is based on a lower bound), and slightly better predictive distributions. It is shown empirically that the approximate methods tested here (message passing with $\alpha = 0, \frac{1}{2}, 1$) are well suited for model selection, and approximating the predictive distribution with high accuracy.

- In this chapter it is also practically shown that EP need not have a unique fixed point; if the fixed points are not unique, they depend on *both* the initialization and the random order in which factor refinements take place. Both these questions were posed by Minka (2001a).

Other points underlined empirically are: the log marginal likelihood estimates increase with α; the number of local solutions depends on the prior width; the discrepancy between the approximate and true log marginal likelihoods increase with model size; the marginal likelihoods give a characteristic 'Ockham hill' over increasing model size, thus providing a useful tool for model selection.

Chapter 4 presents parallel tempering and thermodynamic integration as methods to sample from multimodal posterior distributions, and determine normalizing constants. This chapter presents three main contributions to the field of inference.

- The first of these is a parallel tempered approach to sampling from a mixture of Gaussians posterior through Gibbs sampling (section 4.4.1).

The success of thermodynamic integration—from which we can estimate normalizing constants—depends on the effectiveness of a numerical interpolation of log likelihood averages. The interpolation is sensitive in regions of high temperature averages and around phase transitions.

- A suitable method of interpolation is proposed in section 4.2.1 to get numerically stable estimates for temperatures near infinity (or near-zero inverse temperatures).

Parallel tempering, as used in a Bayesian framework, is based on a careful interpolation between two distributions, slowly ranging from the prior distribution (at zero inverse temperature) to the full posterior distribution (at temperature of one).

- The third contribution made by chapter 4 is to change the interpolation between two distributions to be from a distribution with lower variance at zero inverse temperature, to the posterior. For reasons that follow in section 4.3, this change makes thermodynamic integration easier.

Chapter 5 takes some ideas from variational inference and applies them to the design of MCMC transition densities. We try to address a very basic question: armed with so many elegant methods of deterministic approximate inference, is it possible to build any into Monte Carlo samplers?

- A novel combination of deterministic and stochastic methods is made, and the result is a variational transition kernel for the MH algorithm.

The new kernel is a variational lower bound to an exact transition kernel. Unlike previous variational approaches to MCMC (de Freitas et al., 2001), the kernel is *adaptive*, and depends on a previous sample in a MH algorithm. Although theoretically pleasing, we highlight the apparent dangers of such a variational approach through an investigation into its effectiveness.

- It is finally shown with a discussion and proof in section 5.4.1 that the method need not be geometrically ergodic. This provides theoretical insight into why variational methods haven't made further inroads into Monte Carlo methods.

Chapter 6 provides a summary of contributions made by this thesis, and looks into future directions of research.

Chapter 2

Deterministic Approximate Inference

2.1 Introduction

This chapter focuses on finding an analytically tractable *approximation* to the joint distribution. Omitting the extra $\mathcal{M}$ for brevity (but knowing that we are still working with a chosen model, maybe from a set of models), the task at hand can be summarized with

$$p(\boldsymbol{\theta}, \mathbf{x}) = p(\mathbf{x})p(\boldsymbol{\theta}|\mathbf{x}) \approx sq(\boldsymbol{\theta}) \ . \tag{2.1}$$

In words, we would like to approximate the joint distribution $p(\boldsymbol{\theta}, \mathbf{x})$—a distribution that we cannot typically integrate over—with an easier distribution $q(\boldsymbol{\theta})$, appropriately scaled by s. (As $\mathbf{x}$ is observed, the joint distribution is unnormalized.) The posterior will then be approximated by q, allowing us to use it to make predictions or compute averages (as integrating over q should be an easier task). The scale s gives an approximation to the marginal likelihood, needed for model comparison and selection. We have effectively replaced an integration problem by an *optimization* problem: how to best fit $sq(\boldsymbol{\theta})$ to the joint distribution. A few unanswered questions remain, namely how to choose a parameterized $q(\boldsymbol{\theta})$, how to measure the goodness of fit, and finally how to find such a q.

This chapter approaches the problem from the viewpoint of a generic message-passing algorithm, which is an intuitively appealing way of finding such a q (Minka, 2005). After choosing the functional form of q and an objective function, we are by no means restricted to the scheme presented here. Depending on the objective function, variational Bayes or more sophisticated double loop algorithms (Opper & Winther, 2005a) can also be implemented. Such a discussion is best left to section 2.9.

A mixture of Gaussian distributions is chosen as a running example to first illustrate how one data point likelihood (or factor) can be exactly approximated, and finally how this approach can be extended to many observations, or a general factor graph. The illustrative model is the simplest non-trivial latent variable model, for example giving multimodal posteriors. The approach to latent variable modeling taken here can be traced back to Dempster et al. (1977)'s seminal paper on expectation maximization (EM), which has stimulated many further developments in latent variable modeling. Through data completion, a parameter estimate is found that (locally) maximizes the likelihood. The EM algorithm can be generalized to a variational Bayes (VB) EM algorithm (Neal & Hinton, 1998), allowing us to work with posterior parameter distributions rather than parameter point estimates, and overcoming some possible singularities present in EM. By choosing delta functions as posterior approximating distributions, EM for maximum a posteriori learning can be recovered. A mixture of Gaussians was typically taken as example

implementation (Attias, 1999). The chapter emphasizes VBEM as again being a specific case of the larger class of approximate methods, and we can recover the VB objective in the limiting case $\alpha \to 0$, where the role of α is left to discussion in section 2.2.

The rest of the chapter follows with a review of divergence measures, focussing on the α-divergence. A simple illustration of a one-dimensional mixture of Gaussians (section 2.3), with all parameters but the means known, is taken as running example. By focusing on only one observation, we can derive an *exact* solution for s and $q(\boldsymbol{\theta})$ for different measures of divergence in sections 2.4 to 2.6. The results for tackling this toy problem with VB and EP are both well known, but the use of α-divergences is new in this arena. Except for VB, and exact solution for s and $q(\boldsymbol{\theta})$ cannot typically be found if we are faced with an abundance of data (in the case of mixture models we are faced with an exponential number of terms)—in other words the 'global' divergence cannot be minimized directly. By again focussing on single observations, we can still minimize 'local' divergences. We can therefore add more data to the tractable 'single observation' case, and derive a general optimization scheme over a factor graph. This is illustrated in section 2.7, with variational Bayes and expectation propagation included as special cases. Unwanted model pruning is discussed in section 2.8, while section 2.9 concludes with a discussion on the objective functions of all these algorithms.

2.2 Divergence measures

A divergence measure quantifies the goodness of fit of one distribution to another. The family of divergence measures used here is the α-divergence (Amari, 1985; Minka, 2005), indexed by a continuous parameter $\alpha \in \mathbb{R}$. The global α-divergence is

$$D_\alpha\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) = \frac{\int \alpha p(\mathbf{x},\boldsymbol{\theta}) + (1-\alpha) sq(\boldsymbol{\theta}) - p(\mathbf{x},\boldsymbol{\theta})^\alpha [sq(\boldsymbol{\theta})]^{1-\alpha} d\boldsymbol{\theta}}{\alpha(1-\alpha)} \,. \tag{2.2}$$

Notice that neither p nor sq is normalized; for our purposes we let q remain normalized, so that we can easily read off an approximate posterior and marginal likelihood estimate. A special case of the α-divergence is the Kullback-Leibler (KL) divergence,

$$\mathsf{KL}\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) = \int p(\mathbf{x},\boldsymbol{\theta}) \ln \frac{p(\mathbf{x},\boldsymbol{\theta})}{sq(\boldsymbol{\theta})}\, d\boldsymbol{\theta} + \int \Big(sq(\boldsymbol{\theta}) - p(\mathbf{x},\boldsymbol{\theta}) \Big)\, d\boldsymbol{\theta}, \tag{2.3}$$

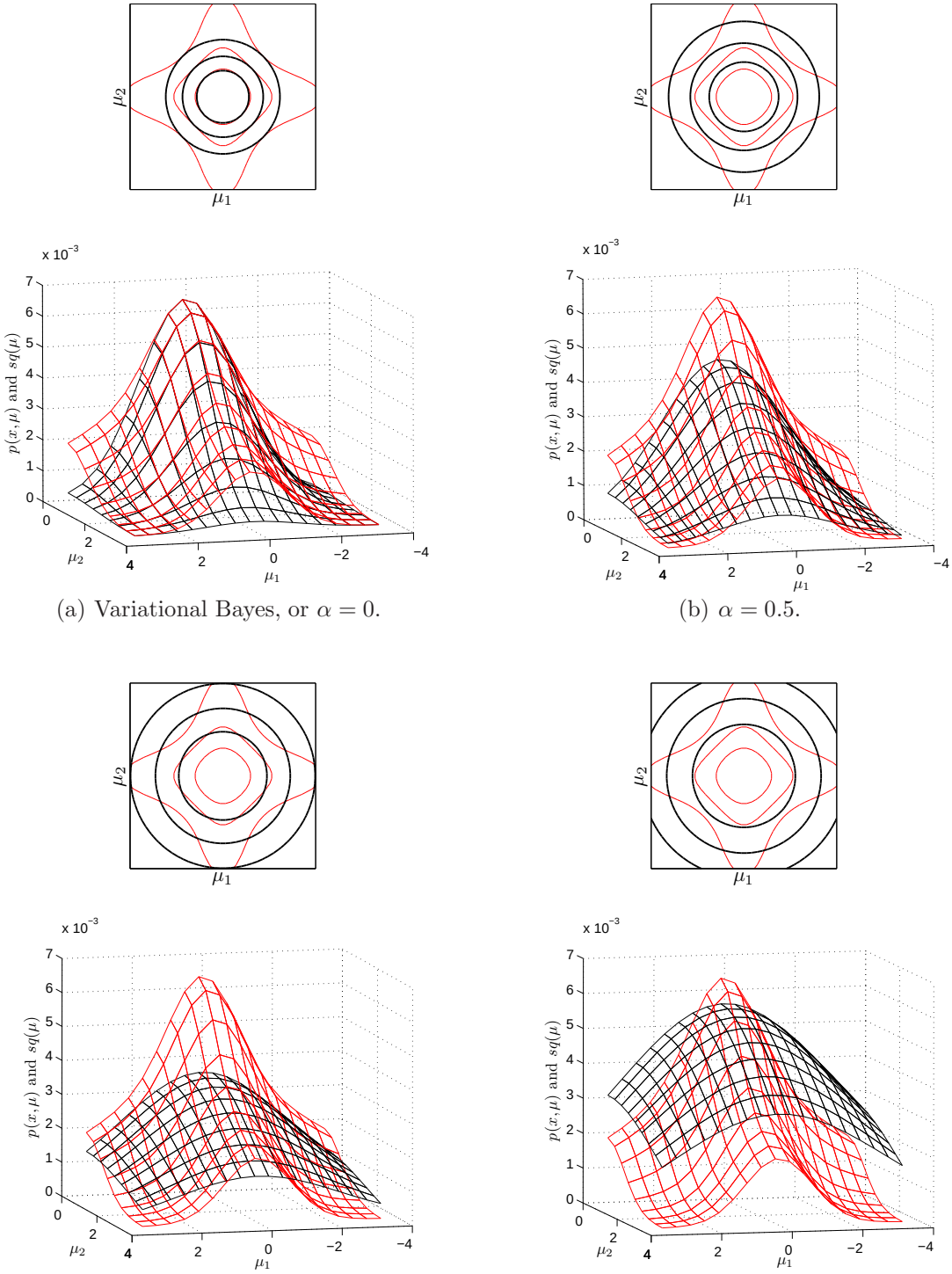
which is asymmetric with respect to p and sq. The correction factor added to the usual KL divergence follows from its application here to unnormalized distributions as well. The divergence follows from taking the limit,

$$\lim_{\alpha\to 1} D_\alpha\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) = \mathsf{KL}\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) \tag{2.4}$$

$$\lim_{\alpha\to 0} D_\alpha\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) = \mathsf{KL}\big(sq(\boldsymbol{\theta}) \,\big\|\, p(\mathbf{x},\boldsymbol{\theta})\big) \,, \tag{2.5}$$

which we formally show in appendix A.1. The divergence measure used in EP, sometimes referred to as the 'inclusive' KL divergence (Frey et al., 2000), is given by (2.4). Taking the limit to zero gives the 'exclusive' KL divergence in (2.5), which is used in VB.

The α-divergence is convex with respect to $p(\mathbf{x},\boldsymbol{\theta})$ and $sq(\boldsymbol{\theta})$, zero if and only if $p(\mathbf{x},\boldsymbol{\theta}) = sq(\boldsymbol{\theta})$, and positive otherwise. Sections 2.4, 2.5, and 2.6 describe how to obtain an exact minimum for a single observation, illustrated with a mixture of Gaussians problem. We preview how such a solution will look in figure 2.1: The joint distribution is fitted with with a product of two Gaussians with adjustable mean, precision (inverse variance) and scale. The unknown



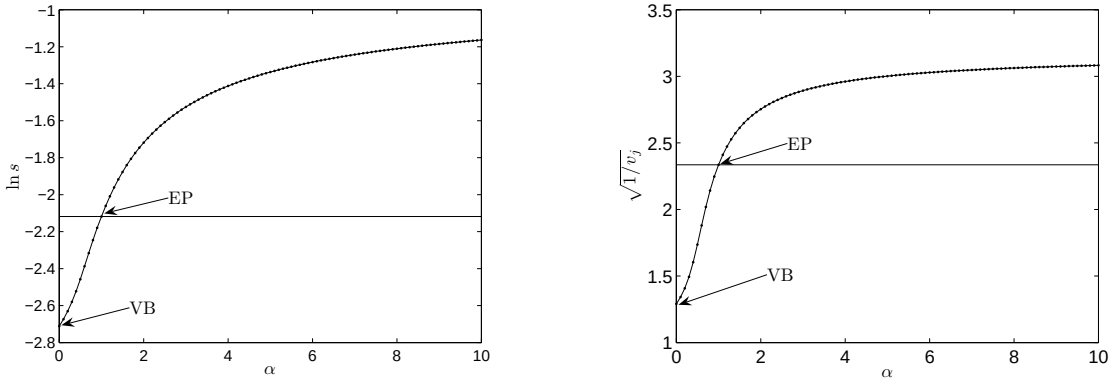
(a) Variational Bayes, or $\alpha = 0$.

(b) $\alpha = 0.5$.

(c) Expectation propagation, or $\alpha = 1$.

(d) $\alpha = 10$.

FIGURE 2.1: With one data point $x_n = 0$, the figures illustrate a simple mixture of two Gaussians with unknown means $\boldsymbol{\mu} = \{\mu_1, \mu_2\}$, as given in (2.9). The complete joint $p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)$ from (2.25) was approximated with $sq(\boldsymbol{\mu})q(\mathbf{z}_n)$. The marginal of interest, $p(x_n, \boldsymbol{\mu})$, is plotted in red, and its approximation $sq(\boldsymbol{\mu})$ is plotted in black. The mixing weights were fixed to $\pi_j = \frac{1}{2}$, and the precisions to $\lambda_j = 1$, for components $j \in \{1, 2\}$. The prior hyperparameter values were set to $v_{0j} = 0.1$ and $m_{0j} = 0$.



(a) $\ln s$, the log marginal likelihood estimate. This illustrates Theorem 1, which states that $\ln s$ is non-decreasing as a function of α, when we can exactly minimize the α-divergence.

(b) $\sqrt{1/v_j}$, the standard deviation estimate. Notice the underestimation of the true variance by Variational Bayes, and zero-forcing divergences in general.

FIGURE 2.2: The log scale $\ln s$ and the standard deviation of the product of two Gaussians that minimize the α-divergence to $p(\mathbf{x}, \boldsymbol{\theta})$. This figure follows the same example of figure 2.1.

parameters in the joint distribution were the two component means in (2.9). From figure 2.2 we observe that $\alpha = 0$, and indeed all $\alpha < 1$, lower bounds the marginal likelihood when an exact minimization is possible. When the minimization of an objective function is performed on a factor graph, $\alpha = 0$ (VB) still provides a bound.

For the case of $\alpha = 0$ in figure 2.1, the approximation $sq(\boldsymbol{\theta})$ also *lower bounds* the function $p(\mathbf{x}, \boldsymbol{\theta})$. It is not a property of the KL divergence, but here comes as a result of explicitly constructing a lower bound that relies on an extra 'variational' distribution $q(\mathbf{z})$. Section 1.4.3 describes the lower bound, and how it relates to VB and the EM algorithm in particular.

A divergence with $\alpha \leq 0$ is referred to a zero-forcing divergence, for when the joint distribution is zero, the scaled approximation is forced to be zero too. Consequently some non-zero parts of the joint distribution may be excluded, hence the name 'exclusive' KL divergence. From figure 2.2 it is evident that zero-forcing divergences tend to underestimate the true variance. As α grows[1], the scaled approximating Gaussian smoothly expands until it covers the entire joint distribution for $\alpha \to \infty$. As the approximation expands as much of the joint distribution as possible is included; $\alpha \geq 1$ requires the approximation to be nonzero whenever the joint is nonzero, hence the KL divergence is 'inclusive'. Varying α between zero and one blends the properties of the inclusive and exclusive KL divergences.

Figure 2.2 illustrates the approximate log marginal likelihood $\ln s$ as a function of α for the two-mean joint distribution of figure 2.1. The scale s monotonically increases with α, with $\alpha = 1$ giving the true marginal likelihood. This result applies only when an *exact* minimization is possible (for many observations we shall later do an approximate minimization over a factor graph). The following result, given without proof, confirms this observation.

Theorem 1. *(Minka, 2005). When $sq(\boldsymbol{\theta})$ minimizes $D_\alpha(p(\mathbf{x}, \boldsymbol{\theta}) \| sq(\boldsymbol{\theta}))$, then s is monotonically*

[1]Section 2.6's optimization routine is only valid for nonnegative α, as it includes $\sqrt{\alpha}$. Therefore, for the mixtures problem we are concerned with, only nonnegative αs are illustrated. This does not preclude the use of $\alpha < 0$ to other problems.

increasing as a function of α. Consequently

$$s \leq p(\mathbf{x}) \qquad \text{if } \alpha < 1 \tag{2.6}$$

$$s = p(\mathbf{x}) \qquad \text{if } \alpha = 1 \tag{2.7}$$

$$s \geq p(\mathbf{x}) \qquad \text{if } \alpha > 1 \tag{2.8}$$

Our attention shall now be turned to figure 2.1 as an illustrative case, and the following sections shall use it as a toy example in aid of explaining methods to minimize $D_\alpha(p(\mathbf{x}, \boldsymbol{\theta}) \| sq(\boldsymbol{\theta}))$.

2.3 A simple mixture of Gaussians

As a simple illustration of different divergence measures, consider a one-dimensional Gaussian mixture with unknown means, so that $\boldsymbol{\theta} \equiv \boldsymbol{\mu}$,

$$p(x_n|\boldsymbol{\mu}) = \sum_{j=1}^{J} \pi_j \mathcal{N}(x_n|\mu_j, \lambda_j^{-1}) \ , \tag{2.9}$$

where

$$\mathcal{N}(x_n|\mu_j, \lambda_j^{-1}) = \left(\frac{\lambda_j}{2\pi}\right)^{1/2} e^{-\frac{1}{2}\lambda_j(x_n-\mu_j)^2} = \frac{1}{\mathcal{Z}_{\mathcal{N}}(\lambda_j)} e^{-\frac{1}{2}\lambda_j(x_n-\mu_j)^2} \ . \tag{2.10}$$

In the above mixture of J Gaussians, we let each precision (inverse variance) λ_j, as well as the mixing weights $\boldsymbol{\pi}$, be known. The unknown parameters $\boldsymbol{\theta}$ are therefore the set of means $\boldsymbol{\mu} = \{\mu_j\}_{j=1}^J$. Let the prior on the means be conjugate and hence Gaussian,

$$p(\boldsymbol{\mu}) = \prod_{j=1}^{J} p(\mu_j) = \prod_{j=1}^{J} \mathcal{N}(\mu_j \mid m_{0j}, v_{0j}^{-1}) \ . \tag{2.11}$$

For q we choose a product of Gaussians, one each to model the mean of a component in the mixture,

$$q(\boldsymbol{\mu}) = \prod_{j=1}^{J} q(\mu_j) = \prod_{j=1}^{J} \mathcal{N}(\mu_j \mid m_j, v_j^{-1}) \ . \tag{2.12}$$

In many cases our choice of approximating distribution will be restricted by the model. For one observation (let it be x_n, for instance) our task is to match $sq(\boldsymbol{\mu}) \approx p(x_n|\boldsymbol{\mu})p(\boldsymbol{\mu})$ We can directly minimize both KL divergences, but need to resort to an iterative method to minimize other α-divergences. As the cases of $\alpha = 1$ and $\alpha = 0$ ultimately expand respectively into expectation propagation and variational Bayes, the sections that follow here are appropriately headed.

As an interlude, it is worthwhile to visualize the joint distribution in a graphical representation, given in figure 2.3. The following two equations are illustrated, where the first gives the parameter dependencies, while the second gives a chosen factorization,

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(x_n|\boldsymbol{\mu}, \mathbf{z}) p(\mathbf{z}) p(\boldsymbol{\mu}) &= \prod_{n=1}^{N} \prod_{j=1}^{J} \mathcal{N}\big(x_n|\mu_j, \lambda_j^{-1}\big)^{z_{nj}} \times \pi_j^{z_{nj}} \times \mathcal{N}\big(\mu_j|m_{0j}, v_{0j}^{-1}\big) \\ &= \prod_{n=1}^{N} f_n(\boldsymbol{\mu}, \mathbf{z}_n) \times f_0(\boldsymbol{\mu}) \ . \end{aligned} \tag{2.13}$$

Each of the factors in the factor graph will ultimately be approximated.

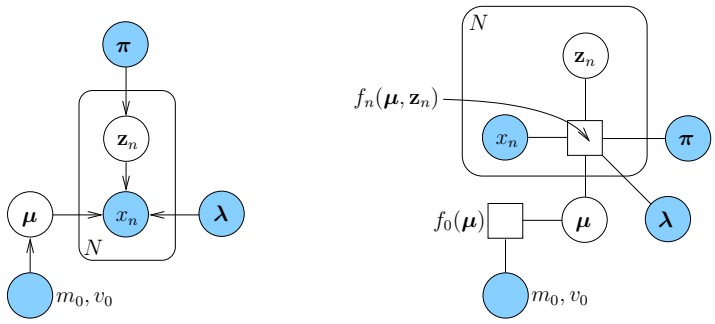


FIGURE 2.3: The structure of a probabilistic model can be made lucid through a graphical representation. On the *left* we have an acyclic graph or Bayesian network, illustrating the parameter dependencies in the joint $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ in (2.13). The box is called a *plate*, and indicates N replicates of the random variables x_n and $\mathbf{z}_n$. Nodes that are coloured indicate observed random variables, and uncoloured nodes indicate variables that we want to average over, marginalize away, etc. The manner in which the factors in our chosen factorization of the joint distribution depend on the parameters can be equally illustrated through a factor graph, shown on the *right*. In this case the square indicates a factor that is dependent on a number of observed and hidden variables.

The direction forward. At this point it is useful to draw an outline of what we are trying to achieve. As we have seen in figure 2.2, we can find an exact solution for $sq(\boldsymbol{\theta})$ for a 'single observation' case, as the partition function (scale) is tractable for $\alpha = 1$. We can exactly minimize D_α, as will be explained in sections 2.4 to 2.6. In section 2.7 we shall motive an algorithm that repeatedly performs these minimizations over a factor graph. When faced with a factor graph, the global D_α (i.e. with many observations in the likelihood) will *not* be minimized, but rather a related objective function given in section 2.9 (the exception to this rule is VB). As a result we are still left with useful algorithms, although convergence cannot always be guaranteed (VB with $\alpha = 0$ is again an exception).

We could have turned the order presented here on its head, and motivated an objective function, for which algorithms can be derived. In section 2.9 we show solid reason behind the $\alpha = 1$ objective function, giving the *expectation consistent* framework. From this perspective EP is but one algorithm to minimize the objective function, and other algorithms (for example double loop, which comes with a convergence guarantee) can be derived with the same objective in mind.

First, though, we concern ourselves with a prior and one likelihood factor.

2.4 Expectation propagation: a single observation

The KL divergence, as a function of s and $\{m_j, v_j\}_{j=1}^J$, the parameters of $q(\boldsymbol{\mu})$, is

$$\mathsf{KL}\big(p(x_n, \boldsymbol{\mu}) \,\big\|\, sq(\boldsymbol{\mu})\big) = s - \int p(x_n, \boldsymbol{\mu}) \ln[sq(\boldsymbol{\mu})] \, d\boldsymbol{\mu} + \mathsf{const} \ . \tag{2.14}$$

For completeness, detailed derivations to minimize the KL divergence are presented here and in appendix A.2. Chapter 3's results, for which full derivations are not given, follow exactly the same style.

2.4.1 The scale

Taking derivatives of (2.14) with respect to s, and equating to zero, gives $s = \int p(x_n, \boldsymbol{\mu})\, d\boldsymbol{\mu}$. From observing one example, we can directly write down the marginal likelihood as a function of the prior parameter values,

$$s = s(m_0, v_0) = \sum_{j=1}^{J} \pi_j \int p(x_n|\mu_j)p(\mu_j)\, d\mu_j = \sum_{j=1}^{J} \pi_j \mathcal{N}(x_n|m_{0j}, \lambda_{0j}^{-1} + v_{0j}^{-1}) \ . \tag{2.15}$$

2.4.2 Parameter updates for the components

The parameter updates of each approximate distributions $q(\mu_j)$ will take the form of a weighted sum of the prior and component-posterior moments, which has an intuitively pleasing explanation: The moments of $q(\mu_j)$ are calculated by determining the probability of component j generating x_n, multiplied by the moment of including x_n into component j, *plus* the probability of component j *not* generating x_n, multiplied by the prior moment of component j. The moment-matching equations can be determined with

$$\partial \mathsf{KL}\big(p(x_n, \boldsymbol{\mu}) \,\big\|\, sq(\boldsymbol{\mu})\big)/\partial m_j = 0 \ , \tag{2.16}$$

which we derive in appendix A.2. In this vein, define the responsibilities as

$$r_{nj} = \frac{\pi_j \int p(\mu_j)p(x_n|\mu_j)\, d\mu_j}{\sum_k \pi_k \int p(\mu_k)p(x_n|\mu_k)\, d\mu_k} = \frac{\pi_j \mathcal{N}(x_n|m_{0j}, \lambda_j^{-1} + v_{0j}^{-1})}{\sum_k \pi_k \mathcal{N}(x_n|m_{0k}, \lambda_k^{-1} + v_{0k}^{-1})} \ , \tag{2.17}$$

so that the mean of the approximation is therefore a responsibility-weighted sum of the prior and component-posterior means, or a *weighted sum of moments*,

$$\begin{aligned} m_j &= (1 - r_{nj}) \int \mu_j p(\mu_j)\, d\mu_j + r_{nj} \int \mu_j p(\mu_j|x_n)\, d\mu_j \\ &= (1 - r_{nj})\langle \mu_j \rangle + r_{nj} \langle \mu_j | x_n \rangle \ . \end{aligned} \tag{2.18}$$

Exactly the same can be done for the precision parameters. Differentiating the KL divergence with respect to v_j gives (following the same type of arrangement of terms as we have done for the means),

$$\frac{1}{v_j} = (1 - r_{nj}) \int (\mu_j - m_j)^2 p(\mu_j)\, d\mu_j + r_{nj} \int (\mu_j - m_j)^2 p(\mu_j|x_n)\, d\mu_j \ . \tag{2.19}$$

By substituting the value of m_j, we arrive at the second of the elegant *weighted moment-matching* equations,

$$\frac{1}{v_j} = (1 - r_{nj})\langle \mu_j^2 \rangle + r_{nj} \langle \mu_j^2 | x_n \rangle - m_j^2 \ . \tag{2.20}$$

The following expectations are used in the update to get an approximation $q(\mu_j) = \mathcal{N}(\mu_j|m_j, v_j^{-1})$,

$$\langle \mu_j \rangle = m_{0j} \tag{2.21}$$

$$\langle \mu_j | x_n \rangle = \frac{\lambda_j x_n + v_{0j} m_{0j}}{\lambda_j + v_{0j}} \tag{2.22}$$

$$\langle \mu_j^2 \rangle = \mathsf{var}(\mu_j) + \langle \mu_j \rangle^2 = \frac{1}{v_{0j}} + m_{0j}^2 \tag{2.23}$$

$$\langle \mu_j^2 | x_n \rangle = \mathsf{var}(\mu_j|x_n) + \langle \mu_j | x_n \rangle^2 = \frac{1}{\lambda_j + v_{0j}} + \Big(\frac{\lambda_{0j} x_n + v_{0j} m_{0j}}{\lambda_j + v_{0j}}\Big)^2 \ . \tag{2.24}$$

2.5 Variational Bayes: a single observation

For VB we introduce latent allocation variables, so that the sum in the joint distribution becomes a product. Let $\mathbf{z}_n \in \{0,1\}^J$ be a binary latent variable, with $\sum_{j=1}^J z_{nj} = 1$, indicating which component in the mixture generated the data point. Therefore

$$p(x_n, \boldsymbol{\mu}, \mathbf{z}_n) = p(x_n|\boldsymbol{\mu}, \mathbf{z}_n)p(\mathbf{z}_n)p(\boldsymbol{\mu}) = \prod_{j=1}^J p(x_n|\mu_j)^{z_{nj}} \times \prod_{j=1}^J \pi_j^{z_{nj}} \times p(\boldsymbol{\mu}) \ . \tag{2.25}$$

For the α-divergence minimization in section 2.6, this method of data completion is also used. We could equally have done it for EP as well: the result will be exactly the same, as substituting $\alpha = 1$ in section 2.6's fixed point scheme clearly shows.

The joint distribution will be approximated with $sq(\boldsymbol{\mu})q(\mathbf{z}_n)$, where $q(\mathbf{z}_n)$ is multinomial,

$$q(\mathbf{z}_n) = \prod_{j=1}^J \gamma_{nj}^{z_{nj}} \tag{2.26}$$

with $\gamma_{nj} \geq 0$ and $\sum_j \gamma_{nj} = 1$.

A mean field approximation to our simple posterior can be found by 'reversing the KL divergence' to its 'exclusive' form. Again, we write the KL divergence as a function of s and the parameters of $q(\boldsymbol{\mu})$ and $q(\mathbf{z}_n)$:

$$\begin{aligned}\mathsf{KL}\big(sq(\boldsymbol{\mu})q(\mathbf{z}_n) \,\|\, p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)\big) &= \int \sum_{\mathbf{z}_n} sq(\boldsymbol{\mu})q(\mathbf{z}_n) \ln \frac{sq(\boldsymbol{\mu})q(\mathbf{z}_n)}{p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)} \, d\boldsymbol{\mu} - s + \mathsf{const} \\ &= s \ln s + s\langle \ln q(\boldsymbol{\mu})\rangle + s\langle \ln q(\mathbf{z}_n)\rangle \\ &\quad - s\langle \ln p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)\rangle - s + \mathsf{const} \ . \end{aligned} \tag{2.27}$$

2.5.1 Parameter updates

To optimize over distributions $q(\boldsymbol{\mu})$ and $q(\mathbf{z}_n)$, take the functional derivative of the KL divergence (2.27) with respect to $q(\boldsymbol{\mu})$ and $q(\mathbf{z}_n)$, in each case equating to zero and solving. It is shown below that we arrive at an *iterative* optimization procedure, which is a single-observation implementation of VBEM. In essence we are doing an iterative coordinate descent procedure over functions (distributions) q, which will converge to a local minimum as each of the subproblems is convex. The following E- and M-steps are repeated until convergence.

E-step. For the *expectation* step, we keep $q(\boldsymbol{\mu})$ fixed. Zeroing the functional derivative of (2.27) with respect to $q(\mathbf{z}_n)$ gives

$$q(\mathbf{z}_n) \propto \exp\Big\{ \int q(\boldsymbol{\mu}) \ln p(x_n, \boldsymbol{\mu}, \mathbf{z}_n) \, d\boldsymbol{\mu} \Big\} \ . \tag{2.28}$$

Because we can rewrite $\ln p(x_n, \mathbf{z}_n, \boldsymbol{\mu})$ as $\ln p(x_n, \mathbf{z}_n|\boldsymbol{\mu}) + \ln p(\boldsymbol{\mu})$, the above equation can be simplified as $q(\mathbf{z}_n) \propto \exp\{\int q(\boldsymbol{\mu}) \ln p(x_n, \mathbf{z}_n|\boldsymbol{\mu}) \, d\boldsymbol{\mu}\}$. If m_j and v_j are the present parameters of $q(\mu_j)$ (we can start with a guess, e.g. set to the prior), then

$$\int q(\boldsymbol{\mu}) \ln p(x_n, \mathbf{z}_n|\boldsymbol{\mu}) \, d\boldsymbol{\mu} = \sum_{j=1}^J z_{nj} \int \Big[\ln \pi_j - \ln \mathcal{Z}_{\mathcal{N}}(\lambda_j) - \frac{\lambda_j}{2}(x_n - \mu_j)^2 \Big] q(\mu_j) \, d\mu_j \ , \tag{2.29}$$

and hence the responsibilities, characterizing $q(\mathbf{z}_n)$, will be

$$\gamma_{nj} = \frac{\pi_j \sqrt{\lambda_j} \exp\{-\frac{\lambda_j}{2}(v_j^{-1} + (m_j - x_n)^2)\}}{\sum_k \pi_k \sqrt{\lambda_k} \exp\{-\frac{\lambda_k}{2}(v_k^{-1} + (m_k - x_n)^2)\}} \ . \tag{2.30}$$

M-step. For the *maximization* step, the derivation of the E-step is repeated, only with $q(\boldsymbol{\mu})$ and $q(\mathbf{z}_n)$ swapping roles,

$$q(\boldsymbol{\mu}) \propto \exp\Big\{ \sum_{\mathbf{z}_n} q(\mathbf{z}_n) \ln p(x_n, \boldsymbol{\mu}, \mathbf{z}_n) \Big\} \ . \tag{2.31}$$

To update the parameters, note that

$$\begin{aligned} \sum_{\mathbf{z}_n} q(\mathbf{z}_n) \ln \big[p(x_n, \mathbf{z}_n|\boldsymbol{\mu}) p(\boldsymbol{\mu})\big] &= -\sum_{\mathbf{z}_n} \prod_{i=1}^{J} \gamma_{ni}^{z_{ni}} \sum_{j=1}^{J} z_{nj} \frac{\lambda_j}{2}(x_n - \mu_j)^2 - \sum_{j=1}^{J} \frac{v_{0j}}{2}(\mu_j - m_{0j})^2 + \mathsf{const} \\ &= -\sum_{j=1}^{J} \frac{1}{2}(v_{0j} + \gamma_{nj}\lambda_j)\Big(\mu_j - \frac{v_{0j}m_{0j} + \gamma_{nj}\lambda_j x_n}{v_{0j} + \gamma_{nj}\lambda_j}\Big)^2 + \mathsf{const} \ . \end{aligned} \tag{2.32}$$

This is in the form of an unnormalized $q(\boldsymbol{\mu})$, and hence the parameter updates are

$$v_j = v_{0j} + \gamma_{nj}\lambda_j \tag{2.33}$$

$$m_j = \frac{v_{0j}m_{0j} + \gamma_{nj}\lambda_j x_n}{v_{0j} + \gamma_{nj}\lambda_j} \ . \tag{2.34}$$

2.5.2 The scale

After optimizing for the parameters of q, we find the matching scale s by taking the partial derivative of (2.27) with respect to s and equating it to zero. In this case log scale $\ln s$ also corresponds to the negative variational free energy from mean field methods or statistical physics. The approximation to the log marginal likelihood is therefore

$$\ln s = \langle \ln p(x_n, \boldsymbol{\mu}, \mathbf{z}_n) \rangle - \langle \ln q(\boldsymbol{\mu}) \rangle - \langle \ln q(\mathbf{z}_n) \rangle \ . \tag{2.35}$$

This we determine from the following equations:

$$\begin{aligned} \langle \ln p(x_n, \boldsymbol{\mu}, \mathbf{z}_n) \rangle &= \Big\langle \sum_{j=1}^{J} z_{nj} \ln p(x_n|\mu_j) \Big\rangle + \Big\langle \sum_{j=1}^{J} z_{nj} \ln \pi_j \Big\rangle + \Big\langle \sum_{j=1}^{J} \ln p(\mu_j) \Big\rangle \\ &= \sum_{j=1}^{J} \gamma_{nj} \Big[-\ln \mathcal{Z}_{\mathcal{N}}(\lambda_j) - \frac{\lambda_j}{2}\big(\frac{1}{v_j} + (m_j - x_n)^2\big) \Big] \\ &\quad + \sum_{j=1}^{J} \gamma_{nj} \ln \pi_j + \sum_{j=1}^{J} \Big[-\ln \mathcal{Z}_{\mathcal{N}}(v_{0j}) - \frac{v_{0j}}{2}\big(\frac{1}{v_j} + (m_j - m_{0j})^2\big) \Big] \end{aligned} \tag{2.36}$$

$$\langle \ln q(\boldsymbol{\mu}) \rangle = \sum_{j=1}^{J} \langle \ln q(\mu_j) \rangle = -\sum_{j=1}^{J} \ln \mathcal{Z}_{\mathcal{N}}(v_j) - \frac{J}{2} \tag{2.37}$$

$$\langle \ln q(\mathbf{z}_n) \rangle = \sum_{j=1}^{J} \gamma_{nj} \ln \gamma_{nj} \ . \tag{2.38}$$

2.6 α-divergence: a single observation

More generally we seek a $sq(\boldsymbol{\theta})$ that minimizes $D_\alpha(p(\mathbf{x},\boldsymbol{\theta}) \parallel sq(\boldsymbol{\theta}))$, and this section presents a review on how an α-divergence can be minimized by repeatedly minimizing a KL divergence. The review is followed by a new contribution in section 2.6.1 on minimizing D_α for a simple mixtures problem. This contribution is reworked in section 3.5 to provide a key step for a new factor graph algorithm for full-blown mixture of Gaussians.

We let $sq(\boldsymbol{\theta})$ be a member of some scaled exponential family $\mathcal{F}$, for example the family of scaled Gaussians. The following theorem provides a key to finding the minimum:

Theorem 2. *(Minka, 2005). If $\alpha \neq 0$ then $sq(\boldsymbol{\theta})$ is a stationary point of*

$$D_\alpha(p(\mathbf{x},\boldsymbol{\theta}) \parallel sq(\boldsymbol{\theta})) \tag{2.39}$$

if and only if $sq(\boldsymbol{\theta})$ is a stationary point of

$$\arg\min_{sq(\boldsymbol{\theta})\in\mathcal{F}} \mathsf{KL}\Big(p(\mathbf{x},\boldsymbol{\theta})^\alpha [sq(\boldsymbol{\theta})]^{1-\alpha} \parallel sq(\boldsymbol{\theta})\Big) \ . \tag{2.40}$$

Proof. We show that the derivatives match at $sq(\boldsymbol{\theta}) = s_* q_*(\boldsymbol{\theta})$, i.e.

$$\left.\frac{\partial \mathsf{KL}(p(\mathbf{x},\boldsymbol{\theta})^\alpha [s_* q_*(\boldsymbol{\theta})]^{1-\alpha} \parallel sq(\boldsymbol{\theta}))}{\partial [sq(\boldsymbol{\theta})]}\right|_{sq(\boldsymbol{\theta})=s_* q_*(\boldsymbol{\theta})} \propto \left.\frac{\partial D_\alpha(p(\mathbf{x},\boldsymbol{\theta}) \parallel sq(\boldsymbol{\theta}))}{\partial [sq(\boldsymbol{\theta})]}\right|_{sq(\boldsymbol{\theta})=s_* q_*(\boldsymbol{\theta})} , \tag{2.41}$$

which we show by taking functional derivatives with respect to $sq(\boldsymbol{\theta})$, and substituting $sq(\boldsymbol{\theta}) = s_* q_*(\boldsymbol{\theta})$,

$$\frac{\partial D_\alpha(p(\mathbf{x},\boldsymbol{\theta}) \parallel sq(\boldsymbol{\theta}))}{\partial [sq(\boldsymbol{\theta})]} = \frac{1}{\alpha}\Big[1 - \int \Big(\frac{p(\mathbf{x},\boldsymbol{\theta})}{sq(\boldsymbol{\theta})}\Big)^\alpha d\boldsymbol{\theta}\Big] \tag{2.42}$$

$$\frac{\partial \mathsf{KL}(p(\mathbf{x},\boldsymbol{\theta})^\alpha [s_* q_*(\boldsymbol{\theta})]^{1-\alpha} \parallel sq(\boldsymbol{\theta}))}{\partial [sq(\boldsymbol{\theta})]} = 1 - \int \frac{p(\mathbf{x},\boldsymbol{\theta})^\alpha [s_* q_*(\boldsymbol{\theta})]^{1-\alpha}}{sq(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \ . \tag{2.43}$$

Consequently when $s_* q_*(\boldsymbol{\theta})$ is a stationary point of D_α (derivative $= 0$), it is also a stationary point of KL, as its derivative must also be zero. □

The proof given above is equivalent to, but marginally different from that given by Minka (2005), where derivatives with respect to the *parameters* of q were used. To find such a stationary point, we turn to a fixed point scheme. Define

$$f(s_* q_*(\boldsymbol{\theta})) = \arg\min_{sq(\boldsymbol{\theta})\in\mathcal{F}} \mathsf{KL}\Big(p(\mathbf{x},\boldsymbol{\theta})^\alpha [s_* q_*(\boldsymbol{\theta})]^{1-\alpha} \parallel sq(\boldsymbol{\theta})\Big) \tag{2.44}$$

so that we need to solve for

$$f(s_* q_*(\boldsymbol{\theta})) = s_* q_*(\boldsymbol{\theta}) \ . \tag{2.45}$$

This can be done with a typical fixed point algorithm to give

$$s_{(t+1)} q_{(t+1)}(\boldsymbol{\theta}) = f(s_{(t)} q_{(t)}(\boldsymbol{\theta})) \tag{2.46}$$

which we repeat until convergence. Minka (2005) suggests the addition of damping to the fixed point iterations, as the scheme is heuristic and not guaranteed to converge, but it is often successful with enough damping. We have the following fixed point scheme, which is illustrated in figure 2.4. We start with a guess of an initial $s_{(0)} q_{(0)}(\boldsymbol{\theta})$, where $q_{(0)}(\boldsymbol{\theta})$ is normalized, and typically take $s_{(0)} = 1$. Starting with $t = 0$, the following two steps are iterated until convergence.

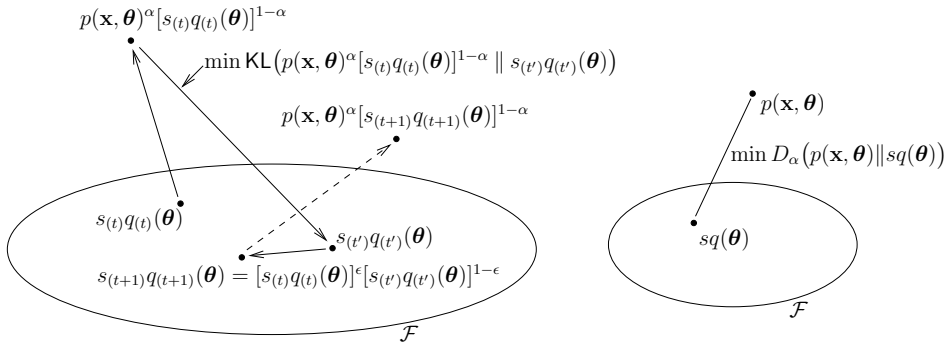


FIGURE 2.4: An α-divergence can be minimized by iteratively minimizing a KL divergence. Our aim, shown on the *right*, is to find an element $sq(\boldsymbol{\theta}) \in \mathcal{F}$ that is closest to $p(\mathbf{x},\boldsymbol{\theta})$ (which we assume is not in a chosen scaled exponential family $\mathcal{F}$) with respect to divergence D_α. This is solved through the iterative method on the *left*. For some $s_{(t)}q_{(t)}(\boldsymbol{\theta}) \in \mathcal{F}$, a function $p(\mathbf{x},\boldsymbol{\theta})^\alpha[s_{(t)}q_{(t)}(\boldsymbol{\theta})]^{1-\alpha}$ that is not in $\mathcal{F}$ is created, and now an element in $\mathcal{F}$ closest to it with respect to the KL divergence is found. After some possible damping, this process is repeated.

Step 1. We find $s_{(t')}q_{(t')}(\boldsymbol{\theta})$,

$$s_{(t')}q_{(t')}(\boldsymbol{\theta}) = \arg\min_{sq(\boldsymbol{\theta})} \mathsf{KL}\big(p(\mathbf{x},\boldsymbol{\theta})^\alpha[s_{(t)}q_{(t)}(\boldsymbol{\theta})]^{1-\alpha} \,\big\|\, sq(\boldsymbol{\theta})\big) \ , \tag{2.47}$$

by first computing the new scale

$$s_{(t')} = \int p(\mathbf{x},\boldsymbol{\theta})^\alpha[s_{(t)}q_{(t)}(\boldsymbol{\theta})]^{1-\alpha}d\boldsymbol{\theta} = s_{(t)}^{1-\alpha}\int p(\mathbf{x},\boldsymbol{\theta})^\alpha q_{(t)}(\boldsymbol{\theta})^{1-\alpha}d\boldsymbol{\theta} \ . \tag{2.48}$$

Now use $p(\mathbf{x},\boldsymbol{\theta})^\alpha q_{(t)}(\boldsymbol{\theta})^{1-\alpha}$ as the 'joint' distribution, and find a normalized $q_{(t')}(\boldsymbol{\theta})$ that minimizes $\mathsf{KL}\big(p(\mathbf{x},\boldsymbol{\theta})^\alpha q_{(t)}(\boldsymbol{\theta})^{1-\alpha} \,\big\|\, q_{(t')}(\boldsymbol{\theta})\big)$ by matching moments, as we have done in section 2.4. (The scale is excluded for simplicity, as we have already found it.)

Step 2. We now have $s_{(t')}q_{(t')}(\boldsymbol{\theta})$. If we had $\alpha = 1$, this would have been perfectly sufficient. We *damp* it with

$$s_{(t+1)}q_{(t+1)}(\boldsymbol{\theta}) = [s_{(t)}q_{(t)}(\boldsymbol{\theta})]^\epsilon[s_{(t')}q_{(t')}(\boldsymbol{\theta})]^{1-\epsilon} \ . \tag{2.49}$$

After implementing the above damping equation, we rearrange the product to keep $q_{(t+1)}$ as a normalized distribution. Therefore set

$$q_{(t+1)}(\boldsymbol{\theta}) = Z^{-1}q_{(t)}(\boldsymbol{\theta})^\epsilon q_{(t')}(\boldsymbol{\theta})^{1-\epsilon} \tag{2.50}$$

$$\text{with} \qquad Z = \int q_{(t)}(\boldsymbol{\theta})^\epsilon q_{(t')}(\boldsymbol{\theta})^{1-\epsilon}d\boldsymbol{\theta} \ , \tag{2.51}$$

$$\text{and then set} \qquad s_{(t+1)} = s_{(t)}^\epsilon s_{(t')}^{1-\epsilon}Z \ . \tag{2.52}$$

This heuristic scheme is iterated until convergence, and the final $s_{(t+1)}$ and $q_{(t+1)}$ are taken as minimizers of α-divergence. Convergence of the heuristic scheme depends on the amount of damping ϵ, and α.

2.6.1 Fixed point iterations

The fixed point scheme involves the 'prior' $p(\boldsymbol{\mu})^\alpha q_{(t)}(\boldsymbol{\mu})^{1-\alpha}$, for which we define the following shorthand parameters:

$$\hat{v}_i = \alpha v_{0i} + (1-\alpha)v_{i(t)} \tag{2.53}$$

$$\hat{m}_i = \frac{\alpha v_{0i} m_{0i} + (1-\alpha) v_{i(t)} m_{i(t)}}{\alpha v_{0i} + (1-\alpha) v_{i(t)}} \ . \tag{2.54}$$

To run the fixed point scheme, start with an initial $s_{(0)} q_{(0)}(\mathbf{z}_n) q_{(0)}(\boldsymbol{\mu})$, where $s_{(0)}$ is set to one, and the parameters γ_{nj} of $q_{(0)}(\mathbf{z}_n)$ possibly set to $1/J$. $q_{(0)}(\boldsymbol{\mu})$ can be set to the prior. Starting with $t = 0$, the following steps are repeated until convergence, or until some maximum number of iterations is reached.

Step 1. Determine the scale, which follows from appendix A.4.1 as

$$\begin{aligned} s_{(t')} &= \int \sum_{\mathbf{z}_n} p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)^\alpha [s_{(t)} q_{(t)}(\mathbf{z}_n) q_{(t)}(\boldsymbol{\mu})]^{1-\alpha} d\boldsymbol{\mu} \\ &= s_{(t)}^{1-\alpha} \prod_{i=1}^{J} \frac{\mathcal{Z}_{\mathcal{N}}(\hat{v}_i)}{\mathcal{Z}_{\mathcal{N}}(v_{0i})^\alpha \mathcal{Z}_{\mathcal{N}}(v_{i(t)})^{1-\alpha}} \exp\Big\{ -\frac{1}{2} \frac{\alpha v_{0i}(1-\alpha) v_{i(t)}}{\hat{v}_i} \big[m_{0i} - m_{i(t)}\big]^2 \Big\} \\ &\quad \times \alpha^{-1/2} \sum_{k=1}^{J} \pi_k^\alpha \gamma_{nk(t)}^{1-\alpha} \mathcal{Z}_{\mathcal{N}}(\lambda_k)^{1-\alpha} \mathcal{N}\Big(x_n \,\Big|\, \hat{m}_k, \frac{1}{\alpha\lambda_k} + \frac{1}{\hat{v}_k}\Big) \ . \end{aligned} \tag{2.55}$$

Following the scale, we find a normalized distribution $q_{(t')}(\mathbf{z}_n) q_{(t')}(\boldsymbol{\mu})$ that will minimize the KL divergence to $p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)^\alpha [q_{(t)}(\mathbf{z}_n) q_{(t)}(\boldsymbol{\mu})]^{1-\alpha}$. We have already solved for the scale and only need to match moments, for which the following expectations are used:

$$\langle \mu_j \rangle = \hat{m}_j \tag{2.56}$$

$$\langle \mu_j | x_n \rangle = \frac{\hat{v}_j \hat{m}_j + \alpha \lambda_j x_n}{\hat{v}_j + \alpha \lambda_j} \tag{2.57}$$

$$\langle \mu_j^2 \rangle = \frac{1}{\hat{v}_j} + \langle \mu_j \rangle^2 \tag{2.58}$$

$$\langle \mu_j^2 | x_n \rangle = \frac{1}{\hat{v}_j + \alpha \lambda_j} + \langle \mu_j | x_n \rangle^2 \ . \tag{2.59}$$

Define

$$r_{nj} = \frac{\pi_j^\alpha \gamma_{nj(t)}^{1-\alpha} \mathcal{N}(x_n \mid \hat{m}_j, \, (\alpha\lambda_j)^{-1} + \hat{v}_j^{-1})}{\sum_k \pi_k^\alpha \gamma_{nk(t)}^{1-\alpha} \mathcal{N}(x_n \mid \hat{m}_k, \, (\alpha\lambda_k)^{-1} + \hat{v}_k^{-1})} \ , \tag{2.60}$$

so that the updates again involve matching weighed moments,

$$m_{j(t')} = (1 - r_{nj}) \langle \mu_j \rangle + r_{nj} \langle \mu_j | x_n \rangle \tag{2.61}$$

$$\frac{1}{v_{j(t')}} = (1 - r_{nj}) \langle \mu_j^2 \rangle + r_{nj} \langle \mu_j^2 | x_n \rangle - m_{j(t')}^2 \tag{2.62}$$

$$\gamma_{nj(t')} = r_{nj} \ . \tag{2.63}$$

The only difference from the standard EP update is that an *exponentiated likelihood* was used; the mean and precision follows the same type of derivation as before. The derivation for $\gamma_{nj(t')}$ follows in appendix A.5.

Step 2. We have $s_{(t')} q_{(t')}(\mathbf{z}_n) q_{(t')}(\boldsymbol{\mu})$, and for the damping step need to do a derivation similar to the first step to find a normalized $q_{(t+1)}$. Define the mean and precision of the new (unscaled) Gaussian $q_{(t)}(\boldsymbol{\mu})^\epsilon q_{(t')}(\boldsymbol{\mu})^{1-\epsilon}$ as

$$v_{j(t+1)} = \epsilon v_{j(t)} + (1-\epsilon) v_{j(t')} \tag{2.64}$$

$$m_{j(t+1)} = \frac{\epsilon v_{j(t)} m_{j(t)} + (1-\epsilon) v_{j(t')} m_{j(t')}}{\epsilon v_{j(t)} + (1-\epsilon) v_{j(t')}} \tag{2.65}$$

for each component j. Then damping gives

$$q_{(t)}(\boldsymbol{\mu})^{\epsilon} q_{(t')}(\boldsymbol{\mu})^{1-\epsilon} = \prod_{j=1}^{J} \frac{1}{\mathcal{Z}_{\mathcal{N}}(v_{j(t)})^{\epsilon}} \frac{1}{\mathcal{Z}_{\mathcal{N}}(v_{j(t')})^{1-\epsilon}} \exp\Big\{ -\frac{1}{2} \frac{\epsilon v_{j(t)}(1-\epsilon) v_{j(t')}}{v_{j(t+1)}} (m_{j(t)} - m_{j(t')})^2 \Big\}$$
$$\times \exp\Big\{ -\frac{1}{2} v_{j(t+1)} (\mu_j - m_{j(t+1)})^2 \Big\} \tag{2.66}$$

$$q_{(t)}(\mathbf{z}_n)^{\epsilon} q_{(t')}(\mathbf{z}_n)^{1-\epsilon} = \prod_{j=1}^{J} [\gamma_{nj(t)}^{\epsilon} \gamma_{nj(t')}^{1-\epsilon}]^{z_{nj}} \ . \tag{2.67}$$

We would like to keep $q_{(t+1)}$ as a normalized distribution, and hence set

$$q_{(t+1)}(\boldsymbol{\mu}) = Z_1^{-1} q_{(t)}(\boldsymbol{\mu})^{\epsilon} q_{(t')}(\boldsymbol{\mu})^{1-\epsilon} \ . \tag{2.68}$$

The means and precisions will not change, but we have to keep track of the scale that we have to divide with to ensure that $q_{(t+1)}(\boldsymbol{\mu})$ remains normalized. That scale is

$$Z_1 = \prod_{j=1}^{J} \frac{\mathcal{Z}_{\mathcal{N}}(v_{j(t+1)})}{\mathcal{Z}_{\mathcal{N}}(v_{j(t)})^{\epsilon} \mathcal{Z}_{\mathcal{N}}(v_{j(t')})^{1-\epsilon}} \exp\Big\{ -\frac{1}{2} \frac{\epsilon v_{j(t)}(1-\epsilon) v_{j(t')}}{v_{j(t+1)}} (m_{j(t)} - m_{j(t')})^2 \Big\} \ . \tag{2.69}$$

Similarly, set the parameters of $q_{(t+1)}(\mathbf{z}_n)$ as

$$\gamma_{nj(t+1)} = Z_2^{-1} \gamma_{nj(t)}^{\epsilon} \gamma_{nj(t')}^{1-\epsilon} \ , \tag{2.70}$$

where the multinomial was normalized with

$$Z_2 = \sum_{k=1}^{J} \gamma_{nk(t)}^{\epsilon} \gamma_{nk(t')}^{1-\epsilon} \ . \tag{2.71}$$

Finally adjust the scale, remembering that we had to *divide* by Z_1 and Z_2 to keep q normalized. Add the appropriate log normalizers to get the updated log marginal likelihood estimate,

$$\ln s_{(t+1)} = \epsilon \ln s_{(t)} + (1-\epsilon) \ln s_{(t')} + \ln Z_1 + \ln Z_2 \ . \tag{2.72}$$

2.7 Minimizing over a factor graph

So far we have a general method of approximating a joint distribution, consisting of a prior and one observation, with a simpler distribution. On a factor graph, as in figure 2.3, this corresponds to approximating a 'prior' and 'data' factor. The approximation scheme can be extended over a full factor graph or full joint distribution. In this section we want to find a scaled distribution $sq(\boldsymbol{\theta})$ to match a full joint,

$$sq(\boldsymbol{\theta}) \approx p(\mathbf{x}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = \prod_{n=0}^{N} f_n(\boldsymbol{\theta}) \ , \tag{2.73}$$

and we hope to again find this by minimizing D_α. We have written the joint distribution as a product of factors, and the way that we choose the factorization need not be unique. In this

chapter a single data likelihood is chosen as a factor. In the full mixture of Gaussians that follows in chapter 3, we could equally have split the data likelihood into a mixture weight factor and a factor modeling the component parameters.

Let $sq(\boldsymbol{\theta})$ be a member of a scaled exponential family $\mathcal{F}$. Our choice is motivated as only a finite number of moments need to be propagated through the factor graph. The family is closed under multiplication, as the product of any number of distributions in $\mathcal{F}$ is also in $\mathcal{F}$. (When speaking about $sq(\boldsymbol{\theta})$, we have previously restricted $\mathcal{F}$ to normalizable exponential functions, for example in the iterative optimization scheme of section 2.6. Here we extend the exponentials to include factor approximations, which aren't necessarily normalizable, as well.) Each factor will be approximated by a member of $\mathcal{F}$,

$$\tilde{f}_n(\boldsymbol{\theta}) = \tilde{s}_n \exp\Big\{\sum_m \phi_m(\boldsymbol{\theta})\eta_{nm}\Big\} = \tilde{s}_n \exp\Big\{\boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\boldsymbol{\theta})\Big\} = \tilde{s}_n \tilde{f}'_n(\boldsymbol{\theta}) \ , \tag{2.74}$$

where $\boldsymbol{\eta}$ is some natural parameter vector, and $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of sufficient statistics or features of the distribution, e.g. $\boldsymbol{\phi}(\theta) = (\theta^2, \theta, 1)$ for a one dimensional Gaussian, and $\boldsymbol{\eta} = (-\frac{1}{2}v, vm, c)$. The factor definition has been made deliberately loosely; '1' has been added as a feature, to allow a choice for extra constant terms c in the exponential. This addition is possible as the exponential is again rescaled by some constant $\tilde{s}_n$, which we are free to choose. In the continuation of the mixtures example, we choose $c = -\frac{1}{2}vm^2$, for example, so that $\boldsymbol{\eta}^\top\boldsymbol{\phi}(\theta)$ factorizes over θ with the usual $-\frac{v}{2}(\theta - m)^2$.

The factor approximations need *not* be normalizable; however, their product must be, and equal to $sq(\boldsymbol{\theta}) \in \mathcal{F}$,

$$sq(\boldsymbol{\theta}) = \prod_{n=0}^{N} \tilde{f}_n(\boldsymbol{\theta}) \ . \tag{2.75}$$

The following section describes how the above approximation can be found by considering single factors at a time.

2.7.1 A generic message passing algorithm

Assume that we have an approximation for all factors, except for a factor n, and that we want to include f_n in the approximation. Now define two distributions, the joint without factor f_n's inclusion,

$$p^{\setminus n}(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i \neq n} f_i(\boldsymbol{\theta}) \ , \tag{2.76}$$

and the approximation excluding approximate factor $\tilde{f}_n$,

$$s^{\setminus n} q^{\setminus n}(\boldsymbol{\theta}) = sq(\boldsymbol{\theta})/\tilde{f}_n(\boldsymbol{\theta}) = \prod_{i \neq n} \tilde{f}_i(\boldsymbol{\theta}) \ . \tag{2.77}$$

In ideal circumstances factor $\tilde{f}_n$ should minimize

$$D_\alpha\big(p^{\setminus n}(\mathbf{x}, \boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \,\big\|\, s^{\setminus n} q^{\setminus n}(\boldsymbol{\theta}) \tilde{f}_n(\boldsymbol{\theta})\big) \ , \tag{2.78}$$

which gives rise to an intractable minimization problem. A tractable road forward exists, and that is to assume that the approximation made to the rest of the factor graph is a good one, so that

$$p^{\setminus n}(\mathbf{x}, \boldsymbol{\theta}) \approx s^{\setminus n} q^{\setminus n}(\boldsymbol{\theta}) \ . \tag{2.79}$$

Algorithm 1 Generic message passing algorithm

1: **initialize:** $\tilde{f}_n(\boldsymbol{\theta})$ for all n.
2: **repeat**
3: pick a factor n.
4: compute $s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})$.
5: update the factor approximation $\tilde{f}_n(\boldsymbol{\theta})$ by getting the new approximation to the joint $s^{\text{new}}q(\boldsymbol{\theta})^{\text{new}}$ with equation (2.81), and solve for the updated term contribution $\tilde{f}_n$ with (2.83).
6: **until** all $\tilde{f}_n(\boldsymbol{\theta})$ converge.

It may be helpful to interpret this assumption in the following way (specific to the examples in this chapter, although the method is much more general): If single data likelihoods are taken as factors, this simply means that all the data that we have already included in the approximation can be summarized in some form of scaled prior distribution $s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})$, which needs to be multiplied by the data term factor f_n (or likelihood $p(\mathbf{x}_n|\boldsymbol{\theta})$), to get a new approximation to the joint distribution.

The problem simplifies to that of finding a $\tilde{f}_n$ that minimizes

$$D_\alpha\big(s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta}) \,\big\|\, s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})\tilde{f}_n(\boldsymbol{\theta})\big) \ . \tag{2.80}$$

In light of the methods presented in sections 2.4 to 2.6, where some $sq(\boldsymbol{\theta})$ is matched to a prior times a likelihood, we find $\tilde{f}_n$ by finding a *new* scaled approximation $s^{\text{new}}q(\boldsymbol{\theta})^{\text{new}}$, with

$$s^{\text{new}}q(\boldsymbol{\theta})^{\text{new}} = \underset{sq(\boldsymbol{\theta})}{\arg\min}\, D_\alpha\big(s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) \ , \tag{2.81}$$

and then using $s^{\text{new}}q(\boldsymbol{\theta})^{\text{new}}$ and $s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})$ to find $\tilde{f}_n(\boldsymbol{\theta})$. This is purely a matter of division, as we are using the exponential family of distributions. We have

$$s^{\backslash n}q^{\backslash n}(\boldsymbol{\theta})\tilde{f}_n(\boldsymbol{\theta}) = s^{\text{new}}q(\boldsymbol{\theta})^{\text{new}} \ , \tag{2.82}$$

and hence

$$\tilde{f}_n(\boldsymbol{\theta}) = \frac{s^{\text{new}}}{s^{\backslash n}}\frac{q(\boldsymbol{\theta})^{\text{new}}}{q^{\backslash n}(\boldsymbol{\theta})} \ . \tag{2.83}$$

The resulting message passing algorithm is summarized in algorithm 1.

2.7.2 Minimizing the mixtures example over a factor graph

For our *practical* implementation it is not necessary to determine $s^{\backslash n}$ when we want to update the scale contribution $\tilde{s}_n$, as

$$Sq(\boldsymbol{\theta})^{\text{new}} = \underset{s'q(\boldsymbol{\theta})}{\arg\min}\, D_\alpha\big(q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta}) \,\big\|\, s'q(\boldsymbol{\theta})\big) \tag{2.84}$$

gives the same minimum as equation (2.81), with $s^{\text{new}} = Ss^{\backslash n}$, and equation (2.83) therefore gives $\tilde{f}_n(\boldsymbol{\theta}) = Sq(\boldsymbol{\theta})^{\text{new}}/q^{\backslash n}(\boldsymbol{\theta})$. This slightly simplified restatement will be used in the minimization routine that follows.

We choose term definitions $\tilde{f}_n$, for Gaussian components with unknown means, as

$$\tilde{f}_n = \tilde{s}_n \prod_{j=1}^{J} e^{-\frac{1}{2}\tilde{v}_{nj}(\mu_j - \tilde{m}_{nj})^2} \ . \tag{2.85}$$

The message passing algorithm is:

- Start by initializing, for $n = 1, \ldots, N$,

$$\tilde{s}_n = 1, \quad \tilde{m}_{nj} = 0, \quad \text{and} \quad \tilde{v}_{nj} = 0 \ , \tag{2.86}$$

so that all the factor approximations are one. Initialize the prior as

$$\tilde{s}_0 = \prod_{j=1}^{J} \frac{1}{\mathcal{Z}_{\mathcal{N}}(\tilde{v}_{0j})} = \prod_{j=1}^{J} (\tilde{v}_{0j}/2\pi)^{1/2}, \quad \tilde{m}_{0j} = m_{0j}, \quad \text{and} \quad \tilde{v}_{0j} = v_{0j} \ . \tag{2.87}$$

- Repeat until all $\tilde{f}_n$ converge:

1. Pick a factor n. This can be done by looping over different random permutations of $1, \ldots, N$.
2. Compute $q^{\setminus n}(\boldsymbol{\mu})$ (for reasons stated above we need not concern ourselves with $s^{\setminus n}$). We take the convention of identifying the parameters of $q^{\setminus n}(\boldsymbol{\mu})$ with a subscript o for 'old', where 'old' refers to our approximation *before* including term n. This convention is also followed because here $q^{\setminus n}(\boldsymbol{\mu})$ takes the role of a prior, allowing a straight-forward implementation of the optimization routines of sections 2.4 to 2.6. Recovering the old distribution—also referred to as the cavity distribution—follows from reversing equations (2.91) and (2.92),

$$v_{\mathrm{o}j} = v_j - \tilde{v}_{nj} \tag{2.88}$$

$$m_{\mathrm{o}j} = \frac{v_j m_j - \tilde{v}_{nj}\tilde{m}_{nj}}{v_{\mathrm{o}j}} \ . \tag{2.89}$$

If a non-normalizable distribution is recovered, the update for factor $\tilde{f}_n$ can be skipped, and we continue again from step 1.

3. Similar to sections 2.4 to 2.6, let S, m_j and v_j be the parameters of $q(\boldsymbol{\mu})^{\mathrm{new}}$ that minimizes equation (2.84). After finding the new approximation, we have to set the factor contribution. After a rearrangement so that a single exponential depends on μ_j, we get

$$\begin{aligned}
\tilde{f}_n &= S \frac{q(\boldsymbol{\mu})^{\mathrm{new}}}{q^{\setminus n}(\boldsymbol{\mu})} \\
&= S \prod_{j=1}^{J} \frac{\mathcal{Z}_{\mathcal{N}}(v_{\mathrm{o}j})}{\mathcal{Z}_{\mathcal{N}}(v_j)} \exp\Big\{ + \frac{1}{2} \frac{v_j v_{\mathrm{o}j}}{v_j - v_{\mathrm{o}j}} (m_j - m_{\mathrm{o}j})^2 \Big\} \\
&\qquad \cdots \prod_{j=1}^{J} \exp\Big\{ - \frac{1}{2}(v_j - v_{\mathrm{o}j}) \big[\mu_j - \frac{v_j m_j - v_{\mathrm{o}j} m_{\mathrm{o}j}}{v_j - v_{\mathrm{o}j}}\big]^2 \Big\} \ .
\end{aligned} \tag{2.90}$$

According to our term definition (2.85), we get the substitution for $\tilde{v}_{nj}$ and $\tilde{m}_{nj}$ for $j = 1, \ldots, J$, and for the change of scale $\tilde{s}_n$, as

$$\tilde{v}_{nj} = v_j - v_{\mathrm{o}j} \tag{2.91}$$

$$\tilde{m}_{nj} = \frac{v_j m_j - v_{\mathrm{o}j} m_{\mathrm{o}j}}{\tilde{v}_{nj}} \tag{2.92}$$

$$\tilde{s}_n = S \prod_{j=1}^{J} \frac{\mathcal{Z}_{\mathcal{N}}(v_{\mathrm{o}j})}{\mathcal{Z}_{\mathcal{N}}(v_j)} \exp\Big\{ + \frac{1}{2} \frac{v_j v_{\mathrm{o}j}}{v_j - v_{\mathrm{o}j}} (m_j - m_{\mathrm{o}j})^2 \Big\} \ . \tag{2.93}$$

- Finally the approximation to the evidence $p(\mathbf{x})$ is determined with

$$p(\mathbf{x}) \approx \int \prod_{n=0}^{N} \tilde{f}_n(\boldsymbol{\mu})\, d\boldsymbol{\mu} = \Big(\prod_{n=0}^{N} \tilde{s}_n\Big) \prod_{j=1}^{J} e^{\frac{1}{2}[v_j m_j^2 - \sum_{n=0}^{N} \tilde{v}_{nj}\tilde{m}_{nj}^2]} \mathcal{Z}_{\mathcal{N}}(v_j) \ . \tag{2.94}$$

Starting with another prior contribution

We need not start by initializing the prior contribution $\tilde{s}_0$, $\tilde{m}_{0j}$ and $\tilde{v}_{0j}$ $\forall j$ to the prior hyperparameter values; sometimes we may wish to break symmetry so that the iterative algorithm converges to one solution rather than another. This may be necessary when the prior is symmetric around zero, causing the responsibilities to always remain equal, stifling any progress. A typical example may be when the posterior has more than one equivalent mode (as often happens in mixtures models, which are invariant with respect to permutations of the component labeling). When the modes are well separated we may want to approximate one of them, i.e. break symmetry and not find a solution with small mass balanced between them.

To break symmetry we start with the 'wrong' prior, and can always go back after the first loop over factors to correct the prior contributions to the correct prior hyperparameter values. (In practice, if $m_{0j} = 0$ $\forall j$, for example, we may start with $\tilde{m}_{0j} = \epsilon_j$ where ϵ_j represents some added symmetry-breaking noise.)

If we did choose to start with the wrong prior contributions, the prior is treated like any other factor. The hyperparameter values need not to be set in any optimization routine, as they are prespecified and fixed. We find $q(\boldsymbol{\mu})^{\text{new}}$ with

$$v_j = v_{\text{o}j} + v_{0j} \quad \text{and} \quad m_j = (v_{\text{o}j} m_{\text{o}j} + v_{0j} m_{0j})/(v_{\text{o}j} + v_{0j}) \ , \tag{2.95}$$

and update the prior contribution to the correct hyperparameters,

$$\tilde{s}_0 = \prod_{j=1}^{J} \frac{1}{\mathcal{Z}_{\mathcal{N}}(\tilde{v}_{0j})}, \quad \tilde{m}_{0j} = m_{0j}, \quad \text{and} \quad \tilde{v}_{0j} = v_{0j} \ . \tag{2.96}$$

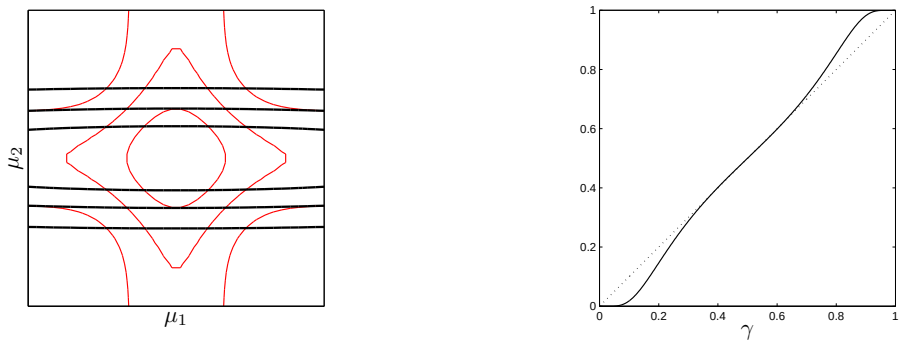
2.8 Model pruning

The aim of this discussion is to look at the behavior of VB and EP on very simple examples. This will mostly be in light of chapter 3, for when we understand the behavior of the algorithms under different settings, we can intuitively explain certain higher-dimensional phenomena.

A problem that may arise when a free energy is minimized in *variational Bayes* ($\alpha = 0$), is that the degrees of freedom in the parameter space may be pruned, perhaps even inappropriately (MacKay, 2001). A mixture model, as is the example in this chapter, may self-prune. If we add more components to the model we believe in, the extra components and parameters may not be used at all. This is contrary to our expectation, where we expect all the parameters to be included in the posterior, maybe with big error bars on them.

Let's first reflect back on a well-behaved example. For figure 2.1(a) we had $J = 2$ components, and found an approximation $sq(\boldsymbol{\mu})$ that gave equal weight to each mixture component. In the approximation $q(\mu_1)$ and $q(\mu_2)$ are equal, and $\gamma_{n1} = \gamma_{n2} = \frac{1}{2}$ (with one observation, $N = 1$). This is what we would expect, as we have no motivation to prefer one component over the other. (All precisions were set to $\lambda_j = 1$, the mixing weights were fixed at $\pi_j = \frac{1}{2}$, and $v_{0j} = 0.1$ and $m_{0j} = 0$ were taken as prior parameter values.)

A typical example of unwanted pruning is found in figure 2.5(a), which is similar to an example encountered before in figure 2.1(a), but with a *broader* prior value $v_{0j} = 0.01$. Apart from



(a) A possible approximate solution for the fixed point equation (2.97), showing one component being pruned.

(b) Possible solutions to equation (2.97), at the intersections of the functions γ and $g(\gamma) = 1/(1+\exp\{\frac{1}{2}(\frac{1}{v_1}-\frac{1}{v_2})\})$. There are in fact five intersections, at $\gamma \approx 0, 0.38, 0.5, 0.62$ and 1.

FIGURE 2.5: Local solutions with a broader prior $v_{0j} = 0.01$. With $J = 2$ components, the other prior parameter values we set to $\pi_j = \frac{1}{2}$, $m_{0j} = 0$ and $\lambda_j = 1$.

$\gamma_{n1} = 0.5$, there are four other local maxima in the free energy, $-\ln s$, with $\gamma_{n1} \approx 0, 0.38, 0.62$ and 1. With $\gamma_{n1} \approx 1$, and its converse, one of the mixture components is lost, with its approximation effectively being equal to the prior. (The values of γ_{nj} are not strictly at zero and one, but asymptotically close as v_{0j} shrinks to give a broader prior.) There would be no reason for us to favor any model, as the single $x_n = 0$ could have been generated by the first or second component, but yet that is what the local minima in the free energy would lead us to believe.

As we have only one data point and two components, let $\gamma \equiv \gamma_{n1}$. From equations (2.30) and (2.33) we have

$$\gamma = \frac{1}{1+\exp\left\{\frac{1}{2}(\frac{1}{v_1}-\frac{1}{v_2})\right\}}, \quad \text{where} \quad v_1 = v_{01} + \gamma\lambda_1 \quad \text{and} \quad v_2 = v_{02} + (1-\gamma)\lambda_2 \ , \tag{2.97}$$

with the values of λ_j fixed at 1. To solve for the extrema, we need to find a value of γ such that equation (2.97) holds. With function $g(\gamma) = 1/(1+\exp\{\frac{1}{2}(\frac{1}{v_1}-\frac{1}{v_2})\})$, this can be solved with a fixed point equation

$$\gamma^{\text{new}} \leftarrow g(\gamma) \ , \tag{2.98}$$

which is merely the VBEM algorithm in another form. Figure 2.5(b) shows a plot of γ and $g(\gamma)$ with a broader prior $v_{0j} = 0.01$. The free energy minima occur at the intersection of these two functions.

If we narrow the prior and make v_{0j} bigger, the symmetry-breaking solutions are lost. Figure 2.6 illustrates this phenomena. There is a critical value of $v_{0j} \approx 0.97$ that separates our notion of symmetry-breaking and symmetry preserving priors in this example, where bigger precisions (smaller variances) preserve the symmetry. This gives some *bifurcation*, as there is a very sudden split between many possible solutions and only one.

The relevance of this discussion finds form in section 3.7, where VB is run over a factor graph. As factors (or data points) are included one at a time, this differs from conventional VB, where all data points are treated together in an EM algorithm. MacKay (2001) illustrates that unwanted model pruning becomes less of a problem with more data present in standard VB. By including one data point at a time, or optimizing on a factor by factor basis, it has been found in practice (section 3.7) that components can be lost on the inclusion of the first data

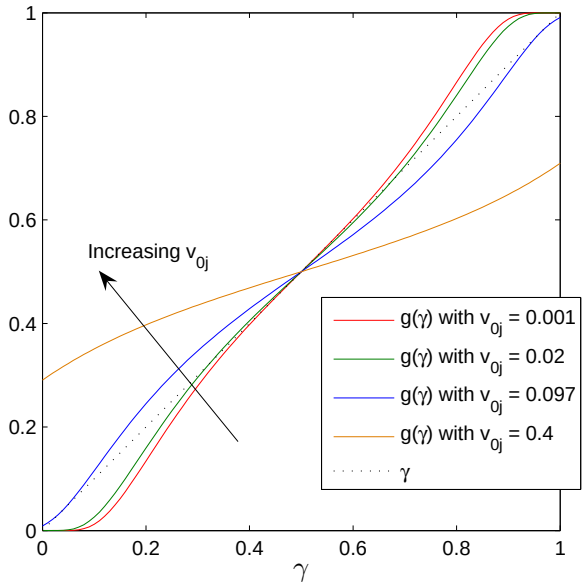


FIGURE 2.6: If the prior from figure 2.5 is narrowed, the symmetry-breaking solutions are lost, so that the only solution to the fixed point equation (2.97) would be a spherical Gaussian centered at zero. The figure also illustrates a critical value $v_{0j} \approx 0.97$ where a bifurcation occurs, as we see a very sudden split between many possible solutions (intersections) and only one.

point to a broad prior. This puts us in a situation where some components are almost zero-weighted after the addition of the first factor, and the inclusion of subsequent factors cannot recover these components. As we will see next, EP doesn't suffer from this drawback, and the solution adopted in section 3.7 is to run one or two EP loops over all factors to give a sensible initialization of approximate factors before the VB loops are run.

Model pruning and expectation propagation

An approximation scheme like EP will not suffer under the same component-pruning behaviour, as the parameter updates take a on a very different form.

- For VB we solved using the EM fixed point equations, and iterated an expectation and maximization step. The E-step determined responsibilities that depend on the present parameter settings and *not* the prior. The M-step then depended on the responsibilities and prior. Equations (2.30), (2.33) and (2.34) give an example of these dependencies.
- EP relied on two steps as well, but they are not iterated. The first was a responsibility update that depends on the prior and *not* any present parameter values. It was followed by a moment matching step that depends on the responsibility-weighed prior and component-posterior moments. Equations (2.17), (2.18) and (2.20) can be taken as an example.

2.9 The objective function

Under a set of constrains, the generic message passing algorithm would attempt to find the largest scale s for a given normalized distribution $q(\boldsymbol{\theta})$ and an associated joint distribution

$p(\mathbf{x}, \boldsymbol{\theta})$. In other words, given a factorized approximation, we can determine the scale, and therefore need to maximize some objective function over the *approximate factors* such that the scale is as big as possible. This was clearly seen in the case of variational Bayes, for example, where s was always a lower bound to the true marginal likelihood $p(\mathbf{x})$, and we had to adjust the approximating distribution so that the scale (as a function of the approximating distribution) could be as big as possible.

We will show here, following (Minka, 2005), that the generic message passing iterations from algorithm 1 always have a fixed point when the approximations are in the exponential family. This does not mean that the fixed point will necessarily be found, even though its objective function is well defined. It is a single loop algorithm, and longer double-loop algorithms, which come with a convergence guarantee, can be created to minimize the same objective function Opper & Winther (2005a).

The scale that will be computed by message-passing can be written as

$$s = \int sq(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int \prod_{n=0}^{N} \tilde{f}_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \Big(\prod_{n=0}^{N} \tilde{s}_n\Big) \int \prod_{n=0}^{N} \tilde{f}'_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \Big(\prod_{n=0}^{N} \tilde{s}_n\Big) \int q'(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{2.99}$$

where we have defined the shortcut $q'(\boldsymbol{\theta}) = \prod_n \tilde{f}'_n(\boldsymbol{\theta})$. The final approximation has the form

$$q(\boldsymbol{\theta}) \propto q'(\boldsymbol{\theta}) = \prod_{n=0}^{N} \tilde{f}'_n(\boldsymbol{\theta}) = \prod_{n=0}^{N} \exp\Big\{\boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\boldsymbol{\theta})\Big\} = \exp\Big\{\sum_{n=0}^{N} \boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\boldsymbol{\theta})\Big\} \tag{2.100}$$

for some $\boldsymbol{\eta} \equiv \{\boldsymbol{\eta}_n\}_{n=0}^{N}$. Each leave-one-out (cavity) distribution has the form

$$q^{\backslash n}(\boldsymbol{\theta}) \propto \prod_{i \neq n} \tilde{f}'_i(\boldsymbol{\theta}) = \exp\Big\{\sum_{i \neq n} \boldsymbol{\eta}_i^\top \boldsymbol{\phi}(\boldsymbol{\theta})\Big\}\ . \tag{2.101}$$

The following two sections treat $\alpha \neq 0$ and $\alpha = 0$ as two separate cases, and the objective function is derived for each.

2.9.1 The objective function for $\alpha \neq 0$

At each step in algorithm 1's main loop, the α-divergence

$$\underset{\tilde{f}_n(\boldsymbol{\theta})}{\arg\min}\, D_\alpha\Big(s^{\backslash n} q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \,\Big\|\, s^{\backslash n} q^{\backslash n}(\boldsymbol{\theta}) \tilde{f}_n(\boldsymbol{\theta})\Big) \tag{2.102}$$

is minimized. To determine the scale from (2.99), we are interested in the different scales $\tilde{s}_n$ that will be computed. The approximate factors divide into $\tilde{f}_n(\boldsymbol{\theta}) = \tilde{s}_n \tilde{f}'_n(\boldsymbol{\theta})$, and we can write the leave-one-out estimates or cavity distributions as

$$s^{\backslash n} q^{\backslash n}(\boldsymbol{\theta}) = \prod_{i \neq n} \tilde{s}_i \tilde{f}'_i(\boldsymbol{\theta})\ . \tag{2.103}$$

To determine each individual scale $\tilde{s}_n$, take the derivative of the α-divergence with respect to $\tilde{s}_n$,

$$\frac{\partial D_\alpha}{\partial \tilde{s}_n} = \frac{\partial}{\partial \tilde{s}_n} \frac{1}{\alpha(1-\alpha)} \Bigg[\int \alpha f_n(\boldsymbol{\theta}) \prod_{i \neq n} \tilde{s}_i \tilde{f}'_i(\boldsymbol{\theta}) + (1-\alpha) \tilde{s}_n \tilde{f}'_n(\boldsymbol{\theta}) \prod_{i \neq n} \tilde{s}_i \tilde{f}'_i(\boldsymbol{\theta})$$

$$
\underbrace{- \Big(f_n(\boldsymbol{\theta}) \prod_{i \neq n} \tilde{s}_i \tilde{f}_i'(\boldsymbol{\theta}) \Big)^{\alpha} \Big(\tilde{s}_n \tilde{f}_n'(\boldsymbol{\theta}) \prod_{i \neq n} \tilde{s}_i \tilde{f}_i'(\boldsymbol{\theta}) \Big)^{1-\alpha} d\boldsymbol{\theta}}_{=\tilde{s}_n^{1-\alpha} \left(\prod_{i \neq n} \tilde{s}_i \right) q'(\boldsymbol{\theta}) \left(f_n(\boldsymbol{\theta}) / \tilde{f}_n'(\boldsymbol{\theta}) \right)^{\alpha}} \Bigg]
$$

$$
= \frac{1}{\alpha} \Big(\prod_{i \neq n} \tilde{s}_i \Big) \int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta} - s_n^{-\alpha} \frac{1}{\alpha} \Big(\prod_{i \neq n} \tilde{s}_i \Big) \int q'(\boldsymbol{\theta}) \left(\frac{f_n(\boldsymbol{\theta})}{\tilde{f}_n'(\boldsymbol{\theta})} \right)^{\alpha} d\boldsymbol{\theta} \ , \tag{2.104}
$$

and equating the partial derivative to zero gives, for $\alpha \neq 0$,

$$
\tilde{s}_n = \left(\frac{\int q'(\boldsymbol{\theta}) \left(\frac{f_n(\boldsymbol{\theta})}{\tilde{f}_n'(\boldsymbol{\theta})} \right)^{\alpha} d\boldsymbol{\theta}}{\int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \right)^{1/\alpha} . \tag{2.105}
$$

Let $N+1$ be the number of terms $f_n(\boldsymbol{\theta})$, as we conventionally count n from zero. If we now substitute $\tilde{s}_n$ back into the equation (2.99) we get a mass estimate

$$
s = \left(\int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \right)^{1-(N+1)/\alpha} \prod_{n=0}^{N} \left(\int q'(\boldsymbol{\theta}) \Big(\frac{f_n(\boldsymbol{\theta})}{\tilde{f}_n'(\boldsymbol{\theta})} \Big)^{\alpha} d\boldsymbol{\theta} \right)^{1/\alpha} . \tag{2.106}
$$

From the approximate and leave-one-out distributions of equations (2.100) and (2.101), the negative log scale gives an objective function or free energy to be minimized,

$$
\begin{aligned}
- \ln s = & \Big(\frac{N+1}{\alpha} - 1 \Big) \ln \int \exp \Big\{ \sum_n \boldsymbol{\eta}_n^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta} \\
& - \sum_{n=0}^{N} \frac{1}{\alpha} \ln \int f_n(\boldsymbol{\theta})^{\alpha} \exp \Big\{ \Big(\sum_{i \neq n} \boldsymbol{\eta}_i + (1-\alpha) \boldsymbol{\eta}_n \Big)^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta} \ .
\end{aligned} \tag{2.107}
$$

The generic message passing scheme from algorithm 1 comes with no guarantee that the minimum of the objective function in (2.107) will be found. It is a single-loop algorithm, which will converge fast in many practical cases. We now turn our discussion to double loop algorithms, which come with a guarantee of convergence.

Expectation consistent inference and double loop algorithms

Expectation consistent (EC) inference (Opper & Winther, 2005a,b) provides an alternative view of the local-consistency approximations made by EP, and generalizes Adaptive TAP (Opper & Winther, 2001a,b), which is used for inference in densely connected graphical models. This section presents a short review, emphasizing that other algorithms can be derived to minimize (2.107).

Taking two terms from (2.107), with $\alpha = 1$, we can write the objective function as the EC approximation to the free energy,

$$
\begin{aligned}
- \ln Z^{\mathrm{EC}} = - \ln s = & \ln \int \exp \Big\{ (\boldsymbol{\eta}_0 + \boldsymbol{\eta}_1)^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta} \\
& - \ln \int f_0(\boldsymbol{\theta}) \exp \Big\{ \boldsymbol{\eta}_1^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta} - \ln \int f_1(\boldsymbol{\theta}) \exp \Big\{ \boldsymbol{\eta}_0^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta} \ ,
\end{aligned} \tag{2.108}
$$

which we write as the sum of three partition functions:

$$
- \ln Z^{\mathrm{EC}}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_1) = \ln Z^q(\boldsymbol{\eta}_0 + \boldsymbol{\eta}_1) - \ln Z_0(\boldsymbol{\eta}_1) - \ln Z_1(\boldsymbol{\eta}_0) \ . \tag{2.109}
$$

This can naturally be extended to a higher number of terms. By writing three distributions,

$$q(\boldsymbol{\theta}) = \frac{1}{Z^q(\boldsymbol{\eta}_0 + \boldsymbol{\eta}_1)} \exp\left\{(\boldsymbol{\eta}_0 + \boldsymbol{\eta}_1)^\top \boldsymbol{\phi}(\boldsymbol{\theta})\right\} \tag{2.110}$$

$$q_0(\boldsymbol{\theta}) = \frac{1}{Z_0(\boldsymbol{\eta}_1)} f_0(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(\boldsymbol{\theta})\right\} \tag{2.111}$$

$$q_1(\boldsymbol{\theta}) = \frac{1}{Z_1(\boldsymbol{\eta}_0)} f_1(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\eta}_0^\top \boldsymbol{\phi}(\boldsymbol{\theta})\right\} , \tag{2.112}$$

the objective function in (2.109) is maximized by solving a nonlinear set of equations such that the moments match,

$$\left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_q = \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_0} = \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_1} . \tag{2.113}$$

In this case we can view EP as a particular algorithm, and compare it to a double loop algorithm, for example given below.

To examine the stationary points of (2.109), notice that the log partition functions, for example $\ln Z_0$, are the *cumulant generating function* of the random variables $\boldsymbol{\phi}(\boldsymbol{\theta})$,

$$\mathbf{H}_0 = \frac{\partial^2 \ln Z_0(\boldsymbol{\eta}_1)}{\partial \boldsymbol{\eta}_1 \partial \boldsymbol{\eta}_1^\top} = \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \boldsymbol{\phi}(\boldsymbol{\theta})^\top \right\rangle - \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle^\top , \tag{2.114}$$

where the expectations are taken under distribution $q_0(\boldsymbol{\theta})$. The log partition functions are differentiable and *convex* functions of their domains. We can similarly define Hessian matrices $\mathbf{H}^q$ and $\mathbf{H}_1$, all of which will be positive semi-definite. By considering $\boldsymbol{\eta} = (\boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$, we can conclude that the EC objective function is a *non-convex* combination of convex functions, and has a Hessian that is not necessarily positive semi-definite,

$$\mathbf{H}^{\mathrm{EC}} = \frac{\partial^2 \ln Z^{\mathrm{EC}}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top} = -\begin{pmatrix} \mathbf{H}^q - \mathbf{H}_1 & \mathbf{H}^q \\ \mathbf{H}^q & \mathbf{H}^q - \mathbf{H}_0 \end{pmatrix} . \tag{2.115}$$

There may therefore be more than one stationary point.

The double loop algorithm (Yuille, 2002; Opper & Winther, 2005a) is guaranteed to converge to a stationary point, assuming that a certain cost function is bounded from below. Define $\boldsymbol{\eta}_{\mathrm{sum}} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1$, so that (2.109) is restated as

$$-\ln Z^{\mathrm{EC}}(\boldsymbol{\eta}_{\mathrm{sum}}, \boldsymbol{\eta}_1) = \ln Z^q(\boldsymbol{\eta}_{\mathrm{sum}}) - \ln Z_0(\boldsymbol{\eta}_1) - \ln Z_1(\boldsymbol{\eta}_{\mathrm{sum}} - \boldsymbol{\eta}_1) . \tag{2.116}$$

The double loop algorithm iterates two steps:

Step 1. When $\boldsymbol{\eta}_{\mathrm{sum}}$ is held fixed, (2.116) is *concave* in $\boldsymbol{\eta}_1$, and can be uniquely maximized with

$$\boldsymbol{\eta}_{1(t)} = \arg\max_{\boldsymbol{\eta}_1} \left[-\ln Z^{\mathrm{EC}}(\boldsymbol{\eta}_{\mathrm{sum}(t-1)}, \boldsymbol{\eta}_1) \right] . \tag{2.117}$$

As a result the moments of $q_{0(t)}(\boldsymbol{\theta})$ and $q_{1(t)}(\boldsymbol{\theta})$ are set to be equal,

$$\left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_{0(t)}} = \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_{1(t)}} . \tag{2.118}$$

Step 2. When $\boldsymbol{\eta}_1$ is held fixed, (2.116) is a sum of a *concave* and a *convex* function of $\boldsymbol{\eta}_{\mathrm{sum}}$, and cannot be directly minimized. What we can do, however, is to update $\boldsymbol{\eta}_{\mathrm{sum}}$ given the moments at a fixed $\boldsymbol{\eta}_1$. This is essentially an EP step, saying

$$q_{(t)}(\boldsymbol{\theta}) = \arg\min_{q(\boldsymbol{\theta})} \mathsf{KL}(q_{0(t)}(\boldsymbol{\theta}) \parallel q(\boldsymbol{\theta})) , \tag{2.119}$$

which will update $\boldsymbol{\eta}_{\text{sum}(t)}$ such that the moments of $q_{(t)}(\boldsymbol{\theta})$ and $q_{0(t)}(\boldsymbol{\theta})$ match:

$$\left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_{(t)}} = \left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_{0(t)}} . \tag{2.120}$$

Although not necessary, the value of $\boldsymbol{\eta}_{0(t)}$ can be determined with $\boldsymbol{\eta}_{0(t)} = \boldsymbol{\eta}_{\text{sum}(t)} - \boldsymbol{\eta}_{1(t)}$.

Our aim here is merely to present an overview of the double loop algorithm and gain insight into the nature of the objective function, and not to provide a detailed description. A thorough account, with convergence proof, is provided by Opper & Winther (2005a).

Finally, we note that the first step differentiates the double loop algorithm from the single loop algorithm that is EP, which is summarized here for comparison:

Step 1 (EP). Update $\boldsymbol{\eta}_{\text{sum}}$ given the moments at a fixed $\boldsymbol{\eta}_0$. This is equivalent to minimizing

$$q_{(t)}(\boldsymbol{\theta}) = \arg\min_{q(\boldsymbol{\theta})} \mathsf{KL}(q_{1(t-1)}(\boldsymbol{\theta}) \parallel q(\boldsymbol{\theta})) , \tag{2.121}$$

which will update $\boldsymbol{\eta}_{\text{sum}(t)}$ such that the moments of $q_{(t)}(\boldsymbol{\theta})$ and $q_{1(t-1)}(\boldsymbol{\theta})$ match:

$$\left\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \right\rangle_{q_{(t)}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{1(t-1)}} . \tag{2.122}$$

On obtaining $\boldsymbol{\eta}_{\text{sum}(t)}$, the value of $\boldsymbol{\eta}_{1(t)}$—to be kept fixed in step 2—is determined with $\boldsymbol{\eta}_{1(t)} = \boldsymbol{\eta}_{\text{sum}(t)} - \boldsymbol{\eta}_{0(t-1)}$.

Step 2 (EP). The double loop algorithm's step 2, or step 1 (EP) with the roles of $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ exchanged.

As the objective function may have more than one stationary point, a small example follows for further insight.

One mode or both? A toy example

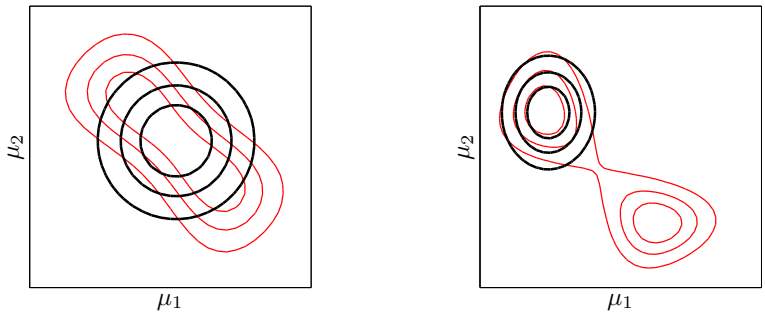
The purpose of this small example is to show that when two well-separated modes exist in the joint distribution, *expectation propagation* may lock to one of them, as the objective function has more than one local minimum. The mode need not be the biggest mode, as the results in section 3.7 will indicate.

For figure 2.7, a data set with twenty examples was created so that the joint distribution has two closely-connected modes. The objective is minimized by a unique solution that includes both modes. When some of the data points are slightly adjusted to separate the modes, the EP loops converge to one of the modes.

2.9.2 The VB objective function

From the same construction as section 2.9.1 we can determine the objective function that is minimized on local divergences with $\alpha = 0$. The message passing fixed points is in this case (uniquely for $\alpha = 0$) equal to the stationary points of the global KL divergence, as the objective functions can be shown to be identical. At each step the KL divergence

$$\mathsf{KL}\Big(s^{\backslash n} q^{\backslash n}(\boldsymbol{\theta}) \tilde{f}_n(\boldsymbol{\theta}) \,\Big\|\, s^{\backslash n} q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \Big) \tag{2.123}$$



(a) A joint distribution with two closely connected modes, created from a data set with $N = 20$ examples, is shown in red. In black is the EP solution, including both modes. Note: this is *not* the global KL solution.

(b) The joint distribution from the same data set as figure 2.7(a) is shown here, except that four of the data points (and hence likelihood terms) were slightly adjusted to allow EP to settle on one mode and break symmetry.

FIGURE 2.7: One mode or both? As the modes in the joint (or posterior) distribution become more separated, EP here chooses one of them.

is minimized over $\tilde{f}_n(\boldsymbol{\theta})$. With the cavity distribution defined as $s^{\setminus n} q^{\setminus n}(\boldsymbol{\theta}) = \prod_{i \neq n} \tilde{s}_i \tilde{f}'_i(\boldsymbol{\theta})$, and shorthand $q'(\boldsymbol{\theta}) = \prod_n \tilde{f}'_n(\boldsymbol{\theta})$, and can write the derivative with respect to an individual scale as

$$\begin{aligned}
\frac{\partial \mathsf{KL}}{\partial \tilde{s}_n} &= \frac{\partial}{\partial \tilde{s}_n} \Bigg[\tilde{s}_n \Big(\prod_{i \neq n} \tilde{s}_i \Big) \int q'(\boldsymbol{\theta}) \ln \frac{\tilde{s}_n \tilde{f}'_n(\boldsymbol{\theta})}{f_n(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \\
&\qquad - \tilde{s}_n \Big(\prod_{i \neq n} \tilde{s}_i \Big) \int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta} + \int f_n(\boldsymbol{\theta}) \prod_{i \neq n} \tilde{s}_i \tilde{f}_i(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \Bigg] \\
&= \ln \tilde{s}_n \Big(\prod_{i \neq n} \tilde{s}_i \Big) \int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta} + \Big(\prod_{i \neq n} \tilde{s}_i \Big) \int q'(\boldsymbol{\theta}) \ln \frac{\tilde{f}'_n(\boldsymbol{\theta})}{f_n(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \, . \qquad (2.124)
\end{aligned}$$

If the above expression is set to zero we find

$$\tilde{s}_n = \exp \left\{ \frac{\int q'(\boldsymbol{\theta}) \ln \frac{f_n(\boldsymbol{\theta})}{\tilde{f}'_n(\boldsymbol{\theta})} \, d\boldsymbol{\theta}}{\int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \right\} . \qquad (2.125)$$

If $\tilde{s}_n$ is now substituted back into the equation (2.99), and the chosen factorizations $p(\mathbf{x}, \boldsymbol{\theta}) = \prod_n f_n(\boldsymbol{\theta})$ and $q'(\boldsymbol{\theta}) = \prod_n \tilde{f}'_n(\boldsymbol{\theta})$ used, we get a mass estimate

$$s = \exp \left\{ \frac{\int q'(\boldsymbol{\theta}) \sum_n \ln \frac{f_n(\boldsymbol{\theta})}{\tilde{f}'_n(\boldsymbol{\theta})} \, d\boldsymbol{\theta}}{\int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \right\} \int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = \exp \left\{ \frac{\int q'(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \boldsymbol{\theta})}{q'(\boldsymbol{\theta})} \, d\boldsymbol{\theta}}{\int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \right\} \int q'(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \, . \qquad (2.126)$$

The objective function, as a function of the parameters of the exponential term approximations, is therefore

$$- \ln s = - \int \exp \Big\{ \sum_{n=0}^{N} \boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \sum_{n=0}^{N} \ln \frac{f_n(\boldsymbol{\theta})}{\exp\{\boldsymbol{\eta}_n \boldsymbol{\phi}(\boldsymbol{\theta})\}} \, d\boldsymbol{\theta} \Big/ \int \exp \Big\{ \sum_{n=0}^{N} \boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta}$$

$$- \ln \int \exp\Big\{ \sum_{n=0}^{N} \boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\boldsymbol{\theta}) \Big\} \, d\boldsymbol{\theta} \; . \tag{2.127}$$

To see that the objective function is equivalent to that of the *global* KL divergence, all we need to do is notice that $q'(\boldsymbol{\theta}) = cq(\boldsymbol{\theta})$ for some constant c, such that $q(\boldsymbol{\theta})$ is a normalized distribution (in particular, from (2.99) the scale is $s/\prod_n \tilde{s}_n$). Substituting $q'(\boldsymbol{\theta}) = cq(\boldsymbol{\theta})$ into (2.126) gives

$$s = \exp\left\{ \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \right\} , \tag{2.128}$$

which is exactly the scale or negative exponentiated free energy that we get when minimizing the KL with VB.

2.9.3 Message passing with $\alpha = 0$ as an EM algorithm

The generic message passing algorithm used to minimize local divergences over a factor graph relies on a specific factorization of the joint distribution of interest. Uniquely for $\alpha = 0$, the objective function of message passing with local divergences and the global divergence objective function will match, and the same stationary points will be reached. Using Jensen's inequality we find a lower bound to the marginal likelihood in the usual way,

$$\begin{aligned} \ln p(\mathbf{x}) &\geq \int q(\boldsymbol{\theta}) q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}) q(\mathbf{z})} \, d\boldsymbol{\theta} d\mathbf{z} \\ &= \int q(\boldsymbol{\theta}) q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} + \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, d\boldsymbol{\theta} = \mathcal{F}[q(\boldsymbol{\theta}), q(\mathbf{z})] \; . \end{aligned} \tag{2.129}$$

From the assumption that the observations are independent and identically distributed, the factorization $q(\mathbf{z}) = \prod_n q(\mathbf{z}_n)$ can be made. In this case it is well known (Neal & Hinton, 1998) that an *incremental* algorithm that updates one $q(\mathbf{z}_n)$ at a time should increase $\mathcal{F}$ (their algorithm 7). In all their algorithms, the *entire* vector $\boldsymbol{\theta}$ is updated in the M step.

EM and generic message passing routines

When the E-step updates a single $q(\mathbf{z}_n)$, and the M-step updates $q(\boldsymbol{\theta})$, then $\mathcal{F}$ increases, as we are doing gradient ascent over functions (distributions). We may ask,

> *will the increase still hold if we factorize $q(\boldsymbol{\theta})$ into a product of functions over $\boldsymbol{\theta}$, and update only one of them in the M-step?*

It turns out that this is indeed true, provided *each* factor i in the product can be defined as

$$\tilde{f}_i(\boldsymbol{\theta}) \propto \exp\Big\{ \int q(\mathbf{z}_i) \ln p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) \, d\mathbf{z}_i \Big\} \; . \tag{2.130}$$

which *only* depends on $q(\mathbf{z}_i)$.

Before presenting a formal argument of this result, a practical example is presented: Figure 2.8 illustrates this behaviour in practice on the **galaxy** data set from section 3.7. An assumed density filtering (ADF) loop—essentially the first EP loop, which includes factors one by one into the approximation—was run to initialize all terms $\tilde{f}_n(\boldsymbol{\theta})$. On initialization, the condition given in (2.130) doesn't hold for any of the factors, and $\ln s$ doesn't show a monotonic increase.

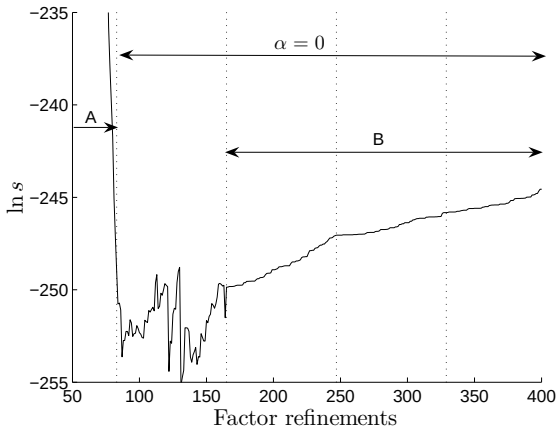


FIGURE 2.8: The **galaxy** data set from section 3.7 is used to illustrate that the objective function $\mathcal{F}$ only increases monotonically when (2.130) holds for all factors. The figure also indicates: A, the end of the ADF loop to get an initialization; B, the factor refinements with $\alpha = 0$ where (2.130) holds.

As soon as (2.139) holds for *all* factors—which can be seen in the figure after one $\alpha = 0$ loop—the objective function monotonically increases, even with partial updates to $q(\boldsymbol{\theta})$. If we solve for $\ln s$ separately in each case, using the present $q(\boldsymbol{\theta})q(\mathbf{z})$ in

$$\ln s = -\mathsf{KL}\big(q(\boldsymbol{\theta})q(\mathbf{z}) \,\big\|\, p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})\big) + \ln p(\mathbf{x}) \ , \tag{2.131}$$

we are still not guaranteed a monotonic increase until (2.130) holds, as figure 2.9 illustrates for the same problem.

To determine formally when $\mathcal{F}$ will always increase, write

$$q(\boldsymbol{\theta}) = \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta}) \ , \tag{2.132}$$

for some factorization of the approximation, and assume $q(\boldsymbol{\theta})$ integrates to one. Let this be *any* factorization—we subscript it with m to indicate that it needn't be over individual data likelihoods.

Now write $\mathcal{F}[q]$ as (from i.i.d. assumption)

$$\mathcal{F}[q] = \sum_{n=1}^{N} \int \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta}) q(\mathbf{z}_n) \ln \frac{p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})}{q(\mathbf{z}_n)} \, d\boldsymbol{\theta} d\mathbf{z}_n + \int \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta})}{\prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \ . \tag{2.133}$$

If we set $\partial\mathcal{F}/\partial q(\mathbf{z}_n)$ to zero, we arrive at the familiar expression

$$q(\mathbf{z}_n) \propto \exp\Big\{ \int q(\boldsymbol{\theta}) \ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \Big\} \ , \tag{2.134}$$

where Lagrange multipliers would give the correct normalization (there is an integration constraint to keep $q(\mathbf{z}_n)$ as a distribution). This is the particular E-step in the subproblem that is solved by the message passing scheme. The *present* factorized approximation $\prod_m \tilde{t}_m(\boldsymbol{\theta})$ is used, and we find that the objective function increases.

Now take the functional derivative with respect to one of the m terms, with an added integration constraint through a Lagrange multiplier ℓ,

$$\frac{\partial\mathcal{F}}{\partial\tilde{t}_i(\boldsymbol{\theta})} = \frac{\partial}{\partial\tilde{t}_i(\boldsymbol{\theta})} \Bigg[\int \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta}) q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \, d\boldsymbol{\theta} d\mathbf{z} + \int \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta}) \ln p(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

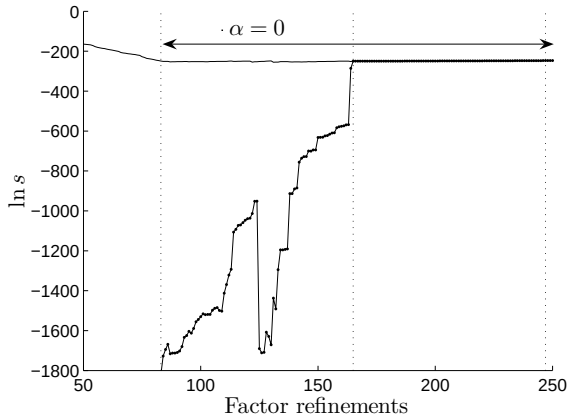


FIGURE 2.9: This graph zooms out on figure 2.8 to show the approximation $\ln s$ as the lower line, if the present $q(\boldsymbol{\theta})$ is used in (2.131). The large difference in the first $\alpha = 0$ loop (an ADF loop was used to initialize the approximation) is due to the fact that (2.131) ensures a lower bound to the joint $p(\mathbf{x}, \boldsymbol{\theta})$. If the present approximation is relatively wide, a small scale is needed to ensure that the bound holds. As soon as (2.130) holds at the end of the first $\alpha = 0$ loop, the message passing scale and the scale determined from (2.131) match.

$$
\begin{aligned}
&- \sum_{k=1}^{M} \int \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta}) \ln \tilde{t}_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} \Bigg] + \ell \frac{\partial}{\partial \tilde{t}_i(\boldsymbol{\theta})} \Bigg[\int \prod_{m=1}^{M} \tilde{t}_m(\boldsymbol{\theta})\, d\boldsymbol{\theta} - 1 \Bigg] \\
&= \prod_{m \neq i} \tilde{t}_m(\boldsymbol{\theta}) \Bigg[\int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})}\, d\mathbf{z} + \ln p(\boldsymbol{\theta}) - \ln \tilde{t}_i(\boldsymbol{\theta}) - \ln \prod_{m \neq i} \tilde{t}_m(\boldsymbol{\theta}) + (\ell - 1) \Bigg] \,.
\end{aligned}
\tag{2.135}
$$

The above derivative is equated to zero and solved, hence

$$
\tilde{t}_i(\boldsymbol{\theta}) \prod_{m \neq i} \tilde{t}_m(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \exp\Big\{ \int q(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\, d\mathbf{z} \Big\} \exp\Big\{ - \int q(\mathbf{z}) \ln q(\mathbf{z})\, d\mathbf{z} \Big\} \exp\{\ell - 1\} \,. \tag{2.136}
$$

Integrating on both sides (the left hand side is one), taking logs, solving for ℓ, and substituting back, gives

$$
\tilde{t}_i(\boldsymbol{\theta}) \prod_{m \neq i} \tilde{t}_m(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \frac{\exp\{\int q(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\, d\mathbf{z}\}}{\int \exp\{\int q(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\, d\mathbf{z}\}\, d\boldsymbol{\theta}} \,. \tag{2.137}
$$

We are again on familiar territory, as the left hand side equals $q(\boldsymbol{\theta})$. But now only one factor $\tilde{t}_i(\boldsymbol{\theta})$ is updated, and the rest is left fixed. The update will still give coordinate ascent, but we need to show an equivalence with the message passing formulation.

To see an equivalence with the generic message passing algotihm, let $m \equiv n$, so that the approximate factors are therefore $\tilde{f}_n(\boldsymbol{\theta}) \propto \tilde{t}_n(\boldsymbol{\theta}) q(\mathbf{z}_n)$. From the i.i.d. assumption we have,

$$
\tilde{t}_i(\boldsymbol{\theta}) \prod_{n \neq i} \tilde{t}_n(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\mathbf{z}}} p(\boldsymbol{\theta}) \prod_{n \neq i} \exp\Big\{ \int q(\mathbf{z}_n) \ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})\, d\mathbf{z}_n \Big\} \exp\Big\{ \int q(\mathbf{z}_i) \ln p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})\, d\mathbf{z}_i \Big\} \,. \tag{2.138}
$$

If we choose $\tilde{t}_0(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})$ and $\tilde{t}_n(\boldsymbol{\theta}) \propto \exp\{\int q(\mathbf{z}_n) \ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta})\, d\mathbf{z}_n\}$ for $n \neq i$, the product $\prod_{n\neq i} \tilde{t}_n(\boldsymbol{\theta})$ divides away on both sides, and the update simplifies to

$$\tilde{t}_i(\boldsymbol{\theta}) \prod_{n\neq i} \tilde{t}_n(\boldsymbol{\theta}) \propto \tilde{t}_0(\boldsymbol{\theta}) \prod_{n\neq i} \tilde{t}_n(\boldsymbol{\theta}) \exp\left\{ \int q(\mathbf{z}_i) \ln p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})\, d\mathbf{z}_i \right\}$$

$$\tilde{t}_i(\boldsymbol{\theta}) \propto \exp\left\{ \int q(\mathbf{z}_i) \ln p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})\, d\mathbf{z}_i \right\}, \tag{2.139}$$

which *only* depends on $q(\mathbf{z}_i)$. This is the key, as the full derivative simplifies as

$$\begin{aligned} \frac{d\mathcal{F}}{d\tilde{t}_i(\boldsymbol{\theta})} &= \frac{\partial \mathcal{F}}{\partial \tilde{t}_i(\boldsymbol{\theta})} + \sum_{n=1}^{N} \frac{\partial \mathcal{F}}{\partial q(\mathbf{z}_n)} \frac{\partial q(\mathbf{z}_n)}{\partial \tilde{t}_i(\boldsymbol{\theta})} \\ &= \frac{\partial \mathcal{F}}{\partial \tilde{t}_i(\boldsymbol{\theta})} + \frac{\partial \mathcal{F}}{\partial q(\mathbf{z}_i)} \frac{\partial q(\mathbf{z}_i)}{\partial \tilde{t}_i(\boldsymbol{\theta})} . \end{aligned} \tag{2.140}$$

In the E-step $\partial\mathcal{F}/\partial q(\mathbf{z}_i)$ was set to zero (and hence the derivative $\partial q(\mathbf{z}_i)/\partial \tilde{t}_i(\boldsymbol{\theta})$ need not be determined). Setting the partial derivative with respect to $\tilde{t}_i(\boldsymbol{\theta})$ to zero should ensure a monotonic increase of $\mathcal{F}$.

Similar algorithms

We are not forced to choose m to match n, but can choose each $\tilde{t}_i(\boldsymbol{\theta})$ to match any set of data items $\mathcal{A}_i$, *on the condition that* the E-step sets $\partial\mathcal{F}/\partial q(\mathbf{z}_n) = 0$ for all $n \in \mathcal{A}_i$. For the variational Bayes EM algorithm the choice $\mathcal{A}_i = \{1, \ldots, N\}$ was made, with only one term $\tilde{t}_i$. For the generalized message passing algorithm above the choice $\mathcal{A}_i = \{i\}$ was implemented, and we had N terms $\tilde{t}_i$. Any grouping of data items to terms is therefore possible. All these algorithms are possible because the use of the exclusive KL divergence allows them to have identical objective functions.

Finding $\boldsymbol{\theta}_{\mathrm{ML}}$ through message passing

In relation to the generic message passing algorithm, Neal & Hinton (1998)'s EM algorithm 9 is of particular interest, and it is briefly described here. It is used to find the *maximum likelihood* parameter setting.

When the *complete* data likelihood $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ falls in the *exponential* family, its sufficient statistics can be summarized as a sum of 'sufficient statistics contributions' from the individual likelihoods $p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}) \propto \exp\{\boldsymbol{\eta}_n^\top \boldsymbol{\phi}(\mathbf{x}_n, \mathbf{z}_n)\}$, for $n = 1, \ldots, N$. Notice that the sufficient statistics are for the complete likelihoods, and will include responsibilities for each example. Therefore, if $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \propto \exp\{\boldsymbol{\eta}^\top \boldsymbol{\phi}(\mathbf{x}, \mathbf{z})\}$, where $\boldsymbol{\eta} = \sum_{n=1}^{N} \boldsymbol{\eta}_n$, the maximum likelihood parameters $\boldsymbol{\theta}_{\mathrm{ML}}$ can be found by iteratively updating data contributions until the parameter estimate converges, akin to what we have done in the generic message passing algorithm. This is done as follows:

In looping over all data points, assume that we are updating data point n at time t. For data point n, we automatically copy $\boldsymbol{\eta}_i^{(t-1)}$ to $\boldsymbol{\eta}_i^{(t)}$, for all $i \neq n$. Given the parameter value $\boldsymbol{\theta}^{(t-1)}$ from the previous iteration, set $\boldsymbol{\eta}_n^{(t)}$ to the sufficient statistics of $p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}^{(t-1)})$. The sufficient statistics of the complete data likelihood is updated with $\boldsymbol{\eta}^{(t)} = \boldsymbol{\eta}^{(t-1)} - \boldsymbol{\eta}_n^{(t-1)} + \boldsymbol{\eta}_n^{(t)}$. Finally we set $\boldsymbol{\theta}^{(t)}$ to the $\boldsymbol{\theta}$ of maximum likelihood given $\boldsymbol{\eta}^{(t)}$.

2.10 Summary

This chapter introduced the α-divergence as a measure of distance between two possibly unnormalized probability distributions. This divergence is typically not tractable, but can be approximately minimized if a particular factorization (or factor graph) of the distribution is considered, and the minimization restricted to local computations on the factor graph. This distributed algorithm is not guaranteed to converge, although it often does and performs well in practice.

A simple mixture of Gaussians was taken as a illustrative running example, and its full multivariate case is taken next as a practical case in chapter 3.

Chapter 3

Approximate inference for multivariate mixtures

3.1 Introduction

The message passing algorithms for approximate inference, as discussed in chapter 2, can be directly extended to a multivariate mixture of Gaussians. Many of the algorithms and derivations presented here directly follow from those in chapter 2, and this chapter is meant to be read in conjunction with its predecessor.

Variational methods ($\alpha = 0$) for mixtures of Gaussians have proved to be worth their salt to the machine learning community, with the work of Attias (2000) leading a wealth of applications. The way has been paved for expectation propagation ($\alpha = 1$) by Minka (2001a)'s 'clutter problem' and treatment of mixtures with unknown weights. Subsequent work included an application of EP to infinite mixtures with known variances (Minka & Ghahramani, 2003). Chang et al. (2005) used EP for Gaussian mixtures with independent dimensions, with added parameters to determine whether a dimension was relevant to the clustering of the data.

New algorithms for mixture modeling

This chapter adds two new approaches to inference for a full multivariate mixture of Gaussians, namely EP and the more general α-divergence message passing scheme. We present a convincing argument for why and when EP should be favoured above VB, but also illustrate when there is little to choose between the two algorithms, leaving a practitioner to his personal preference. This is followed in section 3.7 with a presentation of experimental results on a number of data sets, comparing the approximate predictive distributions and log marginal likelihoods with the results obtained from VB and parallel tempering. Parallel tempering is a state of the art MCMC method, and will be further discussed in chapter 4. From this comparison we can gather that the approximate methods are perfectly suitable for model selection, and approximating the predictive distribution with high accuracy. It is also practically shown that the EP fixed points are not necessarily unique, and a fixed point may depend on both the initialization and the random order in which factor refinements take place. Both these questions were posed by Minka (2001a). Other points underlined empirically are: the log marginal likelihood estimates increase with α; the number of local solutions depends on the prior width; the discrepancy between the approximate and true log marginal likelihoods increase with model size; the marginal likelihoods give a characteristic 'Ockham hill' as model size increase, providing a useful tool for model selection.

3.2 Mixture of Gaussians

Deterministic approximate inference for a multivariate mixture of Gaussians— where the mixing weights, means, and precision matrices are unknown—naturally arises as an extension of the example in chapter 2. The unknown parameters are $\boldsymbol{\theta} \equiv \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\} \equiv \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j\}_{j=1}^J$, with the mixing weights π_j summing to one. The likelihood of observing a single data point is

$$p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) = \sum_{j=1}^J \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1}) \ , \tag{3.1}$$

and under an assumption that the data is independent and identically distributed, the likelihood of the data set is the product of the individual likelihoods.

For the likelihood, we choose conjugate priors, meaning that the posterior distribution will be in the same family as our choice of prior. Therefore, let the prior on the mixing weights and component parameters be Dirichlet and Normal-Wishart respectively. The choice of approximating distribution q will also match the form of the prior distribution.

The Dirichlet distribution is defined, for nonnegative δ_j, as

$$\mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}) = \frac{\Gamma(\sum_{j=1}^J \delta_j)}{\prod_{j=1}^J \Gamma(\delta_j)} \prod_{j=1}^J \pi_j^{\delta_j - 1} = \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta})} \prod_{j=1}^J \pi_j^{\delta_j - 1} \ . \tag{3.2}$$

The Normal and Wishart distributions, where we keep the notation to that of Normal-Wishart to follow in (3.5), are defined as

$$\mathcal{N}\big(\boldsymbol{\mu}_j|\mathbf{m}_j, (v_j\boldsymbol{\Lambda}_j)^{-1}\big) = \Big(\frac{v_j}{2\pi}\Big)^{\frac{d}{2}} |\boldsymbol{\Lambda}_j|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\mathsf{tr}[(\boldsymbol{\mu}_j - \mathbf{m}_j)(\boldsymbol{\mu}_j - \mathbf{m}_j)^\top v_j \boldsymbol{\Lambda}_j]\Big\} \tag{3.3}$$

$$\mathcal{W}(\boldsymbol{\Lambda}_j|a_j, \mathbf{B}_j) = \frac{|\mathbf{B}_j|^{a_j}}{\prod_{i=1}^d \Gamma(a_j + \frac{1-i}{2})} \pi^{\frac{-d(d-1)}{4}} |\boldsymbol{\Lambda}_j|^{a_j - \frac{d+1}{2}} \exp\Big\{ -\mathsf{tr}[\mathbf{B}_j\boldsymbol{\Lambda}_j]\Big\} \ . \tag{3.4}$$

The Wishart's parameterization here is slightly unorthodox, but is chosen so that the one-dimensional case will exactly reduce to the Gamma distribution, $\mathcal{G}(\lambda_j|a_j, b_j) = b_j^{a_j}/\Gamma(a_j) \cdot \lambda_j^{a_j-1} e^{-b_j\lambda_j}$. The Normal-Wishart distribution is defined as

$$\begin{aligned}\mathcal{NW}(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j|\mathbf{m}_j, v_j, a_j, \mathbf{B}_j) &= \mathcal{N}\big(\boldsymbol{\mu}_j|\mathbf{m}_j, (v_j\boldsymbol{\Lambda}_j)^{-1}\big)\mathcal{W}(\boldsymbol{\Lambda}_j|a_j, \mathbf{B}_j) \\ &= \frac{1}{\mathcal{Z}_{\mathcal{NW}}(v_j, a_j, \mathbf{B}_j)} \exp\Big\{ -\frac{1}{2}v_j\boldsymbol{\mu}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j + v_j\mathbf{m}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j \\ &\quad - \mathsf{tr}[(\mathbf{B}_j + \frac{1}{2}v_j\mathbf{m}_j\mathbf{m}_j^\top)\boldsymbol{\Lambda}_j] + (a_j - \frac{d}{2})\ln|\boldsymbol{\Lambda}_j|\Big\}\end{aligned} \tag{3.5}$$

$$\mathcal{Z}_{\mathcal{NW}}(v_j, a_j, \mathbf{B}_j) = \Big(\frac{2\pi}{v_j}\Big)^{\frac{d}{2}} \pi^{\frac{d(d-1)}{4}} \frac{\prod_{i=1}^d \Gamma(a_j + \frac{1-i}{2})}{|\mathbf{B}_j|^{a_j}} \ . \tag{3.6}$$

Let the prior for the components, as well as the approximating distribution, be factorized Normal-Wisharts,

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{j=1}^J \mathcal{NW}(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j|\mathbf{m}_{0j}, v_{0j}, a_{0j}, \mathbf{B}_{0j}) \quad \text{and} \quad q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{j=1}^J \mathcal{NW}(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j|\mathbf{m}_j, v_j, a_j, \mathbf{B}_j) \ . \tag{3.7}$$

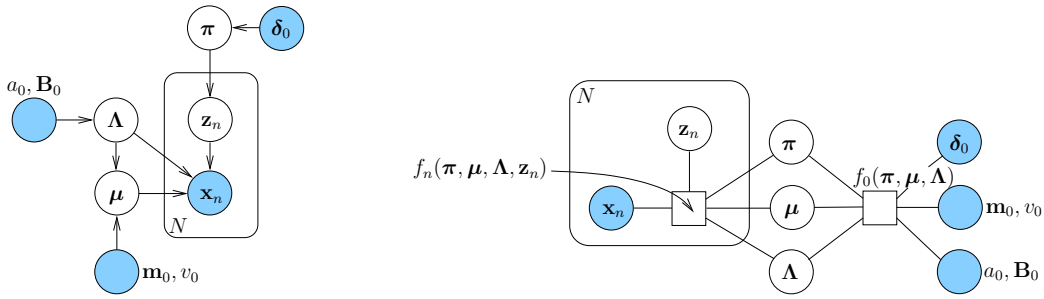


FIGURE 3.1: The Bayesian network, illustrating the parameter dependencies for a mixture of Gaussians, is shown on the *left*. The joint distribution was in this case completed with latent variables $\mathbf{z}_n$. From the Bayesian network we can choose a particular factorization, as illustrated in the factor graph on the *right*. The factor graph clarifies the dependence of the factors on the different variables.

For a convenient notation, the parameters of the *prior* distribution are subscripted with an additional '0', and we assume that these hyperparameters are fixed, as illustrated in figure 3.1. The parameters of the approximating distribution are given without an additional subscript '0', and it is these parameters that we will adjust to find a good approximating distribution q. The prior for the mixing weights, as well as the relevant approximating distribution, is chosen to be Dirichlet,

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}_0) \quad \text{and} \quad q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}) \ . \tag{3.8}$$

We will therefore approximate the joint distribution $p(\mathbf{x}, \boldsymbol{\theta}) \equiv p(\mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ by $sq(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{\pi})$. When a divergence measure other than setting α to one is chosen, say for example in variational Bayes, the likelihood is completed with latent variables $\mathbf{z}$, and the joint distribution $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ will be approximated with $sq(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{\pi})q(\mathbf{z})$.

3.3 Expectation propagation: a single observation

We embark on the EP road in exactly the same way as we have done in the simple case with unknown means in chapter 2, by computing the exact marginal for the scale, and using responsibility-weighted moment-matching equations to find the parameter updates.

If we know which component generated the data point in question, the following integral will prove to be extremely useful. It is the normalizer on observing a data point, or the likelihood averaged over a specific Normal-Wishart prior,

$$\begin{aligned} p_j(\mathbf{x}_n) &= \int \mathcal{W}(\boldsymbol{\Lambda}_j|a_{0j}, \mathbf{B}_{0j}) \int \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_{0j}, (v_{0j}\boldsymbol{\Lambda}_{0j})^{-1})\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})\, d\boldsymbol{\mu}_j\, d\boldsymbol{\Lambda}_j \\ &= \mathcal{T}\Big(\mathbf{x}_n \,\Big|\, \mathbf{m}_{0j}, \frac{v_{0j}+1}{v_{0j}} \frac{2\mathbf{B}_{0j}}{2a_{0j}-d+1}, 2a_{0j}-d+1\Big) \ . \end{aligned} \tag{3.9}$$

The full derivation of this integral is given in appendix A.6. Two more relations will prove useful in later derivations in this section:

$$\Gamma(\delta+1) = \delta\Gamma(\delta) \tag{3.10}$$

$$\Psi(\delta+1) = \frac{d}{d\delta}\ln\Gamma(\delta+1) = \frac{d}{d\delta}[\ln\delta + \ln\Gamma(\delta)] = \frac{1}{\delta} + \Psi(\delta) \ . \tag{3.11}$$

The gamma function $\Gamma(\delta)$ is defined as $\Gamma(\delta) = \int_0^\infty t^{\delta-1}e^{-t}dt$, and $\Psi(\delta)$ is the digamma function,

given by the logarithmic derivative of the gamma function,

$$\Psi(\delta) = \frac{d}{d\delta} \ln \Gamma(\delta) = \frac{\Gamma'(\delta)}{\Gamma(\delta)} \ . \tag{3.12}$$

The following exposition gives the scale and parameters for matching $sq(\boldsymbol{\theta})$ to a prior multiplied by a likelihood for one data point.

3.3.1 The scale

The scale is determined with $s = \int p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi})\, d\boldsymbol{\mu}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\pi}$, and the result

$$s = \frac{1}{\sum_{j=1}^{J} \delta_{0j}} \sum_{k=1}^{J} \delta_{0k} p_k(\mathbf{x}_n) \ , \tag{3.13}$$

is derived in appendix A.3. Define the responsibilities to be used in sections 3.3.2 and 3.3.3 as

$$r_{nj} = \frac{\delta_{0j} p_j(\mathbf{x}_n)}{\sum_{k=1}^{J} \delta_{0k} p_k(\mathbf{x}_n)} = \frac{\delta_{0j} \mathcal{T}\Big(\mathbf{x}_n \,\Big|\, \mathbf{m}_{0j},\ \dfrac{v_{0j}+1}{v_{0j}} \dfrac{2\mathbf{B}_{0j}}{2a_{0j}-d+1},\ 2a_{0j}-d+1\Big)}{\displaystyle\sum_{k=1}^{J} \delta_{0k} \mathcal{T}\Big(\mathbf{x}_n \,\Big|\, \mathbf{m}_{0k},\ \dfrac{v_{0k}+1}{v_{0k}} \dfrac{2\mathbf{B}_{0k}}{2a_{0k}-d+1},\ 2a_{0k}-d+1\Big)} \ . \tag{3.14}$$

3.3.2 Parameter updates for the components

To get the weighed parameter updates for $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, the following expectations under a Normal-Wishart distribution are needed. In each case the 'prior' expectation (the expectation if we know that component j has *not* generated observation $\mathbf{x}_n$), and the 'posterior' expectation (the expectation if we *do know* that component j was responsible for generating observation $\mathbf{x}_n$), is given. The responsibilities (3.14) will blend these expectations when we match moments. The expectations, derived in appendix A.6, are

$$\langle \boldsymbol{\Lambda}_j \rangle = a_{0j} \mathbf{B}_{0j}^{-1} \tag{3.15}$$

$$\langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle = \Big(a_{0j} + \frac{1}{2}\Big) \Big[\mathbf{B}_{0j} + \frac{1}{2} \frac{v_{0j}}{v_{0j}+1} (\mathbf{x}_n - \mathbf{m}_{0j})(\mathbf{x}_n - \mathbf{m}_{0j})^\top\Big]^{-1} \tag{3.16}$$

$$\langle \ln |\boldsymbol{\Lambda}_j| \rangle = \sum_{i=1}^{d} \Psi\Big(a_{0j} + \frac{1-i}{2}\Big) - \ln |\mathbf{B}_{0j}| \tag{3.17}$$

$$\langle \ln |\boldsymbol{\Lambda}_j| \, |\mathbf{x}_n \rangle = \sum_{i=1}^{d} \Psi\Big(a_{0j} + \frac{1}{2} + \frac{1-i}{2}\Big) - \ln \Big|\mathbf{B}_{0j} + \frac{1}{2} \frac{v_{0j}}{v_{0j}+1} (\mathbf{x}_n - \mathbf{m}_{0j})(\mathbf{x}_n - \mathbf{m}_{0j})^\top\Big| \tag{3.18}$$

$$\langle \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j \rangle = \langle \boldsymbol{\Lambda}_j \rangle \mathbf{m}_{0j} \tag{3.19}$$

$$\langle \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j | \mathbf{x}_n \rangle = \langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle \frac{v_{0j} \mathbf{m}_{0j} + \mathbf{x}_n}{v_{0j}+1} \tag{3.20}$$

$$\langle \boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j \rangle = \frac{d}{v_{0j}} + \mathbf{m}_{0j}^\top \langle \boldsymbol{\Lambda}_j \rangle \mathbf{m}_{0j} \tag{3.21}$$

$$\langle \boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j | \mathbf{x}_n \rangle = \frac{d}{v_{0j}+1} + \Big(\frac{v_{0j} \mathbf{m}_{0j} + \mathbf{x}_n}{v_{0j}+1}\Big)^\top \langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle \Big(\frac{v_{0j} \mathbf{m}_{0j} + \mathbf{x}_n}{v_{0j}+1}\Big) \ . \tag{3.22}$$

In the same way as we have done for the one-dimensional component means in chapter 2, we can derive the parameter updates by again minimizing the KL-divergence with respect to the

different parameters of $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. The parameters $\{\mathbf{m}_j, v_j, a_j, \mathbf{B}_j\}_{j=1}^J$ of $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ that minimizes $\mathsf{KL}(p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) \| sq(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{\pi}))$ are

$$\mathbf{m}_j = \tilde{\boldsymbol{\Lambda}}_j^{-1}\big[(1 - r_{nj})\langle \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j \rangle + r_{nj}\langle \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j | \mathbf{x}_n \rangle\big] \tag{3.23}$$

$$\frac{d}{v_j} = (1 - r_{nj})\langle \boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j \rangle + r_{nj}\langle \boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j | \mathbf{x}_n \rangle - \mathbf{m}_j^\top \tilde{\boldsymbol{\Lambda}}_j \mathbf{m}_j \tag{3.24}$$

$$\sum_{i=1}^d \Psi\Big(a_j + \frac{1-i}{2}\Big) - \ln|\mathbf{B}_j| = (1 - r_{nj})\langle \ln|\boldsymbol{\Lambda}_j| \rangle + r_{nj}\langle \ln|\boldsymbol{\Lambda}_j| | \mathbf{x}_n \rangle \equiv c_2 \tag{3.25}$$

$$a_j \mathbf{B}_j^{-1} = \tilde{\boldsymbol{\Lambda}}_j = (1 - r_{nj})\langle \boldsymbol{\Lambda}_j \rangle + r_{nj}\langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle \equiv \mathbf{C}_1 \ . \tag{3.26}$$

(Both $\mathbf{C}_1$ and c_2 are merely shorthands to keep the Newton method that follows below concise.)

Solving for a_j and $\mathbf{B}_j$ in equations (3.25) and (3.26)

We have, for constants $\mathbf{C}_1$ and c_2, $a_j \mathbf{B}_j^{-1} = \mathbf{C}_1$ and $\sum_{i=1}^d \Psi(a_j + (1-i)/2) - \ln|\mathbf{B}_j| = c_2$. Furthermore, $\ln|a_j \mathbf{I}| - \ln|\mathbf{B}_j| = \ln|\mathbf{C}_1|$, and hence $-\ln|\mathbf{B}_j| = \ln|\mathbf{C}_1| - d\ln a_j$. We can therefore solve for a_j with Newton's method by writing (3.25) as

$$\sum_{i=1}^d \Psi\Big(a_j + \frac{1-i}{2}\Big) - d\ln a_j + \ln|\mathbf{C}_1| - c_2 = 0 \ . \tag{3.27}$$

As another shorthand, define constant $c = c_2 - \ln|\mathbf{C}_1|$, and let

$$g(a_j) = \sum_{i=1}^d \Psi\Big(a_j + \frac{1-i}{2}\Big) - d\ln a_j - c \tag{3.28}$$

$$g'(a_j) = \sum_{i=1}^d \Psi'\Big(a_j + \frac{1-i}{2}\Big) - \frac{d}{a_j} \ , \tag{3.29}$$

where $\Psi'(a_j)$ is called the trigamma function, the first derivative of the digamma function. For Newton's method we choose an initial a_j, and update it until convergence with

$$a_j^{\text{new}} = a_j - \frac{g(a_j)}{g'(a_j)} = a_j \left[1 - \frac{\sum_{i=1}^d \Psi(a_j + (1-i)/2) - d\ln a_j - c}{a_j \sum_{i=1}^d \Psi'(a_j + (1-i)/2) - d}\right] \ . \tag{3.30}$$

The Wishart distribution is only defined for $a_j > (d-1)/2$, and there is no guarantee that our update equations will satisfy this constraint, because the gradient in the Newton-Raphson method is taken only locally. The choice of a_{0j} as an initial value of a_j may be good, but comes without guarantee: As the digamma function is concave and monotonically increasing for positive arguments, a choice of an initial value *larger* than the solution may cause a negative argument to be passed to the digamma function in the following iteration, causing havoc in the iterative scheme.

We can reparameterize the fixed point equation to only allow for permissible values with $a_j = \exp\{a_j'\} + (d-1)/2$, and solve for a_j'. Write equation (3.28) as a function of a_j', and divide by the derivative with respect to a_j' to get an update for $a_j'^{\text{new}}$. For brevity, define $k = (d-1)/2$. Some rearrangement allows the update in terms of a_j to be multiplicative,

$$a_j^{\text{new}} = (a_j - k)\exp\left\{-\frac{\sum_{i=1}^d \Psi(a_j + (1-i)/2) - d\ln a_j - c}{(a_j - k)\sum_{i=1}^d \Psi'(a_j + (1-i)/2) - d(a_j - k)/a_j}\right\} + k \ . \tag{3.31}$$

On convergence of the Newton-Raphson method, a_j is used to solve for $\mathbf{B}_j$ in equation (3.26).

3.3.3 Parameter updates for the mixing weights

To update the mixing parameters, we find the parameter setting $\boldsymbol{\delta}$ of $q(\boldsymbol{\pi})$ that minimizes $\mathsf{KL}(p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) \| sq(\boldsymbol{\mu}, \boldsymbol{\Lambda}) q(\boldsymbol{\pi}))$. Equating the derivative of the KL divergence with respect to each δ_j to zero, we find that

$$\begin{aligned} \langle \ln \pi_j \rangle_q &= \langle \ln \pi_j | \mathbf{x}_n \rangle \\ \Psi(\delta_j) - \Psi\Big(\sum_{i=1}^J \delta_i\Big) &= \Psi(\delta_{0j}) - \Psi\Big(\sum_{i=1}^J \delta_{0i}\Big) - \frac{1}{\sum_{i=1}^J \delta_{0i}} + \frac{r_{nj}}{\delta_{0j}} \ . \end{aligned} \tag{3.32}$$

This gives a set of J coupled update equations that we need to solve to find $\boldsymbol{\delta}$. The second expectation is taken given that we observed $\mathbf{x}_n$, i.e. over the posterior distribution, and its moments are derived in appendix A.2. Appendix A.2 also shows that the right hand side of (3.32), when not shown in the simple form given here, also contains a *responsibility-weighted* sum that often occurs when dealing with mixture model moments.

Solving for $\boldsymbol{\delta}$ in equation (3.32)

Parameter vector $\boldsymbol{\delta}$ can again be solved for using Newton's method. Two implementations of Newton's method are presented here; the first requires a matrix inversion, while the second removes the need for a matrix inversion[1].

Method 1. For (3.32) we define a constant c_j with $\Psi(\delta_{0j}) - \Psi(\sum_{i=1}^J \delta_{0i}) - 1/\sum_{i=1}^J \delta_{0i} + r_{nj}/\delta_{0j} = c_j$. For Newton's method we need to solve a system of equations, $\mathbf{g}(\boldsymbol{\delta}) = \mathbf{0}$, and therefore we let $\mathbf{g}$ be a column vector containing the different function evaluations,

$$g(\boldsymbol{\delta})_j = \Psi(\delta_j) - \Psi(\Delta) - c_j \ , \tag{3.33}$$

where $\Delta = \sum_{i=1}^J \delta_i$. Define the Jacobian $\mathbf{J}$ to be a matrix with entries

$$g'(\boldsymbol{\delta})_{ji} = J_{ji} = \frac{\partial g(\boldsymbol{\delta})_j}{\partial \delta_i} = \mathbb{I}_{ji} \Psi'(\delta_j) - \Psi'(\Delta) \ , \tag{3.34}$$

where $\Psi'(a)$ is the trigamma function, and $\mathbb{I}_{ji} = 1$ if $j = i$ and zero otherwise. With Newton's method we choose an initial $\boldsymbol{\delta}$ (e.g. the present approximation's $\boldsymbol{\delta}$ with the responsibility vector $\mathbf{r}_n$ added, may be a good choice), and update it until convergence with

$$\boldsymbol{\delta}^{\text{new}} = \boldsymbol{\delta} - \mathbf{J}^{-1}\mathbf{g} = \boldsymbol{\delta} - \mathbf{J}(\boldsymbol{\delta})^{-1}\mathbf{g}(\boldsymbol{\delta}) \ . \tag{3.35}$$

As shown by Minka (2000), the matrix need not be inverted explicitly. As $\mathbf{J} = \mathbf{D} + \mathbf{1}\mathbf{1}^\top \Psi'(\Delta)$, where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \Psi'(\delta_i)$, and $\mathbf{1}$ is an all-one column vector, the matrix inversion lemma (see appendix A.7) can be used to obtain

$$\mathbf{J}^{-1} = \mathbf{D}^{-1} - \frac{\mathbf{D}^{-1}\mathbf{1}\mathbf{1}^\top\mathbf{D}^{-1}}{1/\Psi'(\Delta) + \mathbf{1}^\top\mathbf{D}^{-1}\mathbf{1}} \tag{3.36}$$

$$(\mathbf{J}^{-1}\mathbf{g})_j = \frac{1}{D_{jj}}\left[g_j - \frac{\sum_{i=1}^J g_i/D_{ii}}{1/\Psi'(\Delta) + \mathbf{1}^\top\mathbf{D}^{-1}\mathbf{1}}\right] = \frac{1}{D_{jj}}\left[g_j - \frac{\sum_{i=1}^J g_i/D_{ii}}{1/\Psi'(\Delta) + \sum_{i=1}^J 1/D_{ii}}\right] \ . \tag{3.37}$$

[1]Thanks to Ole Winther for pointing this out.

Method 2. Instead of solving for each δ_j in (3.32), we can solve for the digamma function of their sum, and from that recover each of the vector components. Define Δ as the sum $\sum_{i=1}^{J} \delta_i$, so that we now have, for each j, $\Psi(\delta_j) - \Psi(\Delta) = c_j$. A unique expression for δ_j arises after taking the inverse of the digamma function,

$$\delta_j = \Psi^{-1}(c_j + \Psi(\Delta)) \ . \tag{3.38}$$

If we now sum over j in (3.38), and take both sides as arguments to the digamma function, the equation

$$\Psi(\Delta) = \Psi\Big(\sum_{j=1}^{J} \Psi^{-1}(c_j + \Psi(\Delta))\Big) \tag{3.39}$$

can be used in Newton's method to solve for $\Psi(\Delta)$ in the usual way. Notice that a solution is obtained for the digamma evaluation of Δ, and not Δ itself, so that the solution can be directly substituted into (3.38). Let

$$g\big(\Psi(\Delta)\big) = \Psi(\Delta) - \Psi\Big(\sum_{j=1}^{J} \Psi^{-1}(c_j + \Psi(\Delta))\Big) \tag{3.40}$$

$$g'\big(\Psi(\Delta)\big) = 1 - \Psi'\Big(\sum_{j=1}^{J} \Psi^{-1}(c_j + \Psi(\Delta))\Big)\left[\sum_{j=1}^{J} \Psi^{-1\prime}(c_j + \Psi(\Delta))\Big)\right] , \tag{3.41}$$

so that

$$\Psi(\Delta)^{\text{new}} = \Psi(\Delta) - \frac{g(\Psi(\Delta))}{g'(\Psi(\Delta))} \ . \tag{3.42}$$

On solving for $\Psi(\Delta)$, each δ_j can be recovered from (3.38).

3.4 Variational Bayes: a single observation

The introduction of latent allocation variables—turning the joint distribution into a product—allows the exclusive KL divergence $\mathsf{KL}(p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) \| sq(\boldsymbol{\theta})q(\mathbf{z}))$ to be minimized with a well-known EM algorithm. The joint distribution becomes a product with

$$\begin{aligned} p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}, \mathbf{z}_n) &= p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \mathbf{z}_n)p(\mathbf{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= \prod_{j=1}^{J} \left[\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)\right]^{z_{nj}} p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \ . \end{aligned} \tag{3.43}$$

The joint can now be approximated with the factorized distribution $sq(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{\pi})q(\mathbf{z}_n)$, where $q(\mathbf{z}_n)$ is a multinomial distribution, $q(\mathbf{z}_n) = \prod_{j=1}^{J} \gamma_{nj}^{z_{nj}}$, with $\gamma_{nj} \geq 1$ and $\sum_j \gamma_{nj} = 1$.

3.4.1 Parameter updates

In a similar fashion to section 2.5, we get an iterative optimization scheme comprising of an expectation and a maximization step.

E-step. For the expectation step the *present* approximation $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{\pi})$ in the loop is used, and the following holds:

$$\int q(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{\pi}) \ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) \, d\boldsymbol{\mu} \, d\boldsymbol{\Lambda} \, d\boldsymbol{\pi}$$

$$= \sum_{j=1}^{J} z_{nj} \Bigg[\Psi(\delta_j) - \Psi\Big(\sum_{i=1}^{J} \delta_i\Big) + \frac{1}{2}\Big(\sum_{i=1}^{d} \Psi\Big(a_j + \frac{1-i}{2}\Big) - \ln|\mathbf{B}_j|\Big)$$
$$- \frac{1}{2}(\mathbf{x}_n - \mathbf{m}_j)^\top a_j \mathbf{B}_j^{-1}(\mathbf{x}_n - \mathbf{m}_j) - \frac{1}{2}\frac{d}{v_j} - \frac{d}{2}\ln 2\pi \Bigg] . \tag{3.44}$$

Equation (3.44) is used in updating the approximation $q(\mathbf{z}_n)$ with

$$\tilde{\gamma}_{nj} = \exp\Bigg\{ \Psi(\delta_j) - \Psi\Big(\sum_{i=1}^{J} \delta_i\Big) + \frac{1}{2}\Big(\sum_{i=1}^{d} \Psi\Big(a_j + \frac{1-i}{2}\Big) - \ln|\mathbf{B}_j|\Big)$$
$$- \frac{1}{2}(\mathbf{x}_n - \mathbf{m}_j)^\top a_j \mathbf{B}_j^{-1}(\mathbf{x}_n - \mathbf{m}_j) - \frac{1}{2}\frac{d}{v_j} \Bigg\} , \tag{3.45}$$

and finally $\gamma_{nj} = \tilde{\gamma}_{nj} / \sum_k \tilde{\gamma}_{nk}$.

M-step. For the maximization step, we update the component parameters, for which we use the *prior*, the data point $\mathbf{x}_n$, and the *present* approximation $q(\mathbf{z}_n)$ (or γ_{nj}). Note that

$$\sum_{\mathbf{z}_n} q(\mathbf{z}_n) \ln\left[p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\pi})\right]$$
$$= \sum_{j=1}^{J} \Bigg[-\frac{1}{2}\Big(\boldsymbol{\mu}_j - \frac{v_{0j}\mathbf{m}_{0j} + \gamma_{nj}\mathbf{x}_n}{v_{0j} + \gamma_{nj}}\Big)^\top (v_{0j} + \gamma_{nj}) \boldsymbol{\Lambda}_j \Big(\boldsymbol{\mu}_j - \frac{v_{0j}\mathbf{m}_{0j} + \gamma_{nj}\mathbf{x}_n}{v_{0j} + \gamma_{nj}}\Big)$$
$$- \mathsf{tr}\big[(\mathbf{B}_{0j} + \frac{1}{2}\frac{v_{0j}\gamma_{nj}}{v_{0j} + \gamma_{nj}}(\mathbf{m}_{0j} - \mathbf{x}_n)(\mathbf{m}_{0j} - \mathbf{x}_n)^\top)\boldsymbol{\Lambda}_j\big] + \big(a_{0j} + \frac{\gamma_{nj}}{2} - \frac{d}{2}\big) \ln|\boldsymbol{\Lambda}_j| \Bigg]$$
$$+ \sum_{j=1}^{J} (\delta_{0j} + \gamma_{nj} - 1) \ln \pi_j + \mathsf{const} , \tag{3.46}$$

and as this is in the Dirichlet and Normal-Wishart forms of (3.2) and (3.5) we can read off the parameter updates as

$$v_j = v_{0j} + \gamma_{nj} \tag{3.47}$$
$$\mathbf{m}_j = \frac{v_{0j}\mathbf{m}_{0j} + \gamma_{nj}\mathbf{x}_n}{v_{0j} + \gamma_{nj}} \tag{3.48}$$
$$\mathbf{B}_j = \mathbf{B}_{0j} + \frac{1}{2}\frac{v_{0j}\gamma_{nj}}{v_{0j} + \gamma_{nj}}(\mathbf{m}_{0j} - \mathbf{x}_n)(\mathbf{m}_{0j} - \mathbf{x}_n)^\top \tag{3.49}$$
$$a_j = a_{0j} + \frac{\gamma_{nj}}{2} \tag{3.50}$$
$$\delta_j = \delta_{0j} + \gamma_{nj} . \tag{3.51}$$

The iterations between the E and M steps are repeated until convergence, and convergence (at least to a local minimum) is guaranteed because both steps are convex.

3.4.2 The scale

The approximate log marginal likelihood $\ln s$—or the negative variational free energy—is computed *after* the iterative EM method has converged. As we can write $\ln p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}, \mathbf{z}_n) =$

$\ln p(\mathbf{x}_n|\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{z}_n) + \ln p(\mathbf{z}_n|\boldsymbol{\pi}) + \ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})$ we have

$$\begin{aligned}\ln s = &\langle \ln p(\mathbf{x}_n|\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{z}_n)\rangle + \langle \ln p(\mathbf{z}_n|\boldsymbol{\pi})\rangle + \langle \ln p(\boldsymbol{\pi})\rangle + \langle \ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})\rangle \\ &- \langle \ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})\rangle - \langle \ln q(\boldsymbol{\pi})\rangle - \langle \ln q(\mathbf{z}_n)\rangle \end{aligned} \tag{3.52}$$

where the expectation is taken over the resulting approximation from the EM steps, $q(\boldsymbol{\mu},\boldsymbol{\Lambda})q(\boldsymbol{\pi})q(\mathbf{z}_n)$. To get $\ln s$, an approximation to the log marginal likelihood, we simply determine these expectations. Notice the difference between prior parameters (starting with a subscript '0'), and the parameters of q, which go without the '0' subscript. The expectations needed in (3.52) are

$$\begin{aligned}\langle \ln p(\mathbf{x}_n|\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{z}_n)\rangle &= \Big\langle \sum_{j=1}^{J} z_{nj} \ln p(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Lambda}_j)\Big\rangle \\ &= \frac{1}{2}\sum_{j=1}^{J}\gamma_{nj}\Big[- d\ln(2\pi) + \sum_{i=1}^{d}\Psi\Big(a_j + \frac{1-i}{2}\Big) - \ln|\mathbf{B}_j| \\ &\quad - (\mathbf{x}_n - \mathbf{m}_j)^\top a_j \mathbf{B}_j^{-1}(\mathbf{x}_n - \mathbf{m}_j) - \frac{d}{v_j}\Big] \end{aligned} \tag{3.53}$$

$$\langle \ln p(\mathbf{z}_n|\boldsymbol{\pi})\rangle = \Big\langle \sum_{j=1}^{J} z_{nj}\ln \pi_j\Big\rangle = \sum_{j=1}^{J}\gamma_{nj}\Big[\Psi(\delta_j) - \Psi\Big(\sum_{i=1}^{J}\delta_i\Big)\Big] \tag{3.54}$$

$$\langle \ln p(\boldsymbol{\pi})\rangle = \ln \mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0) + \sum_{j=1}^{J}(\delta_{0j}-1)\Big[\Psi(\delta_j) - \Psi\Big(\sum_{i=1}^{J}\delta_i\Big)\Big] \tag{3.55}$$

$$\begin{aligned}\langle \ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})\rangle = &\sum_{j=1}^{J}\Big[- \ln \mathcal{Z}_{\mathcal{NW}}(v_{0j},a_{0j},\mathbf{B}_{0j}) - \frac{d}{2}\frac{v_{0j}}{v_j} - \frac{1}{2}v_{0j}(\mathbf{m}_j - \mathbf{m}_{0j})^\top a_j \mathbf{B}_j^{-1}(\mathbf{m}_j - \mathbf{m}_{0j}) \\ &- a_j \mathsf{tr}\big[\mathbf{B}_{0j}\mathbf{B}_j^{-1}\big] + \Big(a_{0j} - \frac{d}{2}\Big)\Big(\sum_{i=1}^{d}\Psi\Big(a_j + \frac{1-i}{2}\Big) - \ln|\mathbf{B}_j|\Big)\Big] \end{aligned} \tag{3.56}$$

$$\begin{aligned}\langle \ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})\rangle = &\sum_{j=1}^{J}\Big[- \ln \mathcal{Z}_{\mathcal{NW}}(v_j,a_j,\mathbf{B}_j) - \frac{d}{2} \\ &- da_j + \Big(a_j - \frac{d}{2}\Big)\Big(\sum_{i=1}^{d}\Psi\Big(a_j + \frac{1-i}{2}\Big) - \ln|\mathbf{B}_j|\Big)\Big] \end{aligned} \tag{3.57}$$

$$\langle \ln q(\boldsymbol{\pi})\rangle = - \ln \mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}) + \sum_{j=1}^{J}(\delta_j - 1)\Big[\Psi(\delta_j) - \Psi\Big(\sum_{i=1}^{J}\delta_i\Big)\Big] \tag{3.58}$$

$$\langle \ln q(\mathbf{z}_n)\rangle = \sum_{j=1}^{J}\gamma_{nj}\ln\gamma_{nj}\ . \tag{3.59}$$

3.5 α-divergence: a single observation

To minimize an α-divergence, we follow the fixed point framework from section 2.6. The 'prior' in each step is the product $[p(\boldsymbol{\pi})p(\boldsymbol{\mu},\boldsymbol{\Lambda})]^{\alpha}[q_{(t)}(\boldsymbol{\pi})q_{(t)}(\boldsymbol{\mu},\boldsymbol{\Lambda})]^{1-\alpha}$, for which we define the shorthand parameters

$$\hat{v}_i = \alpha v_{0i} + (1-\alpha)v_{i(t)} \tag{3.60}$$

$$\hat{\mathbf{m}}_i = \frac{\alpha v_{0i}\mathbf{m}_{0i} + (1-\alpha)v_{i(t)}\mathbf{m}_{i(t)}}{\alpha v_{0i} + (1-\alpha)v_{i(t)}} \tag{3.61}$$

$$\hat{\mathbf{B}}_i = \alpha\mathbf{B}_{0i} + (1-\alpha)\mathbf{B}_{i(t)} + \frac{1}{2}\frac{\alpha(1-\alpha)v_{0i}v_{i(t)}}{\alpha v_{0i} + (1-\alpha)v_{i(t)}}(\mathbf{m}_{0i} - \mathbf{m}_{i(t)})(\mathbf{m}_{0i} - \mathbf{m}_{i(t)})^\top \tag{3.62}$$

$$\hat{a}_i = \alpha a_{0i} + (1-\alpha)a_{i(t)} \tag{3.63}$$

$$\hat{\delta}_j = \alpha\delta_{0j} + (1-\alpha)\delta_{j(t)} \ . \tag{3.64}$$

By examining the above parameters and the scale given in (3.66), we see that the scale may not be finite, and we suddenly find ourselves with a set of practical constrains when α is outside the interval $[0, 1]$. Some of the parameter constraints needed to ensure that the approximating distribution is normalizable, might be violated:

- from (3.60), we require $\hat{v}_i > 0$;
- from (3.62), we require $|\hat{\mathbf{B}}_k| > 0$;
- from (3.63) and the Gamma functions in (3.65), we require both $\hat{a}_k > (d-1)/2$ and $\hat{a}_k > (d-\alpha)/2$;
- from (3.64), we require $\hat{\delta}_j \geq 0$.

The experimental results given in section 3.7 only consider $\alpha = \frac{1}{2}$, the Hellinger distance, giving approximations 'between' those arising from variational Bayes and expectation propagation. For values of $\alpha > 1$, the fixed point iterations often failed on account of the constraints given above.

3.5.1 Fixed point iterations

The iterative scheme outlined in section 2.6 is now implemented on a larger scale. We start with an initial $s_{(0)}q_{(0)}(\mathbf{z}_n)q_{(0)}(\boldsymbol{\pi})q_{(0)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. The prior distribution may provide a good starting point, with the parameters $\gamma_{nj(0)}$ of $q_{(0)}(\mathbf{z}_n)$ all set to $1/J$, and $s_{(0)}$ set to one. Starting with $t = 0$, the following steps are repeated until convergence, or until some maximum number of iterations is reached.

Step 1. Determine the scale, for which we define the unscaled responsibilities r_{nj} as

$$R_{nj} = \gamma_{nj(t)}^{1-\alpha}\frac{\Gamma(\hat{\delta}_j + \alpha)}{\Gamma(\hat{\delta}_j)}|\hat{\mathbf{B}}_j|^{(1-\alpha)/2}\frac{\Gamma(\frac{[2\hat{a}_j+\alpha-d]}{2})}{\Gamma(\frac{[2\hat{a}_j+\alpha-d]+d}{2})}\prod_{l=1}^{d}\frac{\Gamma(\frac{2\hat{a}_j+\alpha+1-l}{2})}{\Gamma(\frac{2\hat{a}_j+1-l}{2})}$$
$$\times\, \mathcal{T}\Big(\mathbf{x}_n \,\Big|\, \hat{\mathbf{m}}_j,\ \frac{\hat{v}_j+\alpha}{\hat{v}_j\alpha}\frac{2\hat{\mathbf{B}}_j}{2\hat{a}_j+\alpha-d},\ 2\hat{a}_j+\alpha-d\Big)\ , \tag{3.65}$$

to get the scale,

$$s_{(t')} = s_{(t)}^{1-\alpha}\frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)^\alpha}\frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t)})^{1-\alpha}}\left(\frac{\prod_j\Gamma(\hat{\delta}_j)}{\Gamma(\alpha+\sum_{j=1}^J\hat{\delta}_j)}\right)$$
$$\times\prod_{j=1}^{J}\frac{\mathcal{Z}_{\mathcal{NW}}(\hat{v}_j, \hat{a}_j, \hat{\mathbf{B}}_j)}{\mathcal{Z}_{\mathcal{NW}}(v_{0j}, a_{0j}, \mathbf{B}_{0j})^\alpha\mathcal{Z}_{\mathcal{NW}}(v_{j(t)}, a_{j(t)}, \mathbf{B}_{j(t)})^{1-\alpha}}$$
$$\times\,(2\pi)^{(1-\alpha)d/2}\alpha^{-d/2}\sum_{k=1}^{J}R_{nk}\ . \tag{3.66}$$

The derivation of the scale is given in appendix A.4.2; to check the correctness of the derivation, substituting $\alpha = 1$ in the above scale again gives us the scale (3.13) derived for expectation propagation.

Now we find a normalized distribution $q_{(t')}(\mathbf{z}_n)q_{(t')}(\boldsymbol{\pi})q_{(t')}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ that minimizes the KL divergence to

$$p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}, \mathbf{z}_n)^\alpha [q_{(t)}(\mathbf{z}_n)q_{(t)}(\boldsymbol{\pi})q_{(t)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})]^{1-\alpha} \ . \tag{3.67}$$

Again, we have already solved for the scale and only need to match moments. The following expectations are derived in appendix A.6, will be used in the (responsibility-weighted) moment matching update equations:

$$\langle \boldsymbol{\Lambda}_j \rangle = \hat{a}_j \hat{\mathbf{B}}_j^{-1} \tag{3.68}$$

$$\langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle = \Big(\hat{a}_j + \frac{\alpha}{2}\Big)\Big[\hat{\mathbf{B}}_j + \frac{1}{2}\frac{\alpha \hat{v}_j}{\hat{v}_j + \alpha}(\mathbf{x}_n - \hat{\mathbf{m}}_j)(\mathbf{x}_n - \hat{\mathbf{m}}_j)^\top\Big]^{-1} \tag{3.69}$$

$$\langle \ln |\boldsymbol{\Lambda}_j| \rangle = \sum_{i=1}^{d} \Psi\Big(\hat{a}_j + \frac{1-i}{2}\Big) - \ln |\hat{\mathbf{B}}_j| \tag{3.70}$$

$$\langle \ln |\boldsymbol{\Lambda}_j| \, | \mathbf{x}_n \rangle = \sum_{i=1}^{d} \Psi\Big(\hat{a}_j + \frac{\alpha}{2} + \frac{1-i}{2}\Big) - \ln \Big|\hat{\mathbf{B}}_j + \frac{1}{2}\frac{\alpha \hat{v}_j}{\hat{v}_j + \alpha}(\mathbf{x}_n - \hat{\mathbf{m}}_j)(\mathbf{x}_n - \hat{\mathbf{m}}_j)^\top\Big| \tag{3.71}$$

$$\langle \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j \rangle = \langle \boldsymbol{\Lambda}_j \rangle \hat{\mathbf{m}}_j \tag{3.72}$$

$$\langle \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j | \mathbf{x}_n \rangle = \langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle \frac{\hat{v}_j \hat{\mathbf{m}}_j + \alpha \mathbf{x}_n}{\hat{v}_j + \alpha} \tag{3.73}$$

$$\langle \boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j \rangle = \frac{d}{\hat{v}_j} + \hat{\mathbf{m}}_j^\top \langle \boldsymbol{\Lambda}_j \rangle \hat{\mathbf{m}}_j \tag{3.74}$$

$$\langle \boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j | \mathbf{x}_n \rangle = \frac{d}{\hat{v}_j + \alpha} + \Big(\frac{\hat{v}_j \hat{\mathbf{m}}_j + \alpha \mathbf{x}_n}{\hat{v}_j + \alpha}\Big)^\top \langle \boldsymbol{\Lambda}_j | \mathbf{x}_n \rangle \Big(\frac{\hat{v}_j \hat{\mathbf{m}}_j + \alpha \mathbf{x}_n}{\hat{v}_j + \alpha}\Big) \ . \tag{3.75}$$

For the responsibilities, define

$$r_{nj} = \frac{R_{nj}}{\sum_{k=1}^{J} R_{nk}} \ . \tag{3.76}$$

Exactly the same weighted sum of moments as equations (3.23) to (3.26) will be used to update $\mathbf{m}_{j(t')}$, $v_{j(t')}$, $a_{j(t')}$ and $\mathbf{B}_{j(t')}$ for all components j, and we do not repeat them here. The only difference lies in the computation of the expectations and responsibilities being used.

By taking a similar route as that taken to derive the mixing weight updates in section 3.3.3, we arrive at a set of update equations,

$$\Psi(\delta_{j(t')}) - \Psi\Big(\sum_{i=1}^{J} \delta_{i(t')}\Big) = (1 - r_{nj})\Psi(\hat{\delta}_j) + r_{nj}\Psi\Big(\hat{\delta}_j + \alpha\Big) - \Psi\Big(\alpha + \sum_{i=1}^{J} \hat{\delta}_i\Big) = c_j. \tag{3.77}$$

The parameter values can be solved for with Newton's method, following the method laid out in section 3.3.3, only with different values for c_j.

Finally, the parameters of $q_{(t')}(\mathbf{z}_n)$ are set as $\gamma_{nj}(t') = r_{nj}$.

Step 2. We have a scaled distribution $s_{(t')}q_{(t')}(\mathbf{z}_n)q_{(t')}(\boldsymbol{\pi})q_{(t')}(\boldsymbol{\mu},\boldsymbol{\Lambda})$, which we damp towards the previous approximation, $s_{(t)}q_{(t)}$, to find an updated $s_{(t+1)}q_{(t+1)}$, where $q_{(t+1)}$ is normalized. Define the parameters of the (unscaled) damped approximating distributions $q_{(t)}(\boldsymbol{\pi})^\epsilon q_{(t')}(\boldsymbol{\pi})^{1-\epsilon}$ and $q_{(t)}(\boldsymbol{\mu},\boldsymbol{\Lambda})^\epsilon q_{(t')}(\boldsymbol{\mu},\boldsymbol{\Lambda})^{1-\epsilon}$ as

$$v_{j(t+1)} = \epsilon v_{j(t)} + (1-\epsilon)v_{j(t')} \tag{3.78}$$

$$\mathbf{m}_{j(t+1)} = \frac{\epsilon v_{j(t)}\mathbf{m}_{j(t)} + (1-\epsilon)v_{j(t')}\mathbf{m}_{j(t')}}{\epsilon v_{j(t)} + (1-\epsilon)v_{j(t')}} \tag{3.79}$$

$$\mathbf{B}_{j(t+1)} = \epsilon\mathbf{B}_{j(t)} + (1-\epsilon)\mathbf{B}_{j(t')} \tag{3.80}$$

$$+\frac{1}{2}\frac{\epsilon v_{j(t)}(1-\epsilon)v_{j(t')}}{\epsilon v_{j(t)} + (1-\epsilon)v_{j(t')}}(\mathbf{m}_{j(t)} - \mathbf{m}_{j(t')})(\mathbf{m}_{j(t)} - \mathbf{m}_{j(t')})^\top \tag{3.81}$$

$$a_{j(t+1)} = \epsilon a_{j(t)} + (1-\epsilon)a_{j(t')} \tag{3.82}$$

$$\boldsymbol{\delta}_{(t+1)} = \epsilon\boldsymbol{\delta}_{(t)} + (1-\epsilon)\boldsymbol{\delta}_{(t')}\ , \tag{3.83}$$

for the mixing weights and each of the mixture components j. Then damping gives

$$q_{(t)}(\boldsymbol{\pi})^\epsilon q_{(t')}(\boldsymbol{\pi})^{1-\epsilon} = \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t)})^\epsilon}\frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t')})^{1-\epsilon}}\prod_{j=1}^{J}\pi_j^{\delta_{j(t+1)}-1} \tag{3.84}$$

$$\begin{aligned} q_{(t)}(\boldsymbol{\mu},\boldsymbol{\Lambda})^\epsilon q_{(t')}(\boldsymbol{\mu},\boldsymbol{\Lambda})^{1-\epsilon} = \prod_{j=1}^{J}&\frac{1}{\mathcal{Z}_{\mathcal{NW}}(v_{(t)},a_{j(t)},\mathbf{B}_{j(t)})^\epsilon}\frac{1}{\mathcal{Z}_{\mathcal{NW}}(v_{(t')},a_{j(t')},\mathbf{B}_{j(t')})^{1-\epsilon}} \\ &\times\exp\Big\{-\frac{1}{2}(\boldsymbol{\mu}_j - \mathbf{m}_{j(t+1)})^\top(v_{j(t+1)}\boldsymbol{\Lambda}_j)(\boldsymbol{\mu}_j - \mathbf{m}_{j(t+1)}) \\ &-\mathsf{tr}[\mathbf{B}_{j(t+1)}\boldsymbol{\Lambda}_j] + \big(a_{j(t+1)} - \frac{d}{2}\big)\ln|\boldsymbol{\Lambda}_j|\Big\} \end{aligned} \tag{3.85}$$

$$q_{(t)}(\mathbf{z}_n)^\epsilon q_{(t')}(\mathbf{z}_n)^{1-\epsilon} = \prod_{j=1}^{J}[\gamma_{nj(t)}^\epsilon\gamma_{nj(t')}^{1-\epsilon}]^{z_{nj}}\ . \tag{3.86}$$

We would like to keep $q_{(t+1)}$ as a normalized distribution, for which we need to divide it by some scale, and multiply $s_{(t+1)}$ by the same scale. The required scales are

$$Z_1 = \int q_{(t)}(\boldsymbol{\pi})^\epsilon q_{(t')}(\boldsymbol{\pi})^{1-\epsilon}\,d\boldsymbol{\pi} = \frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t+1)})}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t)})^\epsilon\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t')})^{1-\epsilon}} \tag{3.87}$$

$$\begin{aligned} Z_2 &= \int q_{(t)}(\boldsymbol{\mu},\boldsymbol{\Lambda})^\epsilon q_{(t')}(\boldsymbol{\mu},\boldsymbol{\Lambda})^{1-\epsilon}\,d\boldsymbol{\mu}\,d\boldsymbol{\Lambda} \\ &= \prod_{j=1}^{J}\frac{\mathcal{Z}_{\mathcal{NW}}(v_{j(t+1)},a_{j(t+1)},\mathbf{B}_{j(t+1)})}{\mathcal{Z}_{\mathcal{NW}}(v_{j(t)},a_{j(t)},\mathbf{B}_{j(t)})^\epsilon\mathcal{Z}_{\mathcal{NW}}(v_{j(t')},a_{j(t')},\mathbf{B}_{j(t')})^{1-\epsilon}}\ . \end{aligned} \tag{3.88}$$

Finally the parameters of $q_{(t+1)}(\mathbf{z}_n)$ are set as

$$\gamma_{nj(t+1)} = Z_3^{-1}\gamma_{nj(t)}^\epsilon\gamma_{nj(t')}^{1-\epsilon}\ , \qquad \text{where} \qquad Z_3 = \sum_{k=1}^{J}\gamma_{nk(t)}^\epsilon\gamma_{nk(t')}^{1-\epsilon} \tag{3.89}$$

To keep $q_{(t+1)}$ as a normalized distribution, a division by Z_1, Z_2, and Z_3 has to be made; therefore their logs are added to $\ln s_{(t+1)}$ with

$$\ln s_{(t+1)} = \epsilon\ln s_{(t)} + (1-\epsilon)\ln s_{(t')} + \ln Z_1 + \ln Z_2 + \ln Z_3\ . \tag{3.90}$$

3.6 Minimizing over a factor graph

For full multivariate mixtures we have chosen to work with Dirichlets (for the mixing weights) and Normal-Wisharts (for the component parameters) as approximating distributions. We let the factor approximations $\tilde{f}_n$ be of exactly the same form,

$$\tilde{f}_n(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \tilde{s}_n \prod_{j=1}^{J} \exp\Big\{ (\tilde{\delta}_{nj} - 1) \ln \pi_j - \frac{1}{2}\tilde{v}_{nj}\boldsymbol{\mu}_j^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j + \tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}^\top \boldsymbol{\Lambda}_j \boldsymbol{\mu}_j$$
$$- \mathsf{tr}[(\tilde{\mathbf{B}}_{nj} + \frac{1}{2}\tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}\tilde{\mathbf{m}}_{nj}^\top)\boldsymbol{\Lambda}_j] + (\tilde{a}_{nj} - \frac{d}{2}) \ln |\boldsymbol{\Lambda}_j| \Big\} \ . \tag{3.91}$$

This choice of $\tilde{f}_n$ also takes the same form as the prior. The message passing algorithm is:

- Start by initializing, for $n = 1, \ldots, N$, and all components j,

$$\begin{aligned} \tilde{\delta}_{nj} &= 1 & \tilde{\mathbf{m}}_{nj} &= \mathbf{0} & \tilde{a}_{nj} &= \frac{d}{2} \\ \tilde{v}_{nj} &= 0 & \tilde{\mathbf{B}}_{nj} &= \mathbf{0} & \tilde{s}_n &= 1 \ . \end{aligned} \tag{3.92}$$

  so that all the factor approximations are one. Initialize the prior, for all j, as

$$\begin{aligned} \tilde{\delta}_{0j} &= \delta_{0j} & \tilde{\mathbf{B}}_{0j} &= \mathbf{B}_{0j} \\ \tilde{v}_{0j} &= \tilde{v}_{0j} & \tilde{a}_{0j} &= a_{0j} \\ \tilde{\mathbf{m}}_{0j} &= \mathbf{m}_{0j} & \tilde{s}_0 &= \frac{1}{\mathcal{Z}_{\mathcal{D}}(\tilde{\boldsymbol{\delta}}_0)} \prod_{j=1}^{J} \frac{1}{\mathcal{Z}_{\mathcal{NW}}(\tilde{v}_{0j}, \tilde{a}_{0j}, \tilde{\mathbf{B}}_{0j})} \ . \end{aligned} \tag{3.93}$$

- Repeat until all $\tilde{f}_n$ converge:

  1. Pick a factor n. This can be done by looping over random permutations of $1, \cdots, N$.
  2. Compute the 'old' approximation $q^{\backslash n}(\boldsymbol{\pi})q^{\backslash n}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, with parameters indexed by an additional 'o', from reversing equations (3.101) to (3.105), hence subtracting the factor contributions from the present approximation,

$$v_{\mathrm{o}j} = v_j - \tilde{v}_{nj} \tag{3.94}$$

$$\mathbf{m}_{\mathrm{o}j} = \frac{v_j \mathbf{m}_j - \tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}}{v_{\mathrm{o}j}} \tag{3.95}$$

$$\mathbf{B}_{\mathrm{o}j} = \mathbf{B}_j - \tilde{\mathbf{B}}_j + \frac{1}{2}v_j \mathbf{m}_j \mathbf{m}_j^\top - \frac{1}{2}v_{\mathrm{o}j}\mathbf{m}_{\mathrm{o}j}\mathbf{m}_{\mathrm{o}j}^\top - \frac{1}{2}\tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}\tilde{\mathbf{m}}_{nj}^\top \tag{3.96}$$

$$a_{\mathrm{o}j} = a_j - \tilde{a}_{nj} + \frac{d}{2} \tag{3.97}$$

$$\delta_{\mathrm{o}j} = \delta_j - \tilde{\delta}_{nj} + 1 \ . \tag{3.98}$$

     If the cavity distribution $q^{\backslash n}$ is not proper (normalizable), a robust heuristic would be to skip the update and continue with the next factor in step 1. Figure 3.5 shows the effect of these update skips.
  3. Again let S, and for all j, δ_j, v_j, $_j$, a_j, $\mathbf{B}_j$ be the parameters of $q(\boldsymbol{\pi})^{\mathrm{new}}q(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{\mathrm{new}}$ that minimizes (2.84), in this particular case

$$Sq(\boldsymbol{\pi})^{\mathrm{new}}q(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{\mathrm{new}} = \underset{s'q(\boldsymbol{\pi})q(\boldsymbol{\mu},\boldsymbol{\Lambda})}{\arg\min} \ D_\alpha\big(q^{\backslash n}(\boldsymbol{\pi})q^{\backslash n}(\boldsymbol{\mu}, \boldsymbol{\Lambda})f_n(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ \big\| \ s'q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\big) \ . \tag{3.99}$$

The new approximation can be found with the methods discussed in sections 3.3, 3.4 and 3.5. Depending on the value of α, latent variables may need to be added to the joint distribution for tractability, and $q(\mathbf{z}_n)$ added to the approximation. As the parameters γ_{nj} of $q(\mathbf{z}_n)$ are related to a specific term f_n and its approximation $\tilde{f}_n$, it is not necessary to keep track of any 'leave-one-out' distributions for the latent variables, and they are merely used as a means to an end.

After solving for the new approximation, we have to set the factor contribution, and hence

$$\begin{aligned}
\tilde{f}_n &= S\frac{q(\boldsymbol{\pi})^{\text{new}}q(\boldsymbol{\mu},\boldsymbol{\Lambda})^{\text{new}}}{q^{\setminus n}(\boldsymbol{\pi})q^{\setminus n}(\boldsymbol{\mu},\boldsymbol{\Lambda})} \\
&= S\frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{\text{o}})}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta})}\prod_{j=1}^{J}\frac{\mathcal{Z}_{\mathcal{NW}}(v_{\text{o}j},a_{\text{o}j},\mathbf{B}_{\text{o}j})}{\mathcal{Z}_{\mathcal{NW}}(v_j,a_j,\mathbf{B}_j)} \\
&\quad \exp\Big\{(\delta_j-\delta_{\text{o}j})\ln\pi_j-\frac{1}{2}(v_j-v_{\text{o}j})\boldsymbol{\mu}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j+(v_j\mathbf{m}_j-v_{\text{o}j}\mathbf{m}_{\text{o}j})^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j \\
&\quad -\mathsf{tr}[(\mathbf{B}_j-\mathbf{B}_{\text{o}j}+\frac{1}{2}v_j\mathbf{m}_j\mathbf{m}_j^\top-\frac{1}{2}v_{\text{o}j}\mathbf{m}_{\text{o}j}\mathbf{m}_{\text{o}j}^\top)\boldsymbol{\Lambda}_j]+(a_j-a_{\text{o}j})\ln|\boldsymbol{\Lambda}_j|\Big\}\ . \qquad (3.100)
\end{aligned}$$

According to factor definition (3.91), we get the following substitutions for the different parameter contributions,

$$\tilde{v}_{nj} = v_j - v_{\text{o}j} \qquad (3.101)$$

$$\tilde{\mathbf{m}}_{nj} = \frac{v_j\mathbf{m}_j - v_{\text{o}j}\mathbf{m}_{\text{o}j}}{\tilde{v}_{nj}} \qquad (3.102)$$

$$\tilde{\mathbf{B}}_j = \mathbf{B}_j - \mathbf{B}_{\text{o}j} + \frac{1}{2}v_j\mathbf{m}_j\mathbf{m}_j^\top - \frac{1}{2}v_{\text{o}j}\mathbf{m}_{\text{o}j}\mathbf{m}_{\text{o}j}^\top - \frac{1}{2}\tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}\tilde{\mathbf{m}}_{nj}^\top \qquad (3.103)$$

$$\tilde{a}_{nj} = a_j - a_{\text{o}j} + \frac{d}{2} \qquad (3.104)$$

$$\tilde{\delta}_{nj} = \delta_j - \delta_{\text{o}j} + 1 \qquad (3.105)$$

$$\tilde{s}_n = S\frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{\text{o}})}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta})}\prod_{j=1}^{J}\frac{\mathcal{Z}_{\mathcal{NW}}(v_{\text{o}j},a_{\text{o}j},\mathbf{B}_{\text{o}j})}{\mathcal{Z}_{\mathcal{NW}}(v_j,a_j,\mathbf{B}_j)}\ . \qquad (3.106)$$

- Finally we determine the approximation to the evidence $p(\mathbf{x})$ with

$$\begin{aligned}
p(\mathbf{x}) &\approx \int\prod_{n=0}^{N}\tilde{f}_n(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})\,d\boldsymbol{\pi}\,d\boldsymbol{\mu}\,d\boldsymbol{\Lambda} \\
&= \Big(\prod_{n=0}^{N}\tilde{s}_n\Big)\int\prod_{j=1}^{J}\pi_j^{\sum_{n=0}^{N}\tilde{\delta}_{nj}-1}e^{-\frac{1}{2}[\sum_{n=0}^{N}\tilde{v}_{nj}]\boldsymbol{\mu}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j+[\sum_{n=0}^{N}\tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}^\top]\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j} \\
&\quad\times e^{-\mathsf{tr}[\sum_{n=0}^{N}(\tilde{\mathbf{B}}_{nj}+\frac{1}{2}\tilde{v}_{nj}\tilde{\mathbf{m}}_{nj}\tilde{\mathbf{m}}_{nj}^\top)\boldsymbol{\Lambda}_j]+\sum_{n=0}^{N}(\tilde{a}_{nj}-\frac{d}{2})\ln|\boldsymbol{\Lambda}_j|}\,d\boldsymbol{\pi}\,d\boldsymbol{\mu}\,d\boldsymbol{\Lambda} \\
&= \Big(\prod_{n=0}^{N}\tilde{s}_n\Big)\int\prod_{j=1}^{J}\pi_j^{\delta_j-1}e^{-\frac{1}{2}v_j\boldsymbol{\mu}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j+v_j\mathbf{m}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\mu}_j} \\
&\quad\times e^{-\mathsf{tr}[(\mathbf{B}_j+\frac{1}{2}v_j\mathbf{m}_j\mathbf{m}_j^\top)\boldsymbol{\Lambda}_j]+(a_j-\frac{d}{2})\ln|\boldsymbol{\Lambda}_j|}\,d\boldsymbol{\pi}\,d\boldsymbol{\mu}\,d\boldsymbol{\Lambda} \\
&= \Big(\prod_{n=0}^{N}\tilde{s}_n\Big)\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta})\prod_{j=1}^{J}\mathcal{Z}_{\mathcal{NW}}(v_j,a_j,\mathbf{B}_j)\ . \qquad (3.107)
\end{aligned}$$

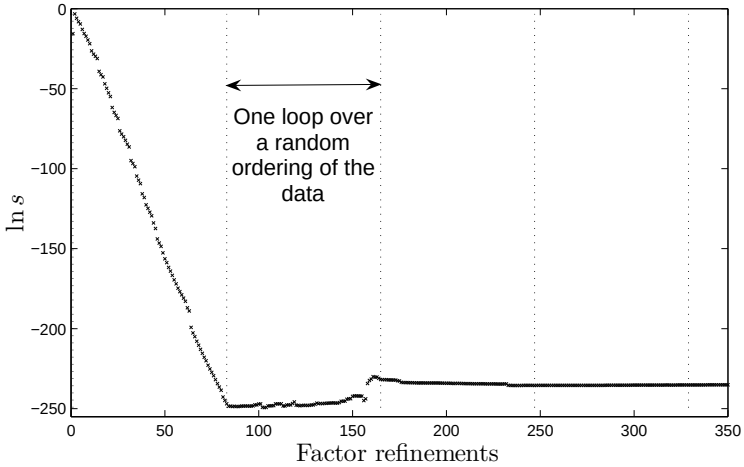


FIGURE 3.2: For the **galaxy** data set, this plot shows the progress of the log marginal likelihood estimate over time. The first loop over $N = 82$ data points is the 'assumed density filtering' (ADF) loop, where observations are included one by one in the approximation. Further loops over random permutations of the data cause the approximation to stabilize to $\ln s = -232.4$. With three components, the prior parameter settings were $\delta_{0j} = 1$, $\mathbf{m}_{0j} = 0$, $v_{0j} = 10^{-2}$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = 0.11$.

Starting with another prior contribution

In the case of mixtures discussed here, a symmetric prior (i.e. the prior is identical for each term in the product over j) leaves the algorithm in a stationary point making no real progress, as all responsibilities r_{nj} remain equal. We may wish to break symmetry by, for example, starting with a random factor initialization. Symmetry is broken here by starting with the 'wrong' prior, and correcting the prior factor contributions to the true prior after a loop over the data. After removing the (wrong) prior's contribution to the approximation, we find the new approximation's parameters by including the (true) prior with

$$v_j = v_{\mathrm{o}j} + v_{0j} \tag{3.108}$$

$$\mathbf{m}_j = \frac{v_{\mathrm{o}j}\mathbf{m}_{\mathrm{o}j} + v_{0j}\mathbf{m}_{0j}}{v_j} \tag{3.109}$$

$$\mathbf{B}_j = \mathbf{B}_{\mathrm{o}j} + \mathbf{B}_{0j} + \frac{1}{2}v_{\mathrm{o}j}\mathbf{m}_{\mathrm{o}j}\mathbf{m}_{\mathrm{o}j}^\top + \frac{1}{2}v_{0j}\mathbf{m}_{0j}\mathbf{m}_{0j}^\top - \frac{1}{2}v_j\mathbf{m}_j\mathbf{m}_j^\top \tag{3.110}$$

$$a_j = a_{\mathrm{o}j} + a_{0j} - \frac{d}{2} \tag{3.111}$$

$$\delta_j = \delta_{\mathrm{o}j} + \delta_{0j} - 1 \ , \tag{3.112}$$

and hence the prior contributions are corrected to $\tilde{v}_{0j} = v_{0j}$, $\tilde{\mathbf{m}}_{0j} = \mathbf{m}_{0j}, \tilde{\mathbf{B}}_{0j} = \mathbf{B}_{0j}, \tilde{a}_{0j} = a_{0j}, \tilde{\delta}_{0j} = \delta_{0j}$. Lastly, the prior's contribution to the scale or approximate marginal likelihood is

$$\tilde{s}_0 = \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)} \prod_{j=1}^{J} \frac{1}{\mathcal{Z}_{\mathcal{NW}}(v_{0j}, a_{0j}, \mathbf{B}_{0j})} \ . \tag{3.113}$$

3.7 Experimental results

This section examines the approximate predictive distributions and log marginal likelihoods found by minimizing the VB, EP, and $\alpha = \frac{1}{2}$ energy (objective) functions over a factor graph, and the results are compared to the 'gold standard' predictions and marginal likelihoods obtained from thermodynamic integration through parallel tempering. Parallel tempering is a Markov chain Monte Carlo method, discussed in greater detail in chapter 4.

Two methods for approximate inference that we discussed in chapter 1 are not included in the comparison. The maximum a posteriori estimate is omitted, for it doesn't provide an estimate of the probability mass and cannot be used for model selection. A Laplace approximation can be used for model selection, as was done with *flat* priors by Roberts et al. (1998). It is excluded as the Hessian matrix couples the component means and variances into a joint covariance matrix of a Gaussian $q(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$, making the approximate predictive distribution intractable.

3.7.1 A toy example

Before turning to larger real-life data sets, we illustrate the difference between EP and VB on a toy example: EP generally gives better predictive distributions than VB on small data sets. As more data is observed, the scale of this improvement is expected to decrease.

Here a data set with $N = 7$ examples were generated from a two-component mixture of Gaussians, and by default we therefore know the correct model class. We have used prior parameter settings $\delta_{0j} = 1$, $\mathbf{m}_{0j} = 0$, $v_{0j} = 10^{-2}$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = 0.11$.

In figures 3.3(a) and 3.3(c) we show a marked difference between the VB and EP predictive distributions on this small data set, both with an incorrect and correct choice of J. Notice that EP gives a broader estimate of $q(\boldsymbol{\theta})$, and hence also broader predictive densities. This difference is most noticeable around the peaks of the distributions, with EP generally giving a closer fit.

As a next illustration the size of the data set was doubled to $N = 14$ by *duplicating* each example. We expect less of a difference in predictive density as the posterior has more concentrated (peaked) modes when there is a greater abundance of data. Figures 3.3(b) and 3.3(d) show a smaller—or even indistinguishable—discrepancy between the EP and VB predictive distributions. In figure 3.3(b) we can still mark EP's slight improvement around the peaks of the density.

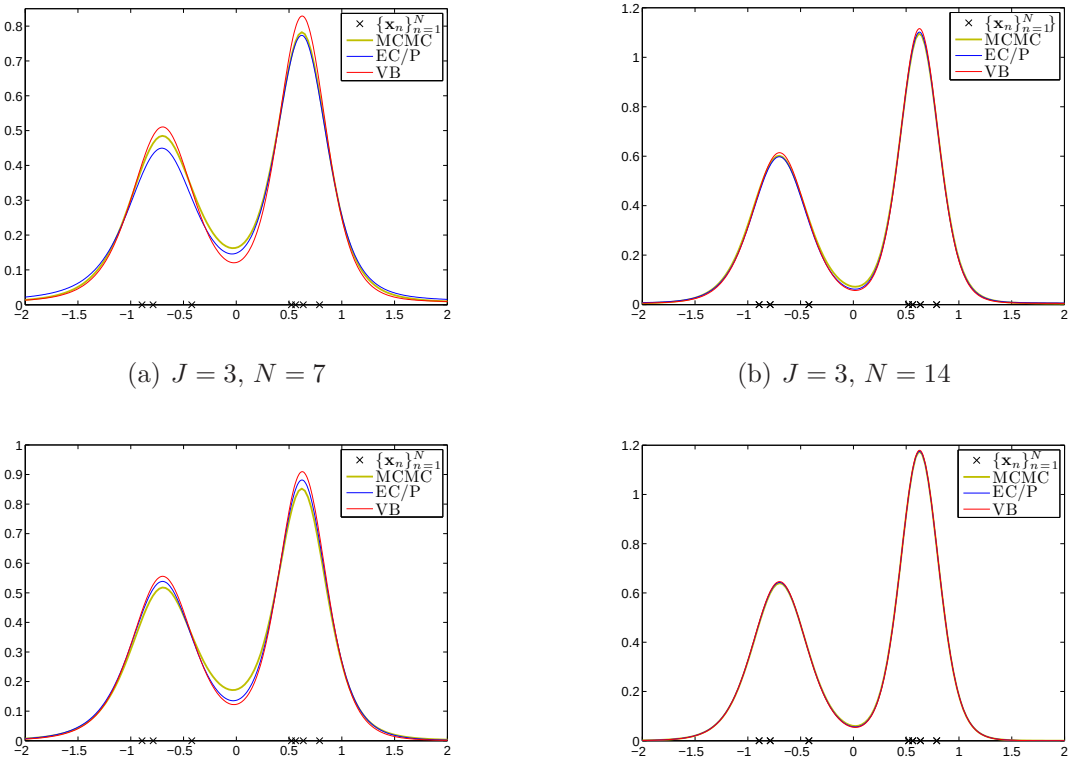
EP gives a better approximation than VB to the log marginal likelihood. This can be expected from the nature of the EP updates; also bear in mind that VB provides a lower bound to the true $\ln Z$. This is illustrated in figure 3.4 for each of the problems in figure 3.3.

3.7.2 Experimental observations

A number of observations from the experiments presented here will confirm the theoretic results given in chapter 2:

- The log marginal likelihood estimates increase with α. The respective objective functions are continuously related through parameter α (see for example equation (2.107)) and therefore we expect related local minima for many data sets. This relation can be observed, with the evaluation of $\ln s$ increasing with α over a local minimum.

- The number of local solutions is influenced by the width of the prior distribution, with more local minima arising in broader priors.



(a) $J = 3$, $N = 7$

(b) $J = 3$, $N = 14$

(c) $J = 2$, $N = 7$

(d) $J = 2$, $N = 14$

FIGURE 3.3: The predictive densities $p(x_{\text{new}}|\mathbf{x}, \mathcal{M}_J)$ given by expectation consistent inference (the expectation propagation algorithm) and variational Bayes. In figures 3.3(a) and 3.3(c) a two-component toy data set was used. For figures 3.3(b) and 3.3(d) the data set size was doubled by duplicating each example. With less data EC/P gives a predictive density that is closer to the truth than VB. With increasing data set size (and hence also more sharply peaked posteriors) this difference becomes less marked to almost indistinguishable.

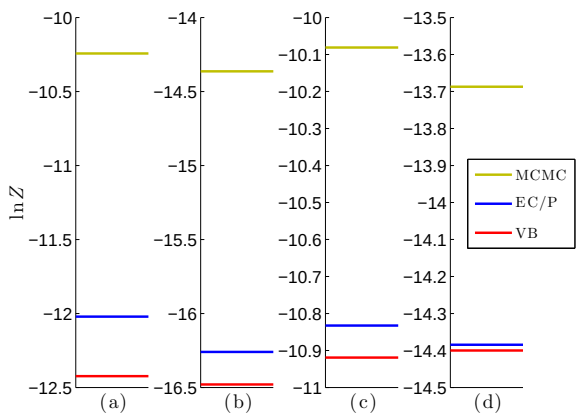


FIGURE 3.4: The log marginal likelihood estimates for the problems shown in figures 3.3(a) to 3.3(d).

- The EP fixed points are not unique, and the fixed point depends on both the initialization and the random order in which factor refinements take place. Both these questions were posed by Minka at the end of his thesis (Minka, 2001a).
- The growth of $\ln s$, as a function of model size, gives a characteristic 'Ockham hill', where the 'peak' of the hill indicates the model with highest approximate $\ln p(\mathbf{x})$. This graph can be used for model comparison or selection, as its form closely matches the MCMC evaluation of $\ln p(\mathbf{x})$.
- The discrepancy between $\ln s$ and the true $\ln p(\mathbf{x})$ grows as the model size is increased. This is mainly due to mixtures being invariant under component relabelling, with the number of permutations increasing as $J!$, with J being the number of components. The true log marginal likelihood will take all such permutations—mostly giving different posterior modes—into account. Provided the modes are well separated, the approximate distribution $q(\boldsymbol{\theta})$ is only able to capture a single.
- As component relabelling gives the same predictive distribution under a symmetric prior, we do not expect a pronounced difference between the true predictive distribution and the predictive distribution obtained from an average over a single-mode $q(\boldsymbol{\theta})$.
- the EC/EP approximation gives a predictive distribution that is closer to the truth than that given by VB.

The message passing algorithm was run for $\alpha = 0$ (VB) , $\alpha = \frac{1}{2}$ (Hellinger distance), and $\alpha = 1$ (EC/EP). The data sets under investigation have been well studied, e.g. by Richardson & Green (1997) for a reversible jump MCMC, and by Corduneanu & Bishop (2001) for variational Bayesian model selection:

galaxy. The galaxy data set contains the velocities (in 1000s of km/second) of 82 galaxies, diverging from our own, in the Corona Borealis region. Multimodal velocity densities provides evidence for clustering of the galaxies into superclusters which are surrounded by large voids, with clusters corresponding to modes in the velocity density (Roeder (1990) provides more detail).

acidity. The acidity data set contains the log measured acid neutralizing capacity indices for 155 lakes in North-central Wisconsin (USA). Identifying groups of lakes at different environmental risk can be useful in determining if any characteristics of a lake can be used to predict higher acidification.

enzyme. The enzyme data set contains enzymatic activity measurements, for an enzyme involved in the metabolism of carcinogenic substances, taken from 245 unrelated individuals. Of interest here is the identification of subgroups of slow or fast metabolisers as a marker of genetic polymorphism in a general population.

old faithful. The data set used contains 222 observations from the Old Faithful Geyser in the Yellowstone National Park. Two measurements—the duration of an eruption, rounded to the nearest 0.1 minutes, and the waiting time to the next eruption, rounded to the nearest minute—constitute a single observation. Although not the largest or most regular, it is the most frequent of the big geysers in the park. The geyser's name comes from the consistency (and predictability) of its eruptions; it was named by the Washburn Expedition in 1870, and is presently still erupting with the same regularity.

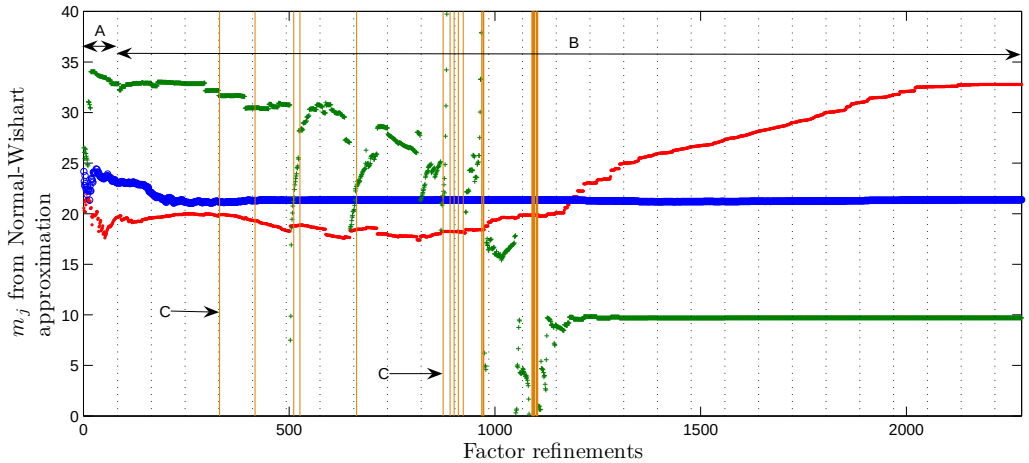


FIGURE 3.5: The convergence need not be fast; this plot is for the better **galaxy** approximation on the *right* in figure 3.7. With three components, the prior parameter settings were $\delta_{0j} = 1$, $\mathbf{m}_{0j} = 0$, $v_{0j} = 10^{-2}$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = 0.11$. The figure also indicates: A, the ADF loop; B, the refinement loops; C, cases where a factor approximation was skipped because constraint-violating parameters (e.g. a negative variance) were recovered for the cavity (leave-one-out) distributions $q^{\setminus n}(\boldsymbol{\theta})$. The dotted lines in each case show the end of one loop of factor refinements.

Discussion

There are a few algorithmic details that are worth considering, and figures 3.2 and 3.5 aim to illucidate the message passing scheme over time.

The progress of $\ln s$, from an algorithmic point of view, is shown in figure 3.2. The algorithm starts with $\ln s$ being equal to the prior distributions's normalizing constant, and it decreases as more factors are included in the ADF loop. The following loops over factors all involve refinements. The figure is given for EP, with $\ln s$ both increasing and decreasing over the refinement loops. In contrast the refinements of VB should, by virtue of the EM algorithm used, give a monotonic increase of $\ln s$. Section 2.9.3 gives greater detail to this statement.

In the practical implementation presented here, a factor approximation $\tilde{f}_n$'s update is skipped if 'illegal' parameters are recovered for the 'leave-one-out' distribution $q^{\setminus n}(\boldsymbol{\theta})$, leaving $q^{\setminus n}(\boldsymbol{\theta})$ unnormalizable. In that case we choose to move on to the next factor, hoping that the problem will have self-alleviated when we return to the particular factor n in the next refinement loop. Figure 3.5's purpose is twofold: it serves as an illustration that the convergence times of EP need not be fast, and it also shows that convergence is sometimes possible when updates are skipped.

EP is a single loop algorithm that aims to minimize a free energy. It comes with no guarantee of convergence, and it was often found that some parameters in figures similar to figure 3.5 (with a larger number of mixture components) never reached a stable solution. A particular example can be seen for a six component mixture under a narrow prior in figure 3.10. More sophisticated double loop algorithms can be implemented to minimize the free energy, and the interested reader is referred to Opper & Winther (2005a)'s detailed discussion on Expectation Consistent (EC) approximate inference (EP can be seen as a particular algorithm minimizing the EC free energy). Alternatively, one may argue, as is done by Minka (2001c), that the approximating distribution is probably not a good choice if EP doesn't converge.

3.7.3 The predictive distribution

For a specific model $\mathcal{M}$, in this case indexed by the number J of Gaussians, the predictive distribution can be approximated by using $q(\boldsymbol{\theta})$ as an approximation to the posterior $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})$. After running the generic message passing scheme, be it for VB or EP or some other divergence measure that was minimized, we have a set of parameters $\boldsymbol{\delta}$, $\{\mathbf{m}_j, v_j, a_j, \mathbf{B}_j\}_{j=1}^J$ governing the shape of the approximation. The approximate predictive distribution is determined by the integral, with shorthands $\boldsymbol{\mu} = \{\boldsymbol{\mu}_j\}_{j=1}^J$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_j\}_{j=1}^J$,

$$p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}) = \int p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})\, d\boldsymbol{\theta} \tag{3.114}$$

$$\approx \int p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta}, \mathcal{M}) q(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

$$= \int \sum_{j=1}^J \pi_j \mathcal{N}(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1}) \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}) \prod_{j=1}^J \mathcal{NW}(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j|\mathbf{m}_j, v_j, a_j, \mathbf{B}_j)\, d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda}$$

$$= \sum_{j=1}^J \frac{\delta_j}{\sum_{k=1}^J \delta_k} \mathcal{T}\left(\mathbf{x}_{\text{new}} \,\middle|\, \mathbf{m}_j,\ \frac{v_j+1}{v_j} \frac{2\mathbf{B}_j}{2a_j - d + 1},\ 2a_j - d + 1\right). \tag{3.115}$$

The exact details of the simplification of an integral similar to the one above follows in appendix A.6.

The predictive distributions for each of the one dimensional data sets in question can be found in figures 3.6 and 3.7. The final predictive distribution strongly depends on whether or not a local minimum in the objective function has been found, as is clear from figure 3.7.

Through the eyes of the predictive distribution, we are averaging over many *equivalent* modes of the posterior distribution in (3.114), as the likelihood is invariant under permutations of the mixture component labels. If the unimodal approximation $q(\boldsymbol{\theta})$ fits one of these modes, we can still expect the predictive distribution to match the truth. To illustrate how much the predictive distribution from (3.115) differs from the true predictive distribution, the figures show the average in (3.114) obtained through a MCMC method. We shall leave the exact details of MCMC used (here the zero-temperature samples from a parallel tempered chain) to chapter 4.

It is interesting to note the difference in predictive distributions given by VB and EP/C, which we compare to the true predictive distribution in figure 3.8. The 'truth' is taken as average over 10,000 samples from a parallel tempered chain (see chapter 4). Expectation consistent inference works on the principle of finding a $q(\boldsymbol{\theta})$ that matches the moments of the predictive distributions for all $\mathbf{x}_n$ under the cavity distributions $q^{\backslash n}(\boldsymbol{\theta})$. We can therefore expect a better predictive distribution than that given by VB, which is based on a lower bound to the posterior.

3.7.4 Ockham hills and the approximate log marginal likelihood

Some insight into the difference between various divergence measures can be gained by examining their Ockham hills. The term 'Ockham hill' is here used loosely as a plot of the log marginal likelihood for a varying number of models. (Rasmussen & Ghahramani (2001) give an insightful account of Ockham's razor, showing that 'plateaus', where the log marginal likelihood flattens with increasing complexity, are also possible.) Each plot is complemented with an estimate of $\ln p(\mathbf{x}|\mathcal{M}_J)$ for different model sizes J. In each case the estimate was obtained form an average over ten MCMC simulations, using 10,000 samples, with two standard deviation error bars also being shown. The exact details of the MCMC method used is presented in chapter 4, where section 4.5 in particular relates to these results.

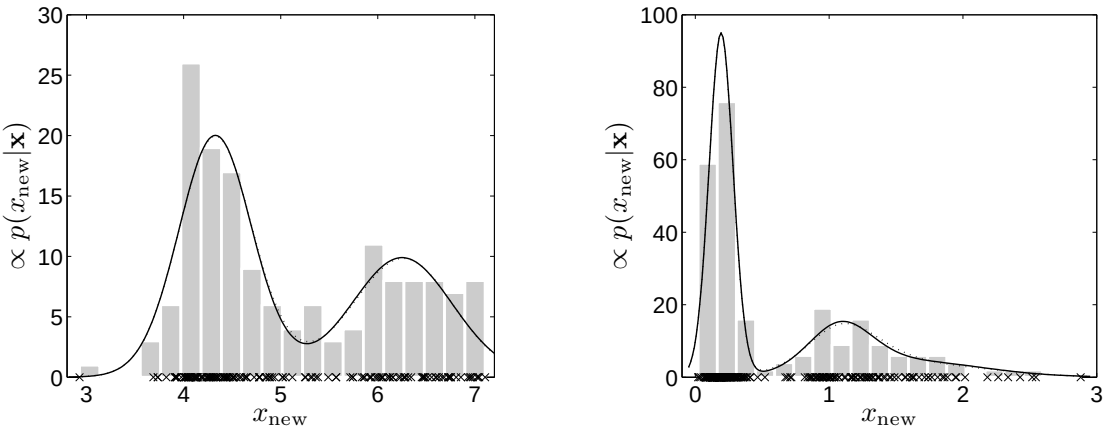


FIGURE 3.6: The predictive distribution $p(x_{\text{new}}|\mathbf{x}, \mathcal{M}_J)$ for the **acidity** data set (*left*) and the **enzyme** data set (*right*). The chosen model $\mathcal{M}_J$ was the one giving the highest log marginal likelihood estimate in figures 3.11 and 3.12, with $J = 2$ and $J = 3$ components respectively. The two log marginal likelihood approximations were $\ln s = -200.3$ (*left*) and $\ln s = -82.4$ (*right*). The prior parameter settings were $\delta_{0j} = 1$, $\mathbf{m}_{0j} = 0$, $v_{0j} = 10^{-2}$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = 0.11$. The true predictive distribution, obtained from an average over a MCMC sample, is shown with a dotted line.

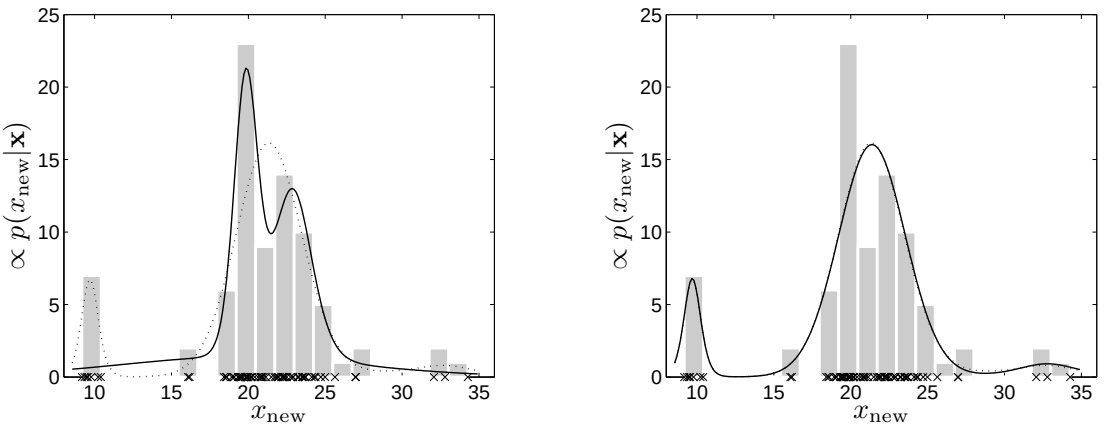


FIGURE 3.7: The predictive distribution $p(x_{\text{new}}|\mathbf{x}, \mathcal{M}_3)$, from two different approximations found by EP for the **galaxy** dataset. This clearly illustrates the non-uniqueness of the EP fixed points. The chosen model was the one giving the highest log marginal likelihood estimate in figure 3.10, for three components. The approximation on the *left* gave $\ln s = -243.8$, whereas the approximation on the *right* gave a much higher $\ln s = -232.4$. The prior parameter settings were $\delta_{0j} = 1$, $\mathbf{m}_{0j} = 0$, $v_{0j} = 10^{-2}$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = 0.11$. The true predictive distribution, obtained from an average over a MCMC sample, is shown with a dotted line.

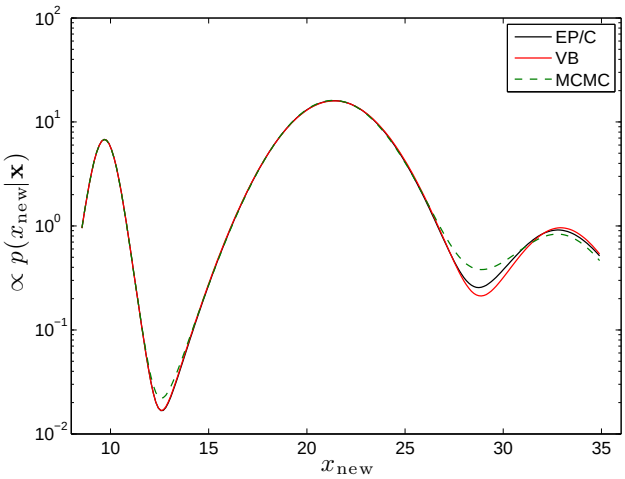


FIGURE 3.8: A comparison between the EP, VB and MCMC log predictive distributions $p(x_{\text{new}}|\mathbf{x}, \mathcal{M}_3)$ for the **galaxy** data set from figure 3.7. With N being large, the approximate predictive distribution from EP is marginally closer to the truth than the approximate predictive distribution given by VB. Especially notice the gain where data is sparse. We've taken the true predictive distribution as an average over a sample of 10,000 points from a parallel tempered Markov chain. The EP and VB approximations with the highest marginal likelihoods were used here, and the parameter settings matched that of figure 3.7.

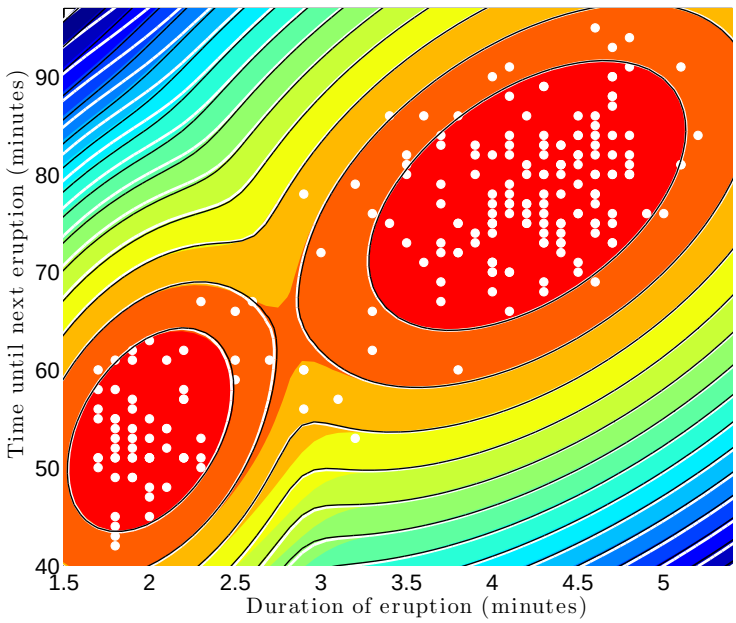


FIGURE 3.9: The log predictive distribution $p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}_2)$ for the **old faithful** data set. The prior parameter settings were $\delta_{0j} = 1$, $\mathbf{m}_{0j} = \mathbf{0}$, $v_{0j} = 10^{-2}$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = [0.11, 0.01; 0.01, 0.11]$. The true predictive distribution, obtained from an average over a parallel tempered MCMC sample, is shown in coloured contours. The EP estimate is shown in white contours, and the VB estimate is overlaid in black contours. The posterior has two sharply peaked modes (as there are two permutations of component labeling), and the difference between the EP and VB predictive densities are virtually indistinguishable under this large data set.

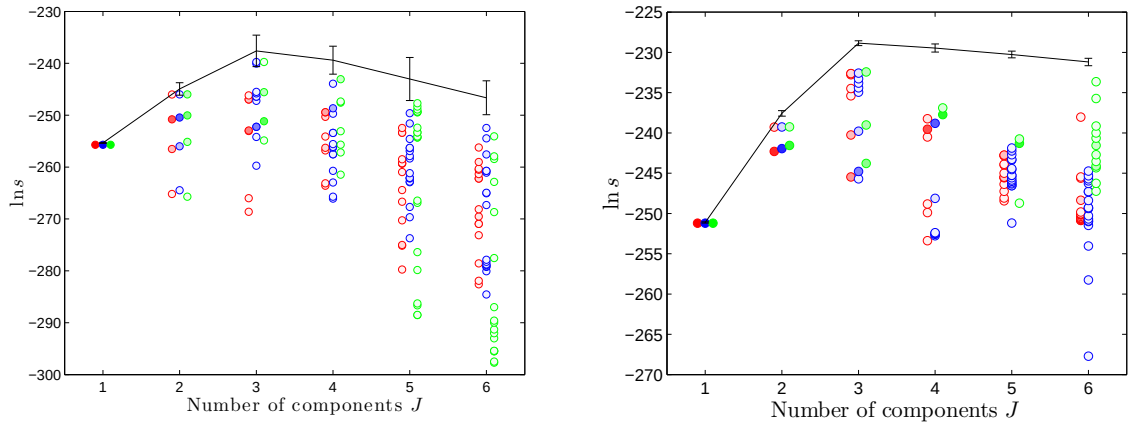


FIGURE 3.10: Ockham hill for the **galaxy** data set, for two different prior settings. For the *left* figure a broad prior with $v_{0j} = 10^{-6}$ was used, for the *right* figure a much narrower prior with $v_{0j} = 10^{-2}$ was used. The other prior hyperparameters were $\delta_{0j} = 1$, $\mathbf{m}_{0j} = 0$, $a_{0j} = 1$ and $\mathbf{B}_{0j} = 0.11$. VB ($\alpha = 0$) is shown in red; $\alpha = \frac{1}{2}$ is shown in blue; EP ($\alpha = 1$) is shown in green. For each model $\mathcal{M}_J$ containing J mixture components, the figure shows twenty runs over different starting values of the means, or different starting 'priors' (later corrected). The colour intensity of the plot corresponds to the frequency of reaching different solutions. Also shown are the values of $\ln p(\mathbf{x}|\mathcal{M}_J)$ found by parallel tempering and thermodynamic integration, averaged for each J over 10 MCMC simulations, with two standard deviation error bars. For the *left* figure with the much broader prior, a generalized version of parallel tempering (see section 4.3) was used.

Model selection and averaging

In the case of VB, the approximation $\ln s$ provides a lower bound to the marginal likelihood $p(\mathbf{x}|\mathcal{M})$, and this quantity is often used for model selection (Beal & Ghahramani, 2003; Bishop & Svensén, 2003; Corduneanu & Bishop, 2001). The model with the largest bound is typically kept, although the bound can also be used for model averaging. In the case of averaging over models $\{\mathcal{M}_i\}$,

$$p(\mathbf{x}_{\text{new}}|\mathbf{x}) = \sum_{\mathcal{M}_i} p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}_i) p(\mathcal{M}_i|\mathbf{x}) \ , \tag{3.116}$$

the approximation given in (3.115) can be used as a substitute for $p(\mathbf{x}_{\text{new}}|\mathbf{x}, \mathcal{M}_i)$. There are usually many local minima in the objective function, and for each model $\mathcal{M}_i$ the resulting model with the largest approximation $\ln s_i$ can be kept. The posterior model distribution $p(\mathcal{M}_i|\mathbf{x}) = p(\mathbf{x}|\mathcal{M}_i)p(\mathcal{M}_i)/p(\mathbf{x})$ can then be approximated with $q(\mathcal{M}_i) \propto s_i p(\mathcal{M}_i)$. Regardless of our choice of divergence measure, poor local minima in the objective function have to avoided in order to obtain meaningful results.

Practical results

Figures 3.10, 3.11 and 3.12 illustrate different values for the approximate evidence for mixture models with a growing number of components for the galaxy, acidity and enzyme data sets. The results for different α-divergences with $\alpha = 0, \frac{1}{2}, 1$ are shown. For VB there are many local maxima in the evidence lower bound, many of which are clearly in some correspondence with the solutions given by $\alpha = \frac{1}{2}$ and EP.

All results obtained here used an initialization of the prior factor to the true prior, *except* for the prior factor mean being initialized to the mean of the data, plus some additive Gaussian

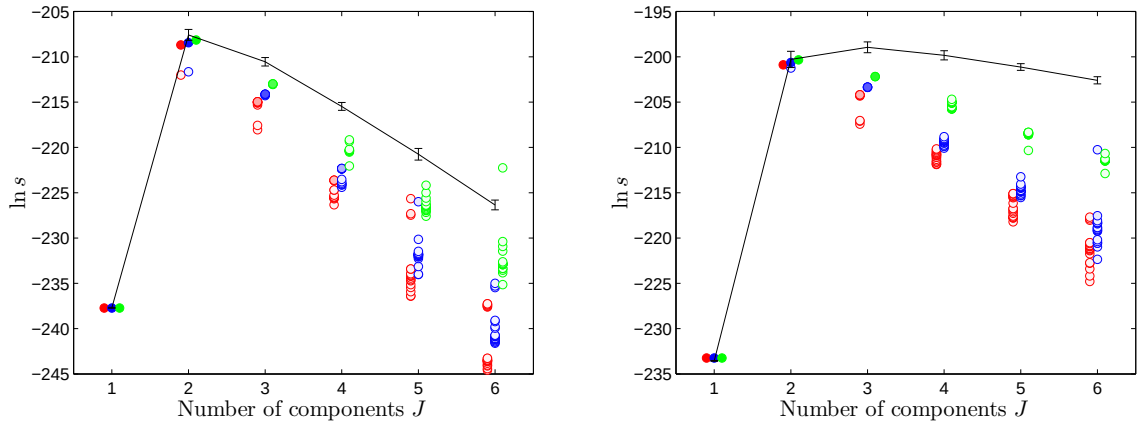


FIGURE 3.11: Ockham hill for the **acidity** dataset, for two different prior settings. For the *left* figure a broad prior with $v_{0j} = 10^{-6}$ was used, for the *right* figure a much narrower prior with $v_{0j} = 10^{-2}$ was used. VB ($\alpha = 0$) is shown in red; $\alpha = \frac{1}{2}$ is shown in blue; EP ($\alpha = 1$) is shown in green. The rest of the experimental setup matched that of figure 3.10. For the *left* figure a generalized version of parallel tempering was again needed to obtain a numerically stable solution.

noise to break symmetry in the following factor updates. The prior factor was corrected to the true prior after the first loop over the data. With these slightly random initializations we find convergence of the message passing scheme to different local minima in the objective function. The random data ordering from the refinement loops may also play a role in convergence to one solution or another. A simple test, by comparing a number of EP runs with the *same* prior factor initialization, showed that all local minima for $J = 3$ in figure 3.10 (narrow prior) could indeed be reached purely based on the random presentation of the factors.

The largest value of $\ln s$ for each model forms the classical 'Ockham hill', with a peak for the optimal model. As models become *less* complex, the hill falls steeply due to a poorer explanation of the data. For *more* complex models the plots show a slower downward trend, as an improvement in data fit is counterbalanced by a penalty from a larger parameter space in Bayesian marginalization. The downward trend for more complex models is even slower when the true log marginal is considered; this is mainly due to the number of modes in the true posterior increasing with the number of components, with an approximation possibly only capturing one of them.

As expected, and as is also visible from the plots, the value for $\ln s$ increases with α, with the difference being greater with increased model size. For all figures, results were obtained by doing factor refinements for a maximum of twenty loops over random orderings of the data set (or factors). This usually proved more than sufficient for convergence. For larger models, EP (and indeed $\alpha = \frac{1}{2}$) may not converge to a stable solution, but iterate around the EP fixed point. This has been observed in practice (Minka, 2001c): when canonical EP does not converge, the reason can be traced back to the approximating family being a poor match to the exact posterior distribution. Intuitively we may think of it this way: EP approximates one of a number of well separated modes in a posterior; when the modes become highly overlapping, as may be the case when using too many mixture components, the approximation may not 'settle down'.

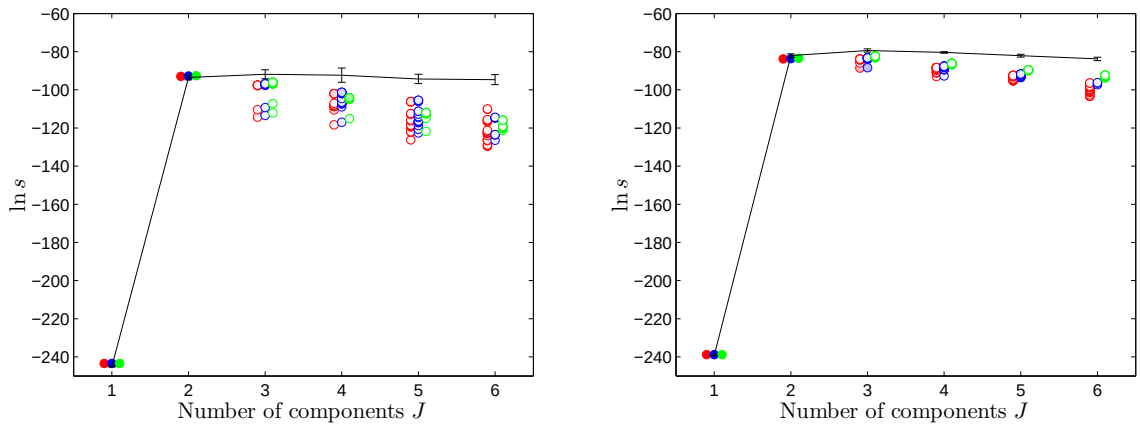


FIGURE 3.12: Ockham hill for the **enzyme** data set, for two different prior settings. For the *left* figure a broad prior with $v_{0j} = 10^{-6}$ was used, for the *right* figure a much narrower prior with $v_{0j} = 10^{-2}$ was used. VB ($\alpha = 0$) is shown in red; $\alpha = \frac{1}{2}$ is shown in blue; EP ($\alpha = 1$) is shown in green. The rest of the experimental setup matched that of figure 3.10.

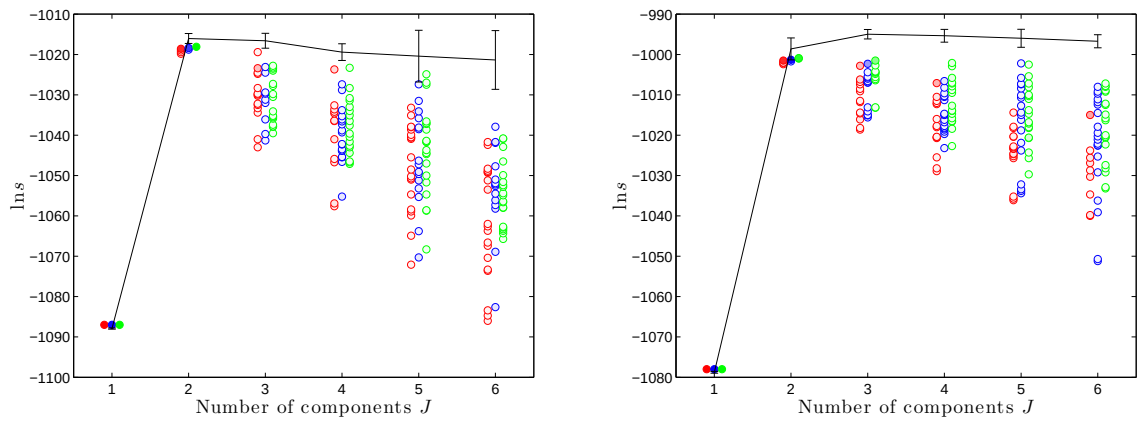


FIGURE 3.13: Ockham hill for the **old faithful** data set, for two different prior settings. For the *left* figure a broad prior with $v_{0j} = 10^{-6}$ was used, for the *right* figure a much narrower prior with $v_{0j} = 10^{-2}$ was used. VB ($\alpha = 0$) is shown in red; $\alpha = \frac{1}{2}$ is shown in blue; EP ($\alpha = 1$) is shown in green. Apart from $\mathbf{B}_{0j} = [0.11, 0.01; 0.01, 0.11]$, the rest of the experimental setup matched that of figure 3.10.

Variational Bayes and symmetry breaking

An observation in section 2.8 was made that under a 'broad enough' prior distribution, the fixed point scheme that is used to minimize the VB objective function contains 'symmetry-breaking' local minima, where a component weight is forced to a near-zero value, and the component in question collapses to a distribution close to its prior. This happened in practice when the ADF loop was implemented with $\alpha = 0$ (VB): as the data points (and hence factors) are included one-by-one into the approximation, it often happens that one of the first inclusions will unnecessarily break symmetry. Some mixture components will be pronounced, while others are allocated a near-zero weight, causing them to never recover their part in the approximation. The ADF loop contrasts with traditional VB, where all the data points are presented together in the same EM algorithm.

The message passing algorithm can be altered in a number of ways to bypass this effect. One approach is to randomly initialize all the factor approximations before any updates take place. Another solution, more coherent with the algorithms presented in this chapter, is to run the ADF loop and possibly the first refinement loop using $\alpha = 1$ (EP). This should give a reasonable starting position, after which the value of α can be flipped back to the desired value, and factor refinements continue as normal. The second solution is the one adopted here, and proved successful in preventing undesired initial model pruning.

3.8 Summary and outlook

Multivariate Gaussian mixtures were put under the microscope in this chapter, and three approximate methods of inference, variational Bayes, expectation propagation, and $\alpha = \frac{1}{2}$ message passing, were compared. Both the predictive densities and marginal likelihoods were compared to the results obtained from parallel tempering, which is discussed in chapter 4. This chapter has mostly been practical, involving the derivation of message passing algorithms for α-divergence. Examples were seen where these algorithms run into trouble, where unnormalizable distributions can be recovered, and updates skipped and possibly recovered from. We have seen how approximate methods can be a useful tool in model selection, and give accurate results if we are interested in the predictive distribution.

The algorithm implemented was a single loop algorithm, and it is not guaranteed to converge. Double loop algorithms, as mentioned in section 2.9.1, can provide a guarantee, but were not implemented for this thesis.

The techniques and derivations presented here can be readily extended to other models that use latent variables, provided that moment-matching is a tractable operation, or that we can analytically write down the partition function or predictive density. We shall here look at a simple extension, and extensions that may need even further approximations:

3.8.1 Hidden Markov models

We can view each component in this chapter's mixture of Gaussians as a particular state. Having chosen state j, observation $\mathbf{x}_n$ is then 'emitted' from the state with probability $p(\mathbf{x}_n|\boldsymbol{\theta}_j)$. Our choice of the particular state is independent of its predecessors, giving a zeroth-order Markov model. We can extend this to a first-order Markov model by choosing the present state to be dependent on the previous state. Let $\mathbf{z}_n$ again be a variable that indicates the unobserved state. The joint probability of a sequence of states and observations—which can be similarly extended

to be higher-order Markov—is

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\pi}) = p(\mathbf{x}_1|\mathbf{z}_1, \boldsymbol{\theta})p(\mathbf{z}_1|\boldsymbol{\pi}) \prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})p(\mathbf{z}_n|\mathbf{z}_{n-1}, \boldsymbol{\pi}) \ . \tag{3.117}$$

In the case of a *hidden Markov model* (Rabiner & Juang, 1986), the states are discrete. Similar to the mixture model, a hidden state can be modeled with a binary latent variable $\mathbf{z}_n \in \{0,1\}^J$ that sums to one. The transition probabilities $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ can be modeled by a transition matrix with entries (i,j) representing $p(z_{nj} = 1|z_{n-1,i} = 1)$, although we can equally represent the joint distribution $p(z_{nj} = 1, z_{n-1,i} = 1)$ with a $J \times J$ matrix $\boldsymbol{\pi}$, and place a prior on the initial state.

With reference to the mixture model addressed in this chapter, we can imagine a temporal component correlating the frequency and length between eruptions of the Old Faithful Geyser (see section 3.7), and the emission distributions $p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})$ may be modeled as Gaussian. Other models are possible: A more traditional view is to let a state emit one of K discrete symbols, which we can again model with a binary variable $\mathbf{x}_n \in \{0,1\}^K$ with entries summing to one. The probabilities $p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})$ can be described by a $J \times K$ symbol emission matrix, with entries (j,k) being $p(x_{nk} = 1|z_{nj} = 1)$.

The joint distribution of a data point and its associated hidden state variables is

$$p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})p(\mathbf{z}_n, \mathbf{z}_{n-1}|\boldsymbol{\pi}) = \prod_{i=1}^{J}\prod_{j=1}^{J}[\pi_{ij}p(\mathbf{x}_n|\boldsymbol{\theta}_j)]^{z_{nj}\times z_{n-1,i}} \ , \tag{3.118}$$

such that $\sum_{i=1}^{J}\sum_{j=1}^{J}\pi_{ij} = 1$. In the simplest example we can choose the prior $p(\boldsymbol{\pi})$ and $q(\boldsymbol{\pi})$ to be Dirichlet. If for example $q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ now has J^2 terms, the partition function and predictive density

$$\begin{aligned} p(\mathbf{x}_n) &= \int \sum_{i=1}^{J}\sum_{j=1}^{J} \pi_{ij}p(\mathbf{x}_n|\boldsymbol{\theta}_j)q(\boldsymbol{\pi})q(\boldsymbol{\theta})\, d\boldsymbol{\pi}\, d\boldsymbol{\theta} \\ &= \frac{1}{\sum_{i=1}^{J}\sum_{j=1}^{J}\delta_{ij}} \sum_{j=1}^{J}\left(\sum_{i=1}^{J}\delta_{ij}\right)p_j(\mathbf{x}_n) \ , \end{aligned} \tag{3.119}$$

is tractable. In first line of (3.119) we have already summed (3.118) over $\mathbf{z}_n$ and $\mathbf{z}_{n-1}$, and as a result get a mixture of distributions. Moment matching is possible, and EP/C updates can be derived.

3.8.2 Latent variable models requiring further approximations

A large number of latent variable models give rise to intractable predictive distributions. A simple example is a *factor analysis*, which models high dimensional data $\mathbf{x}_n$ in terms of a smaller number of latent factors $\mathbf{z}_n$:

$$\mathbf{x}_n = \boldsymbol{\Lambda}\mathbf{z}_n + \boldsymbol{\epsilon}_n \quad \text{and} \quad p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{I}) \ , \quad p(\boldsymbol{\epsilon}_n) = \mathcal{N}(\boldsymbol{\epsilon}_n|\mathbf{0}, \boldsymbol{\Gamma}) \ , \tag{3.120}$$

with the noise covariance $\boldsymbol{\Gamma}$ constrained to be diagonal. We therefore have

$$p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\Lambda}\mathbf{z}_n, \boldsymbol{\Gamma}) \quad \text{and} \quad p(\mathbf{x}_n|\boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Gamma}) \ . \tag{3.121}$$

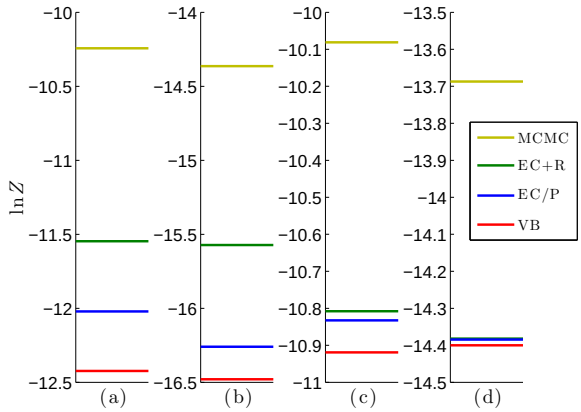


FIGURE 3.14: Following figure 3.4, the log marginal likelihood estimates for the problems shown in figures 3.3(a) to 3.3(d) are shown here. A second order perturbative correction was made to the EC estimate of $\ln Z$, and is indicated by EC+R. There is a clear gain in computing the correction; note in (d) that it is not always big.

A Gaussian prior is typically placed on the rows of $\mathbf{\Lambda}$ (Ghahramani & Beal, 2000), but this renders the derivation of EP updates impossible, as the partition function or predictive distribution

$$p(\mathbf{x}_n) = \int \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Gamma})p(\mathbf{\Lambda})\, d\mathbf{\Lambda} \tag{3.122}$$

is not analytically tractable. Further approximations are therefore necessary, but given the ease in which a Gibbs sampler can be derived for the same problem, it is not immediately clear if a venture into the land of more approximations will be fruitful.

We find ourselves in a similar situation with a *Gaussian linear state space model*, where $\mathbf{z}_n$ is a k-dimensional real valued hidden state variable, and the sequence of $\mathbf{z}_n$s follow a first-order Markov process like equation (3.117). With linear state- and observation equations,

$$\begin{aligned} \mathbf{z}_n &= \mathbf{A}\mathbf{z}_{n-1} + \mathbf{v}_n \ , & p(\mathbf{v}_n) &= \mathcal{N}(\mathbf{v}_n|\mathbf{0}, \mathbf{\Gamma}) \ , \\ \mathbf{x}_n &= \mathbf{B}\mathbf{z}_n + \mathbf{w}_n \ , & p(\mathbf{w}_n) &= \mathcal{N}(\mathbf{w}_n|\mathbf{0}, \mathbf{\Sigma}) \ , \end{aligned} \tag{3.123}$$

and possibly a Gaussian prior on the rows of the observation matrix $\mathbf{B}$ or the state dynamics matrix $\mathbf{A}$, the moments are, very similar to (3.122), analytically intractable. In systems where these matrices are known, or where both $\mathbf{z}_{n-1}$ and $\mathbf{z}_n$ are passed through *known* nonlinear transformations, EP has proved to be a successful engine for approximate inference (Heskes & Zoeter, 2002; Ypma & Heskes, 2003).

3.8.3 Perturbative corrections

Opper (2006) showed how perturbative corrections can be used to improve expectation consistent (EC) approximations. This involves writing the true log partition function (marginal likelihood) as a sum of the EC partition function and a log difference, i.e.

$$\ln Z = \ln Z^{\mathrm{EC}} + \ln R \ , \tag{3.124}$$

where the difference R, with density $q_n(\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})q^{\setminus n}(\boldsymbol{\theta}) \big/ \int p(\mathbf{x}_n|\boldsymbol{\theta}')q^{\setminus n}(\boldsymbol{\theta}')d\boldsymbol{\theta}'$, is

$$R = \int q(\boldsymbol{\theta}) \prod_{n=1}^{N} \left(\frac{q_n(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} \ . \tag{3.125}$$

The difference can be expanded as a series of small parameters, e.g. $\epsilon_n(\boldsymbol{\theta}) = 1 - q_n(\boldsymbol{\theta})/q(\boldsymbol{\theta})$. We therefore have a product $\prod_n(1 - \epsilon_n(\boldsymbol{\theta}))$ in equation (3.125); when this is expanded we can drop higher order terms and compute a tractable correction.

The expansion can be computed up to any order for the mixture model examined in this thesis. We do not present the full derivation and justification here, but show in figure 3.14 the second-order corrections for figure 3.4. The corrections to EC give better log evidence estimates than both EC/P and VB. This is without doubt an exciting area of research to pursue.

Chapter 4

Parallel Tempering

4.1 Introduction

Parallel tempering, or replica exchange, is an efficient method of combining separate Monte Carlo simulations to sample across different modes of a target distribution. As a by-product the normalizing constant of the distribution can also be estimated.

This simulation technique has independently been rediscovered in the 1990s by different authors, and has consequently been referred to by a number of names: the exchange MC-algorithm, the Metropolis-coupled chain algorithm, time-homogeneous parallel annealing, and the multiple Markov chain algorithm (Ferkinghoff-Borg, 2002). Its origins can be traced back to the work of Swendsen & Wang (1986), where a method was introduced where *replicas* of a system of interest were simulated at a series of temperatures, and replicas at adjacent temperatures allowed to exchange partial configuration information. We can also consider parallel tempering as a descendant of the simulated annealing algorithm. Annealing here means that the search for the minimum of some function is conducted from a high to a low temperature, with a temperature parameter gradually being decreased to zero. The use of parallel tempering was initially restricted to problems in statistical physics, but has since found its way into many fields. For an overview of diverse applications of parallel tempering to polymeric systems, proteins and biological molecules, crystalline structures, spin glasses and quantum level systems, see (Earl & Deem, 2005).

4.1.1 Replicas at temperatures

A single Markov chain Monte Carlo simulation may run into difficulties if the target distribution is multi-modal. The chain may get stuck in a local mode, and fail to fully explore other areas of parameter space that have significant probability. One conceptual solution to this problem is to create a series of progressively flatter distributions using some temperature parameter. Systems at higher temperatures, with flatter distributions, should be able to sample from a greater range of parameter space. At lower temperature systems may have precise sampling in local ranges of parameter space, but may become trapped in modes that are difficult to escape from within the run time taken by a typical simulation. We include a temperature parameter through its inverse β, where β ranges from zero to one. With the inverse temperature set to one we can write the posterior distribution—implicitly giving model $\mathcal{M}$, of course—in the usual way,

$$p(\boldsymbol{\theta}|\mathbf{x},\beta) = \frac{1}{\mathcal{Z}(\beta)} p(\mathbf{x}|\boldsymbol{\theta})^{\beta} p(\boldsymbol{\theta}) \ , \tag{4.1}$$

where the partition function is

$$\mathcal{Z}(\beta) = \int p(\mathbf{x}|\boldsymbol{\theta})^{\beta} p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \ . \tag{4.2}$$

The prior is recaptured with an infinite temperature $\beta = 0$, and the log marginal likelihood that we are interested in is given by $p(\mathbf{x}) = \mathcal{Z}(1)$. If the likelihood is flattened with β, the data is effectively gradually introduced from complete absence at $\beta = 0$, to complete presence at $\beta = 1$. The correspondence between (4.1) and a Gibbs distribution comes from defining an 'energy' as $E(\boldsymbol{\theta}) = -\ln p(\mathbf{x}|\boldsymbol{\theta})$, and then

$$p(\boldsymbol{\theta}|\mathbf{x}, \beta) = \frac{1}{\mathcal{Z}(\beta)} \exp\big\{ -\beta E(\boldsymbol{\theta})\big\} p(\boldsymbol{\theta}) \quad \text{and} \quad \mathcal{Z}(\beta) = \int \exp\big\{ -\beta E(\boldsymbol{\theta})\big\}\, dp(\boldsymbol{\theta}) \ . \tag{4.3}$$

We now simulate K replicas of the original system of interest (4.1), each at a different temperature. A set of reciprocal temperatures $\{\beta_k\}_{k=1}^K$ are chosen, and we let the set be ordered as a ladder of increasing inverse temperature distributions with $\beta_k < \beta_{k+1}$. Here we choose $\beta_1 = 0$, corresponding to the prior, and $\beta_K = 1$, corresponding to the posterior distribution.

As the simulation of K replicas instead of just one requires K times the computational effort, there must be some solid reasoning for expending this effort. Firstly, an estimate of the marginal likelihood $\mathcal{Z}(1)$ can be found. Secondly, it has been observed in practice that a parallel tempering simulation is more than $1/K$ times more efficient than a standard, single-temperature Monte Carlo simulation (Earl & Deem, 2005). This is because the replicas allow sampling from low temperature systems, for example the true posterior with $\beta = 1$, to reach regions of parameter space that would not otherwise have been practically accessible had we run a single chain at $\beta = 1$ for K times as long.

The key to parallel tempering is that chains at different temperatures are allowed to exchange complete configurations or states. Through these exchanges low temperature systems access different modes or regions of parameter space via the higher temperature systems.

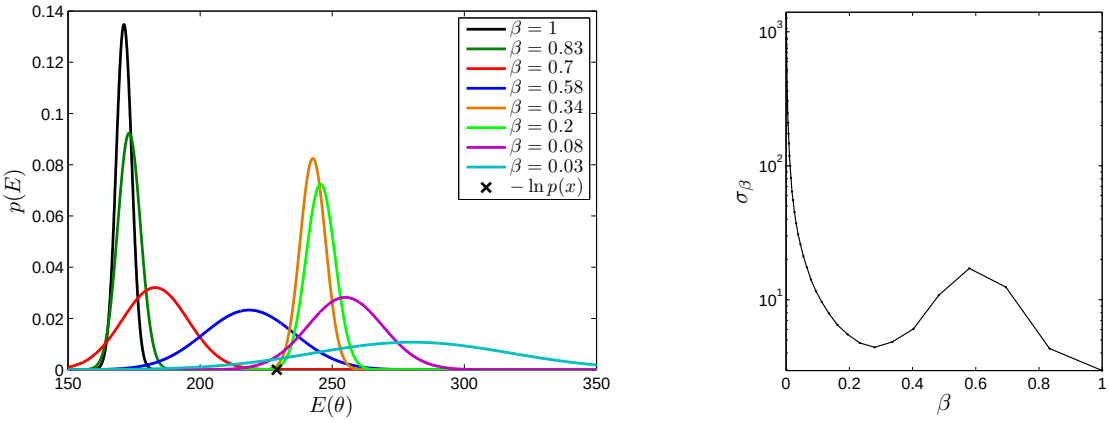
4.1.2 Extended ensembles and replica exchange

As multiple copies of the simulation are run in parallel, each at different temperatures, we therefore have an extended ensemble, where the parameter space is replicated K times to $\{\boldsymbol{\theta}_k\}_{k=1}^K$. With $\beta \in \{\beta_k\}_{k=1}^K$ we run K systems in parallel, and the full target distribution that is being sampled from is

$$p(\{\boldsymbol{\theta}_k\}_{k=1}^K) = \prod_{k=1}^K \frac{1}{\mathcal{Z}(\beta_k)} \exp\Big\{\beta_k \ln p(\mathbf{x}|\boldsymbol{\theta}_k)\Big\} p(\boldsymbol{\theta}_k) \ . \tag{4.4}$$

We run the K chains independently to sample from distributions $p(\boldsymbol{\theta}|\mathbf{x}, \beta_k)$, and add an additional Metropolis Hastings move to swap two βs between chains, or equivalently swap parameters between chains. This is the *replica-exchange* move. Parallel tempering can be done *complementary* to any Monte Carlo method at a single temperature, as long as the exchanges satisfy detailed balance. Having chosen two chains i and j for which we want to swap parameters, let $\{\boldsymbol{\theta}_k\}^{\text{new}}$ be the parameter set with parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ swapped. The acceptance probability for the move is

$$\alpha(\{\boldsymbol{\theta}_k\}^{\text{new}}|\{\boldsymbol{\theta}_k\}) = \min\left(1, \frac{p(\{\boldsymbol{\theta}_k\}^{\text{new}})}{p(\{\boldsymbol{\theta}_k\})}\right) \ . \tag{4.5}$$



(a) The distribution of $E(\boldsymbol{\theta}) = -\ln p(\mathbf{x}|\boldsymbol{\theta})$ under replicas at different temperatures, $p(\boldsymbol{\theta}|\mathbf{x}, \beta)$. These distributions correspond to 'energy histograms', and following (4.7) there should be an overlap between adjacent replicas at different temperatures, so that acceptance of configuration or parameter swaps is allowed for. For interest, the negative log marginal likelihood $-\ln p(\mathbf{x})$ is also indicated on the plot.

(b) The standard deviation σ_β, from (4.9), as a function of β.

FIGURE 4.1: These plots show a selection of reciprocal temperatures for the **galaxy** data set with $J = 3$ components, a prior $v_{0j} = 0.01$, and all other prior parameters following section 3.7. The averages $\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \rangle_{\beta_k}$ for this particular problem and temperature set is illustrated in figure 4.2(a).

As the prior distributions cancel, the ratio between the distributions is

$$\frac{p(\{\boldsymbol{\theta}_k\}^{\text{new}})}{p(\{\boldsymbol{\theta}_k\})} = \frac{\exp\{\sum_{k \neq i,j} \beta_k \ln p(\mathbf{x}|\boldsymbol{\theta}_k) + \beta_i \ln p(\mathbf{x}|\boldsymbol{\theta}_j) + \beta_j \ln p(\mathbf{x}|\boldsymbol{\theta}_i)\}}{\exp\{\sum_{k \neq i,j} \beta_k \ln p(\mathbf{x}|\boldsymbol{\theta}_k) + \beta_i \ln p(\mathbf{x}|\boldsymbol{\theta}_i) + \beta_j \ln p(\mathbf{x}|\boldsymbol{\theta}_j)\}} \ , \tag{4.6}$$

which simplifies as

$$\frac{p(\{\boldsymbol{\theta}_k\}^{\text{new}})}{p(\{\boldsymbol{\theta}_k\})} = \exp\left\{(\beta_i - \beta_j)\big(\ln p(\mathbf{x}|\boldsymbol{\theta}_j) - \ln p(\mathbf{x}|\boldsymbol{\theta}_i)\big)\right\} \ . \tag{4.7}$$

To fully satisfy detailed balance, the swap moves must be performed with a certain probability. Equally, swap moves can be proposed after a fixed number of single temperature Monte Carlo moves. The temperatures of the two replica i and j have to be close to each other to ensure non-negligible acceptance rates, and in practice only neighbouring temperature pairs are taken as candidates for replica exchanges. For detailed balance a pair $\{k, k+1\}$ can be chosen by uniformly choosing a k from between 1 and $K - 1$.

With this formulation the states of the replicas are effectively propagated from high to lower temperatures, and the mixing of the Markov chain is facilitated by the fast relaxation at higher temperatures.

4.1.3 Choosing a temperature set

A good choice of $\{\beta_k\}_{k=1}^K$ is according to a geometric progression, which we first motivate intuitively. Consider a replica exchange between chains k and $k+1$. From equation (4.7), the acceptance probability depends on the difference between $\ln p(\mathbf{x}|\boldsymbol{\theta}_k)$ and $\ln p(\mathbf{x}|\boldsymbol{\theta}_{k+1})$, and for

some swaps to be accepted this difference should not be 'too big'. For a simulation at in inverse temperature β, define the mean evaluation of the log likelihood as

$$\left\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \right\rangle_\beta = \int \ln p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \beta)\, d\boldsymbol{\theta} \ . \tag{4.8}$$

The entire distribution of $\ln p(\mathbf{x}|\boldsymbol{\theta})$ needs to be considered, and there should be an *overlap* of some of the log likelihood evaluations given by samples for adjacent chains, as shown by figure 4.1. The distribution of $-\ln p(\mathbf{x}|\boldsymbol{\theta})$ under different temperatures is shown in figure 4.1, and we consider the distribution of the negative log likelihood evaluations for an analogy with an energy function in a Gibbs distribution. The variance in chain β is

$$\sigma_\beta^2 = \mathsf{var}[\ln p(\mathbf{x}|\boldsymbol{\theta})]_\beta = \left\langle [\ln p(\mathbf{x}|\boldsymbol{\theta})]^2 \right\rangle_\beta - \left\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \right\rangle_\beta^2 \ , \tag{4.9}$$

and for the distributions between adjacent temperatures to have a 'meaningful' overlap, Iba (2001) has shown that the temperature points should be chosen according to the density

$$Q(\beta) \propto \sigma_\beta \ . \tag{4.10}$$

We give (4.10) in the simplest form amenable to applications in Bayesian inference; in (Ferkinghoff-Borg, 2002) and (Iba, 2001) this density is given in terms of the heat capacity, which also includes the system size. The following thought experiment should motivate an increase in replicas as the system size increases: it may be noted that with more data the likelihoods become increasingly peaked, and the 'energy' distributions become narrower and farther apart, and may lose their overlap. The number of replicas needed should therefore increase, at rate $\mathcal{O}(\sqrt{N})$, with the data set size N (the width of the distributions of energies sampled increases as the square root of the system size).

An obvious difficulty arises, as the variance of the log likelihood as a function of β is not known in advance, and must be estimated. The minimum required number K and the distribution of temperatures can therefore not be known a priori.

In practice a set of temperatures $\{\beta_k\}_{k=1}^K$ can be chosen according to a geometric series, and the average acceptance rate tracked over a short simulation. A rough plot of the estimated log likelihood averages and distributions can also give intuition on whether the chain will mix well over replicas. From (4.10), a larger number of replicas is needed in the region where the variance of the energy σ_β^2 takes larger values. Kofke (2002) notes that if the heat capacity is assumed to be constant, the average acceptance probability of a swap depends on temperatures only through their ratio, and that a geometric progression with $\beta_k/\beta_{k+1} = \mathsf{const}$ across all temperatures should result in equal acceptance ratios (see also (Kofke, 2004)).

4.2 Thermodynamic integration and the marginal likelihood

The samples from parallel tempering can be used for model comparison (Gregory, 2005; Skilling, 1998), as the marginal likelihood can be obtained from tempering. Firstly, notice that the integral

$$\int_0^1 d \ln \mathcal{Z}(\beta) = \int_0^1 \frac{d \ln \mathcal{Z}(\beta)}{d\beta}\, d\beta = \ln \mathcal{Z}(1) - \ln \mathcal{Z}(0) = \ln \mathcal{Z}(1) = \ln p(\mathbf{x}) \tag{4.11}$$

is equal to the marginal likelihood, as $\beta = 0$ gives the prior, which integrates to one. We therefore have to determine the derivative $\frac{d}{d\beta} \ln \mathcal{Z}(\beta)$, and it evaluates as an average over the posterior. Recall that

$$\mathcal{Z}(\beta) = \int p(\mathbf{x}|\boldsymbol{\theta})^\beta p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \ . \tag{4.12}$$

By taking the derivative of the log of the partition function,

$$\begin{aligned}\frac{d\ln\mathcal{Z}(\beta)}{d\beta} &= \frac{1}{\mathcal{Z}(\beta)}\int \ln p(\mathbf{x}|\boldsymbol{\theta}) \times p(\mathbf{x}|\boldsymbol{\theta})^{\beta} p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int \ln p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x},\beta)\, d\boldsymbol{\theta} = \left\langle \ln p(\mathbf{x}|\boldsymbol{\theta})\right\rangle_{\beta} , \end{aligned} \tag{4.13}$$

the log marginal likelihood can be evaluated with

$$\ln p(\mathbf{x}) = \int_0^1 \left\langle \ln p(\mathbf{x}|\boldsymbol{\theta})\right\rangle_{\beta} d\beta . \tag{4.14}$$

The integral in (4.14) can be numerically estimated from the Markov chain samples. Let $\{\boldsymbol{\theta}_k^{(t)}\}$ represent the samples for tempering parameter β_k, so that the expectation is approximated with

$$\left\langle \ln p(\mathbf{x}|\boldsymbol{\theta})\right\rangle_{\beta_k} \approx \frac{1}{T}\sum_{t=1}^{T} \ln p(\mathbf{x}|\boldsymbol{\theta}_k^{(t)}) . \tag{4.15}$$

We assume that a burn-in sample has been discarded in the sum over t. As we have run a set of chains in parallel at different inverse temperatures $0 = \beta_1 < \cdots < \beta_K = 1$, the integral can be evaluated numerically by interpolating the K expectations between zero and one (say with a piecewise cubic Hermite interpolation, available as part of `matlab` and other standard software packages), and using for example the trapesium rule to obtain the desired result.

4.2.1 The correct interpolation, or glitches at $\beta \approx 0$

To find a numeric approximation to (4.14), a set of log likelihood estimates are interpolated. In practice it may happen that we mark a huge factor of difference between the expectations at β_1 and β_2[1]. As an example, for the **galaxy** data set (with $v_{0j} = 0.01$ and $J = 3$ components) in sections 3.7 and 4.4.1, $\beta_1 = 0$ and $\beta_2 = 0.001$ gave log likelihood averages of -1.72×10^5 and -1.71×10^3 respectively—roughly a difference of a factor of a hundred for a very small change in β.[2] If the variance of the prior is increased to $v_{0j} = 10^{-6}$, this factor grows to roughly 300, giving a large average slope of 3×10^5 between $\langle \ln p(\mathbf{x}|\boldsymbol{\theta})\rangle_{\beta=0}$ and $\langle \ln p(\mathbf{x}|\boldsymbol{\theta})\rangle_{\beta=0.001}$. Figure 4.3 shows the latter log likelihood estimations.

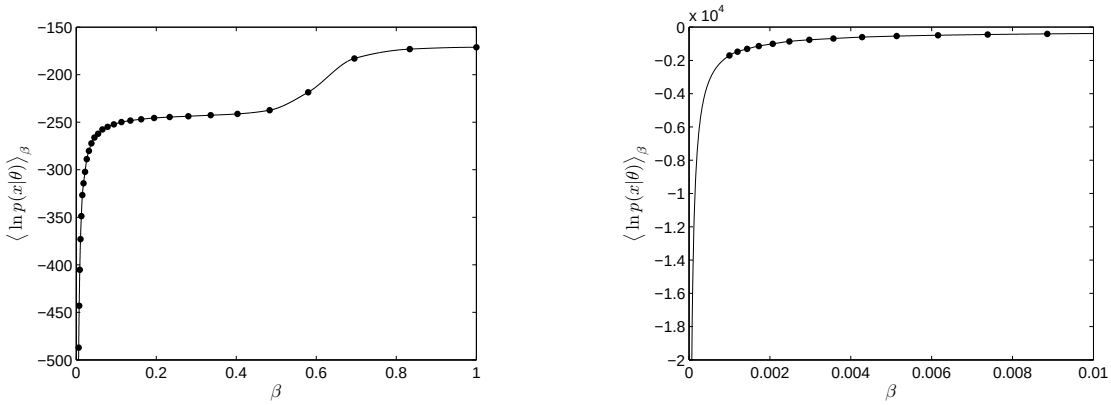
This will differ from problem to problem, and at $\beta = 0$ it depends on how likely the data is under the prior. In this case the log likelihood averages close to zero plays a crucial role in the integral evaluation, as the tail close to zero can grow like $-1/\beta$, which has an infinite integral between zero and one.

This problem can be either addressed by simulating at a much finer grid of temperatures in this interval, or by interpolating the tail with a guess of the exact functional form. The solution proposed here to interpolating correctly between zero and β_2 is to define a function of the form

$$f(\beta) = -\frac{1}{a\beta + b} + c , \tag{4.16}$$

[1]The term 'glitches at $\beta \approx 0$' is taken from a talk by David MacKay, presented at a *Recent Advances in Monte Carlo Based Inference* workshop at the Isaac Newton Institute, Cambridge, 2006.

[2]With Gibbs sampling, described in section 4.4.1, the averages were determined from 9000 samples, after a 1000 sample burn-in. With Gibbs sampling the averages $\langle \ln p(\mathbf{x}|\boldsymbol{\theta})\rangle_{\beta}$ are determined over latent variables $\mathbf{z}$ as well. For brevity here we assume $\boldsymbol{\theta}$ includes these extra variables that are averaged over.



(a) The estimate log likelihood averages $\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \rangle_\beta$ as a function of β. The values in this plot correspond to the (negative) *means* of the energy distributions from figure 4.1(a), i.e. $-\langle E(\boldsymbol{\theta}) \rangle_\beta$ as temperature $1/\beta$ decreases.

(b) A zoom for $\beta \approx 0$, showing the need for a careful interpolation. $\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \rangle_{\beta=0}$ is approximated as -1.72×10^5 (the intercept is outside the scale of the plot), while the same expectation at $\beta = 0.001$ is roughly a factor of a hundred bigger at -1.71×10^3.

FIGURE 4.2: The log likelihood averages $\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \rangle_{\beta_k}$ are estimated from each of the MCMC simulations at temperatures $\{\beta_k\}_{k=1}^K$, and interpolated so that the integral (4.14) can be evaluated numerically. In the problems (see section 4.5) evaluated here, the tail of the interpolation at $\beta \approx 0$ can have a marked difference in the evaluation of the integral, as it grows like $-1/\beta$. A good interpolation can be found with an increasingly fine temperature ladder close to zero. The alternative solution proposed here considers interpolating the tail with a $1/\beta$-like function. The specific problem illustrated here is again the **galaxy** data set, with $J = 3$ components, and $v_{0j} = 0.01$, and all other prior values set as usual.

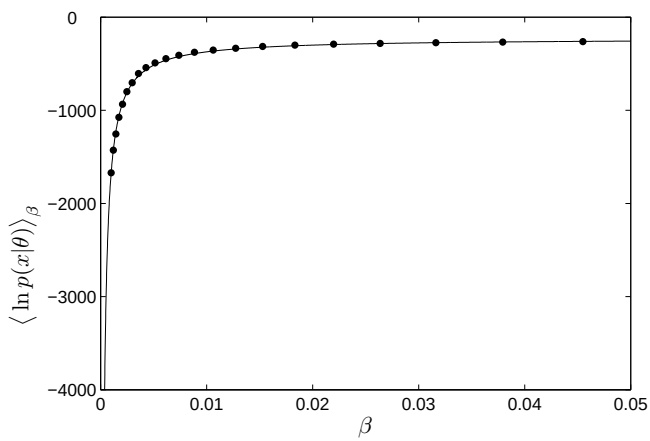


FIGURE 4.3: Motivation for using (4.16) to determine the value of the integral in (4.14) between $\beta_0 = 0$ and β_1. An interpolation with $f(\beta)$ from (4.16) is shown, passing through β_1, β_2, and a small β_k. The value $\beta_1 \approx -4 \times 10^{-7}$ is outside the range of the figure. *The interpolation is only used to evaluate the integral between* $\beta_1 = 0$ *and* β_2. The rest of the integral is determined numerically from a piecewise cubic Hermite interpolation and trapezium rule. The specific problem illustrated here is again the **galaxy** data set, with $J = 3$ components, and $v_{0j} = 10^{-6}$, and all other prior values set as usual.

and take three values to define the interpolation, $\beta_1 = 0$, β_2, and a β_k close to zero. Let the corresponding log likelihood expectations be L_1, L_2 and L_k. The coefficients are found by taking the log likelihood average of chain β_k an initial value for c, and then repeatedly solving

$$a = \Big(\frac{1}{c - L_1} - \frac{1}{c - L_2}\Big)\Big/(\beta_1 - \beta_2) \ , \tag{4.17}$$

$$b = \frac{1}{c - L_2} - a\beta_2 \ , \tag{4.18}$$

$$\text{and} \quad c = L_k + \frac{1}{a\beta_k + b} \ . \tag{4.19}$$

With $\beta_1 = 0$ and $\beta_K = 1$, integral (4.14) can in practice be written as

$$\begin{aligned} \ln p(\mathbf{x}) &= \int_{\beta_1}^{\beta_2} \big\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \big\rangle_\beta \, d\beta + \int_{\beta_2}^{\beta_K} \big\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \big\rangle_\beta \, d\beta \\ &\approx \int_{\beta_1}^{\beta_2} f(\beta) \, d\beta + \int_{\beta_2}^{\beta_K} \big\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \big\rangle_\beta \, d\beta \\ &= \frac{1}{a}\Big(\ln(a\beta_1 + b) - \ln(a\beta_2 + b) \Big) + c(\beta_2 - \beta_1) + \int_{\beta_2}^{\beta_K} \big\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \big\rangle_\beta \, d\beta \ . \end{aligned} \tag{4.20}$$

The second integral does not include averages around zero, and lends itself to a numerically stable solution.

4.3 A practical generalization of parallel tempering

The success of the interpolation obtaining $\langle \log p(\mathbf{x}|\boldsymbol{\theta}) \rangle_\beta$, illustrated in figure 2, is dependent on the slope

$$\frac{d\langle \log p(\mathbf{x}|\boldsymbol{\theta}) \rangle_\beta}{d\beta} = \frac{d^2 \log \mathcal{Z}(\beta)}{d\beta^2} = \sigma_\beta^2 \tag{4.21}$$

at $\beta \approx 0$. Consider the following thought exercise: Imagine a non-informative (infinitely wide) prior at $\beta = 0$. Samples from this prior will strictly speaking have an infinite variance σ_0^2. With $\beta \approx 0$ we introduce the likelihood, practically infinitely decreasing the variance of our samples, causing $\langle \log p(\mathbf{x}|\boldsymbol{\theta}) \rangle_\beta$ to asymptotically diverge at zero. As we narrow our prior the change in this mean should be less rapid, and this motivates a generalization of parallel tempering and thermodynamic integration such that we get a more stable interpolation.

An ingenious idea proposed by Winther (2007) is to introduce a new distribution $q(\boldsymbol{\theta})$, which might be a narrower version of the prior, so that equation (4.1) can be modified to

$$p(\boldsymbol{\theta}|\mathbf{x}, \beta) = \frac{1}{\mathcal{Z}(\beta)} \left[p(\mathbf{x}|\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right]^\beta q(\boldsymbol{\theta}) \ . \tag{4.22}$$

At $\beta = 0$ we are therefore substituting the prior with $q(\boldsymbol{\theta})$. As our 'effective prior' is $p(\boldsymbol{\theta})^\beta q(\boldsymbol{\theta})^{1-\beta}$, the prior's influence is gradually increased, while q's role is decreased until only the posterior remains at $\beta = 1$. With an informed choice of q, which should be closer than the prior to the true posterior, we hope to decrease σ_β^2. The log marginal likelihood can, as before, be determined with

$$\ln p(\mathbf{x}) = \int_0^1 \Big\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) + \ln \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \Big\rangle_\beta \, d\beta \ . \tag{4.23}$$

It does not escape our attention that setting $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x})$ gives $\ln p(\mathbf{x}) = \int_0^1 \langle \ln p(\mathbf{x}) \rangle_\beta \, d\beta$. This suggests a wealth of possibilities of approximating $p(\boldsymbol{\theta}|\mathbf{x})$ with $q(\boldsymbol{\theta})$ to effectively combine deterministic methods of inference with Markov chains. This comes with a cautionary note as Variational Bayes, for example, may give a $q(\boldsymbol{\theta})$ that captures (lower-bounds) a mode of a possibly multimodal posterior, causing PT to lose its pleasing property of fast relaxation at high temperatures. In the results presented in section 3.7, setting q to a narrower version of the prior, where necessary, was found to give adequate results.

In section 4.4.2 a short generalization is given to sample from (4.22) for the mixture of Gaussians problem.

4.4 Gibbs sampling for parallel tempering

Parallel tempering of a mixure of Gaussian distributions $p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) = \sum_{j=1}^J \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})$ require a Monte Carlo simulation at inverse temperature β. As the results obtained here complements that of chapter 3, let the mixing weights and component priors again be Dirichlet and Normal-Wishart, with

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}_0) = \frac{\Gamma(\sum_{j=1}^J \delta_{0j})}{\prod_j \Gamma(\delta_{0j})} \prod_j \pi_j^{\delta_{0j}-1} \tag{4.24}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{j=1}^J \mathcal{N}\big(\boldsymbol{\mu}_j|\mathbf{m}_{0j}, (v_{0j}\boldsymbol{\Lambda}_j)^{-1}\big)\mathcal{W}(\boldsymbol{\Lambda}_j|a_{0j}, \mathbf{B}_{0j}) \ , \tag{4.25}$$

where

$$\mathcal{N}\big(\boldsymbol{\mu}_j|\mathbf{m}_{0j}, (v_{0j}\boldsymbol{\Lambda}_j)^{-1}\big) = \Big(\frac{v_{0j}}{2\pi}\Big)^{\frac{d}{2}} |\boldsymbol{\Lambda}_j|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\mathsf{tr}[(\boldsymbol{\mu}_j - \mathbf{m}_{0j})(\boldsymbol{\mu}_j - \mathbf{m}_{0j})^\top v_{0j}\boldsymbol{\Lambda}_j]\Big\} \tag{4.26}$$

$$\mathcal{W}(\boldsymbol{\Lambda}_j|a_{0j}, \mathbf{B}_{0j}) = \frac{|\mathbf{B}_{0j}|^{a_{0j}}}{\prod_{i=1}^d \Gamma(a_j + \frac{1-i}{2})} \pi^{\frac{-d(d-1)}{4}} |\boldsymbol{\Lambda}_j|^{a_{0j} - \frac{d+1}{2}} \exp\Big\{ -\mathsf{tr}[\mathbf{B}_{0j}\boldsymbol{\Lambda}_j]\Big\} \ . \tag{4.27}$$

In the following two sections describe an implementation of Gibbs sampling to sample from firstly $p(\boldsymbol{\theta}|\mathbf{x}, \beta) \propto p(\mathbf{x}|\boldsymbol{\theta})^\beta p(\boldsymbol{\theta})$, and then from the tempered posterior $p(\boldsymbol{\theta}|\mathbf{x}, \beta) \propto p(\mathbf{x}|\boldsymbol{\theta})^\beta p(\boldsymbol{\theta})^\beta q(\boldsymbol{\theta})^{1-\beta}$ needed for the generalized version of parallel tempering.

4.4.1 Gibbs sampling at β

To implement a Gibbs sampler, we extend the parameter space to include latent allocation variables $\mathbf{z}_n$ for each data point n, to indicate which mixture component was responsible for generating it (Diebolt & Robert, 1994). Consequently $z_{nj} \in \{0, 1\}$, and $\sum_{j=1}^J z_{nj} = 1$. The complete joint distribution is therefore

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{n=1}^N \prod_{j=1}^J \Big[\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})\Big]^{z_{nj}} p(\boldsymbol{\theta}) \ , \tag{4.28}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$. We can write the complete data likelihood as $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z})p(\mathbf{z}|\boldsymbol{\theta})$, and in this form the likelihood, to the power β, is

$$p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z})^\beta = \prod_{n=1}^N \prod_{j=1}^J \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})^{\beta z_{nj}} \tag{4.29}$$

with the prior over $\boldsymbol{\theta}, \mathbf{z}$ being

$$p(\boldsymbol{\theta}, \mathbf{z}) = \prod_{n=1}^{N} \prod_{j=1}^{J} \pi_j^{z_{nj}} p(\boldsymbol{\theta}) \ . \tag{4.30}$$

With inverse temperature parameter β the tempered posterior distribution is

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x}, \beta) \propto p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z})^{\beta} p(\boldsymbol{\theta}, \mathbf{z}) = \prod_{n=1}^{N} \prod_{j=1}^{J} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})^{\beta z_{nj}} \pi_j^{z_{nj}} p(\boldsymbol{\theta}) \ , \tag{4.31}$$

and can be treated as any missing-value Gibbs sampling problem. The allocation variables are sampled with

$$z_{nj}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda} \sim \frac{\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})^{\beta}}{\sum_{k=1}^{J} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{\beta}} \ . \tag{4.32}$$

Given the allocation variables, we define

$$\begin{aligned} \gamma_{nj} &= \beta z_{nj} & \bar{\mathbf{x}}_j &= \frac{1}{N_j} \sum_{n=1}^{N} \gamma_{nj} \mathbf{x}_n \\ N_j &= \sum_{n=1}^{N} \gamma_{nj} & \boldsymbol{\Sigma}_j &= \frac{1}{N_j} \sum_{n=1}^{N} \gamma_{nj} (\mathbf{x}_n - \bar{\mathbf{x}}_j)(\mathbf{x}_n - \bar{\mathbf{x}}_j)^{\top} \ . \end{aligned} \tag{4.33}$$

to give the conditional distributions needed for sampling the mixture parameters as

$$\boldsymbol{\pi}|\mathbf{z} \sim \mathcal{D}\big(\delta_{01} + \frac{1}{\beta} N_1, \ldots, \delta_{0J} + \frac{1}{\beta} N_J\big) \tag{4.34}$$

$$\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j|\mathbf{z} \sim \mathcal{NW}\big(v_j, \mathbf{m}_j, a_j, \mathbf{B}_j\big) \ , \tag{4.35}$$

with

$$v_j = v_{0j} + N_j \tag{4.36}$$

$$\mathbf{m}_j = \frac{v_{0j}\mathbf{m}_{0j} + N_j \bar{\mathbf{x}}_j}{v_{0j} + N_j} \tag{4.37}$$

$$a_j = a_{0j} + \frac{N_j}{2} \tag{4.38}$$

$$\mathbf{B}_j = \mathbf{B}_{0j} + \frac{1}{2} N_j \boldsymbol{\Sigma}_j + \frac{1}{2} \frac{N_j v_{0j} (\bar{\mathbf{x}}_j - \mathbf{m}_{0j})(\bar{\mathbf{x}}_j - \mathbf{m}_{0j})^{\top}}{v_{0j} + N_j} \ . \tag{4.39}$$

As $p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{z})\, d\boldsymbol{\theta} d\mathbf{z}$, we use the samples over $\boldsymbol{\theta}$ and $\mathbf{z}$ to estimate the average log likelihood. If $\{\boldsymbol{\pi}_k^{(t)}, \{\boldsymbol{\mu}_{j,k}^{(t)}, \boldsymbol{\Lambda}_{j,k}^{(t)}\}_{j=1}^{J}, \{z_{n,k}^{(t)}\}_{n=1}^{N}\}_{t=1}^{T}$ indicates the samples of chain k (after a burn-in period), then

$$\big\langle \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z}) \big\rangle_{\beta_k} \approx \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{j=1}^{J} z_{nj,k}^{(t)} \ln \mathcal{N}\big(\mathbf{x}_n|\boldsymbol{\mu}_{j,k}^{(t)}, \boldsymbol{\Lambda}_{j,k}^{(t)\,-1}\big) \ . \tag{4.40}$$

Notice that the samples of the mixing weights $\boldsymbol{\pi}_k^{(t)}$ are not used in estimating the log likelihood average over the posterior, but occur in the prior.

Algorithm 2 Parallel tempering

initialize: $\boldsymbol{\pi}_k^{(0)}$ and $\{\boldsymbol{\mu}_{j,k}^{(0)}, \boldsymbol{\Lambda}_{j,k}^{(0)}\}_{j=1}^J$ for all chains k; tempering sequence $\{\beta_k\}_{k=1}^K$ according to section 4.1.3; $t = 0$.
repeat
  for $k = 1$ to K **do**
    for $n = 1$ to N **do**
      sample $\mathbf{z}_{n,k}^{(t+1)} | \boldsymbol{\pi}_k^{(t)}, \{\boldsymbol{\mu}_{j,k}^{(t)}, \boldsymbol{\Lambda}_{j,k}^{(t)}\}_{j=1}^J$ according to (4.32).
    end for
    sample $\boldsymbol{\pi}_k^{(t+1)} | \mathbf{z}_k^{(t+1)}$ according to (4.34).
    for $j = 1$ to J **do**
      sample $\boldsymbol{\mu}_{j,k}^{(t+1)}, \boldsymbol{\Lambda}_{j,k}^{(t+1)} | \mathbf{z}_k^{(t+1)}$ according to (4.35).
    end for
  end for
  uniformly choose a chain $i \in \{1, \ldots, K-1\}$.
  sample a uniform random variable $u \sim \mathcal{U}(0,1)$.
  if $u \leq \alpha(\{\boldsymbol{\theta}_k^{\text{new}}\} | \{\boldsymbol{\theta}_k^{(t+1)}\})$ *(see (4.5))* **then**
    swap $\{\mathbf{z}_i^{(t+1)}, \boldsymbol{\theta}_i^{(t+1)}\}$ and $\{\mathbf{z}_{i+1}^{(t+1)}, \boldsymbol{\theta}_{i+1}^{(t+1)}\}$
  end if
  $t \leftarrow t + 1$
until $t = t_{\max}$
for $k = 1$ to K **do**
  estimate $\left\langle \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z}) \right\rangle_{\beta_k}$ with (4.40), using samples after some burn-in period *(this step can be included in the main loop over t).*
end for
interpolate $\{\left\langle \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z}) \right\rangle_{\beta_k}\}_{k=1}^K$ between 0 and 1, possibly following section 4.2.1.
numerically estimate $\ln p(\mathbf{x})$ as the volume under the interpolation.

4.4.2 Gibbs sampling at β for generalized parallel tempering

We have a prior distribution $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{j=1}^J p(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$, and to implement a Gibbs sampler for generalized parallel tempering, we choose $q(\boldsymbol{\theta})$ to be of the same form as the prior. As in (4.30), both the prior and $q(\boldsymbol{\theta})$ are extended to include latent allocation variables $\mathbf{z}$. The tempered posterior distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x}, \beta) &= \frac{1}{\mathcal{Z}(\beta)} \Big[p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z}) \frac{p(\boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}, \mathbf{z})} \Big]^\beta q(\boldsymbol{\theta}, \mathbf{z}) \\ &\propto \prod_{n=1}^N \prod_{j=1}^J \Big[\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})^{z_{nj}} \frac{\pi_j^{z_{nj}}}{\pi_j^{z_{nj}}} \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \Big]^\beta \times \pi_j^{z_{nj}} q(\boldsymbol{\theta}) \\ &= \prod_{n=1}^N \prod_{j=1}^J \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})^{\beta z_{nj}} \pi_j^{z_{nj}} p(\boldsymbol{\theta})^\beta q(\boldsymbol{\theta})^{1-\beta} \ . \end{aligned} \tag{4.41}$$

The tempered posterior can be factorized into a 'likelihood' and a prior:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x}, \beta) = \frac{1}{\mathcal{Z}(\beta)} \Big[\prod_{n=1}^N \prod_{j=1}^J \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})^{z_{nj}} \frac{p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda})} \Big]^\beta$$

$$\times \prod_{n=1}^{N} \prod_{j=1}^{J} \pi_j^{z_{nj}} q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \ . \tag{4.42}$$

To implement a Gibbs sampling as before, we have to determine the parameters of the 'effective' prior $p(\boldsymbol{\theta}, \mathbf{z})^{\beta} q(\boldsymbol{\theta}, \mathbf{z})^{1-\beta}$. To differentiate between the parameters of p and q, superscripts p and q will be used. The 'prior' parameters, to be used in equations (4.34) to (4.39), are

$$\boldsymbol{\delta}_0 = \beta \boldsymbol{\delta}_0^p + (1-\beta)\boldsymbol{\delta}_0^q \tag{4.43}$$

$$v_{0j} = \beta v_{0j}^p + (1-\beta) v_{0j}^q \tag{4.44}$$

$$\mathbf{m}_{0j} = \frac{\beta v_{0j}^p \mathbf{m}_{0j}^p + (1-\beta) v_{0j}^q \mathbf{m}_{0j}^q}{\beta v_{0j}^p + (1-\beta) v_{0j}^q} \tag{4.45}$$

$$a_{0j} = \beta a_{0j}^p + (1-\beta) a_{0j}^q \tag{4.46}$$

$$\mathbf{B}_{0j} = \beta \mathbf{B}_{0j}^p + (1-\beta)\mathbf{B}_{0j}^q \tag{4.47}$$

$$+ \frac{1}{2} \frac{\beta v_{0j}^p (1-\beta) v_{0j}^q}{\beta v_{0j}^p + (1-\beta) v_{0j}^q} (\mathbf{m}_{0j}^p - \mathbf{m}_{0j}^q)(\mathbf{m}_{0j}^p - \mathbf{m}_{0j}^q)^{\top} \ . \tag{4.48}$$

The empirical expectation given in (4.40) should be generalized to

$$\left\langle \ln p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z}) + \ln \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\rangle_{\beta_k} \approx \frac{1}{T} \sum_{t=1}^{T} \Bigg[\ln p(\boldsymbol{\pi}_k^{(t)}) - \ln q(\boldsymbol{\pi}_k^{(t)}) + \sum_{j=1}^{J} \Big[\ln p(\boldsymbol{\mu}_{j,k}^{(t)}, \boldsymbol{\Lambda}_{j,k}^{(t)})$$
$$- \ln q(\boldsymbol{\mu}_{j,k}^{(t)}, \boldsymbol{\Lambda}_{j,k}^{(t)}) + \sum_{n=1}^{N} z_{nj,k}^{(t)} \ln \mathcal{N}\big(\mathbf{x}_n | \boldsymbol{\mu}_{j,k}^{(t)}, \boldsymbol{\Lambda}_{j,k}^{(t)\,-1}\big) \Big] \Bigg] \ . \tag{4.49}$$

4.5 Experimental results

For a practical evaluation of tempered Gibbs sampling, the reader is asked to return to section 3.7, where the marginal likelihoods of parallel tempering was compared to that of different deterministic methods. This section gives the details of the method in algorithm 2, together with a discussion on some practicalities.

The parameter values were initialized by randomly drawing $\boldsymbol{\pi}_k^{(0)}$ from the prior $p(\boldsymbol{\pi})$ for each chain k. To initialize the means, a k-means algorithm was run for each chain k, and the means found assigned to the means in the set $\{\boldsymbol{\mu}_{j,k}^{(0)}\}_{j=1}^{J}$. The ordering of the assignment of means is random, and hence different chains should start in different areas of the parameter space, or different modes (we want the chains to be in different modes for good mixing, as we want all modes to be equally visited). The precision matrices $\{\boldsymbol{\Lambda}_{j,k}^{(0)}\}_{j=1}^{J}$ were all initialized to the inverse covariance matrix of the entire data set.

One of the biggest difficulties encountered was to reasonably interpolate $\langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \rangle_{\beta}$ around $\beta \approx 0$. For the broad prior used in the experimental section 3.7, there is a factor of around 300 difference between $\beta_1 = 0$ and $\beta_2 = 0.001$. This gives an average slope of $\Delta \langle \ln p(\mathbf{x}|\boldsymbol{\theta}) \rangle_{\beta}$ as 300,000 for β close to zero. This slope is dependent on the prior width, and decreases as the prior gets narrower. The final estimate to $\ln p(\mathbf{x})$ is *very* sensitive to the numerical method around zero and form of interpolation used to evaluate the integral; where MCMC in section 3.7 gives less desirable results, this is usually the cause, and not the chain not mixing well.

4.6 Discussion: annealed importance sampling

Annealed importance sampling (AIS) (Neal, 2001) is a method closely related to PT. It was previously used by Beal & Ghahramani (2003) in a very similar context as section 3.7, namely as a sampling standard in scoring different marginal likelihood likelihood approximations. We therefore aim to present a brief overview, following the introduction given in section 1.3.2. If we are presented with a sample $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T$ from some distribution with at least the same support as the posterior, with associated importance weights $\{\omega^{(t)}\}_{t=1}^T$, we can approximate the expected value of a function with

$$\langle \phi(\boldsymbol{\theta}) \rangle = \int \phi(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \approx \frac{\sum_{t=1}^T \omega^{(t)} \phi(\boldsymbol{\theta}^{(t)})}{\sum_{t=1}^T \omega^{(t)}} \ . \tag{4.50}$$

The accuracy of this estimate depends on the variability of the importance weights, and for the method to work well, we are faced with the difficult task of finding an importance sampling distribution that approximates the posterior well. (See section 5.4.2 for a practical discussion.)

AIS works by constructing a series of distributions that progressively approximates the posterior well. As in PT, we create a series of distributions $p(\boldsymbol{\theta}|\mathbf{x}, \beta)$, which we only need to know up to a normalizing constant. In this setting we shall call it

$$p^*_\beta(\boldsymbol{\theta}) = p^*(\boldsymbol{\theta}|\mathbf{x}, \beta) = p(\mathbf{x}|\boldsymbol{\theta})^\beta p(\boldsymbol{\theta}) \ . \tag{4.51}$$

A temperature ladder $\{\beta_k\}_{k=1}^K$, with $\beta_k < \beta_{k+1}$, is again constructed. We let $\beta_1 = 0$ recapture the prior and $\beta_K = 1$ recapture the posterior. AIS produces a sample $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T$ with associated importance weights $\{\omega^{(t)}\}_{t=1}^T$ as follows: For each distribution $k = 2, \ldots, K-1$ at inverse temperature β_k (therefore excluding the prior at $\beta_1 = 0$ and the posterior at $\beta_K = 1$) we define a transition kernel $\mathcal{K}_k(\boldsymbol{\theta}|\boldsymbol{\theta}')$. This can be a standard Metropolis-Hastings transition kernel or a Gibbs sampling update.

To generate a sample $\boldsymbol{\theta}^{(t)}$ and its associated weight $\omega^{(t)}$ we first generate a sequence of points, walking down the ladder of distributions from an 'infinite' temperature (the prior) to the posterior:

$$\begin{aligned}
&\text{Generate } \boldsymbol{\theta}_2 \text{ from } p(\boldsymbol{\theta}|\mathbf{x}, \beta_1) = p(\boldsymbol{\theta}) \ . \\
&\text{Generate } \boldsymbol{\theta}_3 \text{ from } \mathcal{K}_2(\boldsymbol{\theta}_3|\boldsymbol{\theta}_2) \ . \\
&\qquad \ldots \\
&\text{Generate } \boldsymbol{\theta}_{K-1} \text{ from } \mathcal{K}_{K-2}(\boldsymbol{\theta}_{K-1}|\boldsymbol{\theta}_{K-2}) \ . \\
&\text{Generate } \boldsymbol{\theta}_K \text{ from } \mathcal{K}_{K-1}(\boldsymbol{\theta}_K|\boldsymbol{\theta}_{K-1}) \ .
\end{aligned} \tag{4.52}$$

Finally set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_K$, and let its associated weight be

$$\omega^{(t)} = \frac{p^*_2(\boldsymbol{\theta}_2)}{p^*_1(\boldsymbol{\theta}_2)} \frac{p^*_3(\boldsymbol{\theta}_3)}{p^*_2(\boldsymbol{\theta}_3)} \cdots \frac{p^*_{K-1}(\boldsymbol{\theta}_{K-1})}{p^*_{K-2}(\boldsymbol{\theta}_{K-1})} \frac{p^*_K(\boldsymbol{\theta}_K)}{p^*_{K-1}(\boldsymbol{\theta}_K)} \ . \tag{4.53}$$

The marginal likelihood can be estimated from the importance weights, as the average of the weights converges to the ratio of normalizers,

$$\frac{1}{T} \sum_{t=1}^T \omega^{(t)} \to \frac{\int p^*_K(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p^*_1(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \quad \text{as } T \to \infty \ . \tag{4.54}$$

In a Bayesian setting this ratio will be $\int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta} / \int p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, which is equal to the marginal likelihood. Here we do not even require that the prior is normalized, as all constant factors will cancel in this ratio. For the marginal likelihood to be correct, though, the likelihood has to include constant factors.

4.7 Summary and outlook

We have seen how parallel tempering can be an effective tool to both sample from multimodal posterior distributions, and estimate the log marginal likelihood.

4.7.1 Other MCMC schemes

A practical problem arises at near-infinite temperatures, or $\beta \approx 0$, as the log likelihood estimates grow rapidly as a function of β. This problem is aggrevated by the fact that we have increased the dimensionality that we are averaging over with the inclusion of latent variables. A MH method could have been used instead, with

$$p(\boldsymbol{\theta}|\mathbf{x},\beta) = \frac{1}{\mathcal{Z}(\beta)} p(\mathbf{x}|\boldsymbol{\theta})^{\beta} p(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(\beta)} \prod_{n=1}^{N} \Big[\sum_{j=1}^{J} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1}) \Big]^{\beta} p(\boldsymbol{\theta}) \ , \tag{4.55}$$

as we do not have to increase the space that we have to average over. However, the β power makes the problem ideally suited to Gibbs sampling, as we can formulate the problem such that all the conditional distributions are tractable.

It would have been equally possible to integrate out $\boldsymbol{\theta}$, and sample over $p(\mathbf{z}|\mathbf{x},\beta)$, where MH proposals could be 'bit flips' on the discrete indicator vectors $\{\mathbf{z}_n\}_{n=1}^{N}$. Each $\mathbf{z}_n$ would still be constrained to have one non-zero bit, indicating to which of the mixture components an observed $\mathbf{x}_n$ belong. The density that we would sample from in this case would be

$$\begin{aligned} p(\mathbf{z}|\mathbf{x},\beta) &= \frac{1}{\mathcal{Z}(\beta)} p(\mathbf{x}|\mathbf{z})^{\beta} p(\mathbf{z}) \\ &= \frac{1}{\mathcal{Z}(\beta)} \prod_{j=1}^{J} \frac{(2\pi)^{-dN_j/2} \mathcal{Z}_{\mathcal{NW}}(v_j, a_j, \mathbf{B}_j)}{\mathcal{Z}_{\mathcal{NW}}(v_{0j}, a_{0j}, \mathbf{B}_{0j})} \times \frac{\mathcal{Z}_{\mathcal{D}}(\delta_{01} + \frac{1}{\beta}N_1, \ldots, \delta_{0J} + \frac{1}{\beta}N_J)}{\mathcal{Z}_{\mathcal{D}}(\delta_{01}, \ldots, \delta_{0J})} \ , \end{aligned} \tag{4.56}$$

where v_j, a_j and $\mathbf{B}_j$ are defined in equations (4.36) to (4.39), with dependance on $\mathbf{z}$ through equation (4.33).

4.7.2 Choices for $q(\boldsymbol{\theta})$

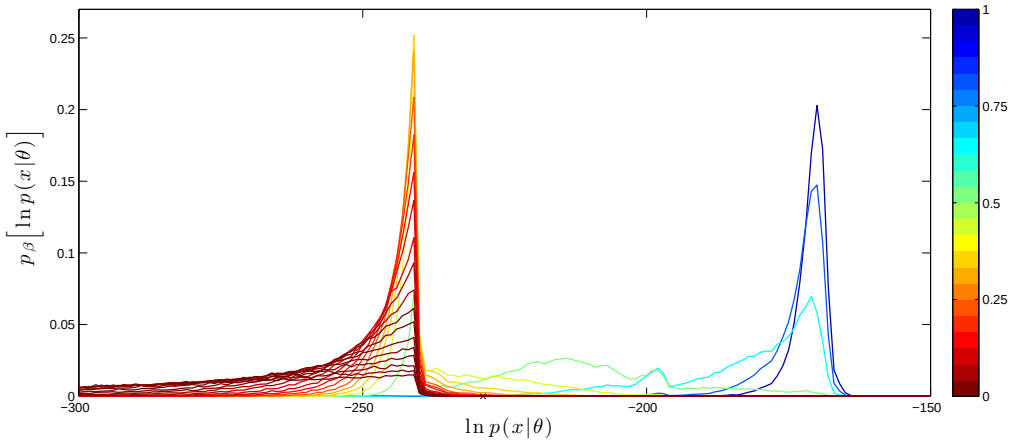
The choice of surrogate prior $q(\boldsymbol{\theta})$ has an effect on the performance of PT, and was originally introduced to reduce σ_β^2 at small values of β. In section 4.3 we argued that $q(\boldsymbol{\theta})$ should match the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ as closely as possible.

One choice of $q(\boldsymbol{\theta})$ might be a narrower version of the prior. As the posterior is typically multimodal, we have to ensure that in some way all the possible posterior modes are still 'well covered' by $q(\boldsymbol{\theta})$. We can make an intelligent choice based on knowledge of the scale of the data set.

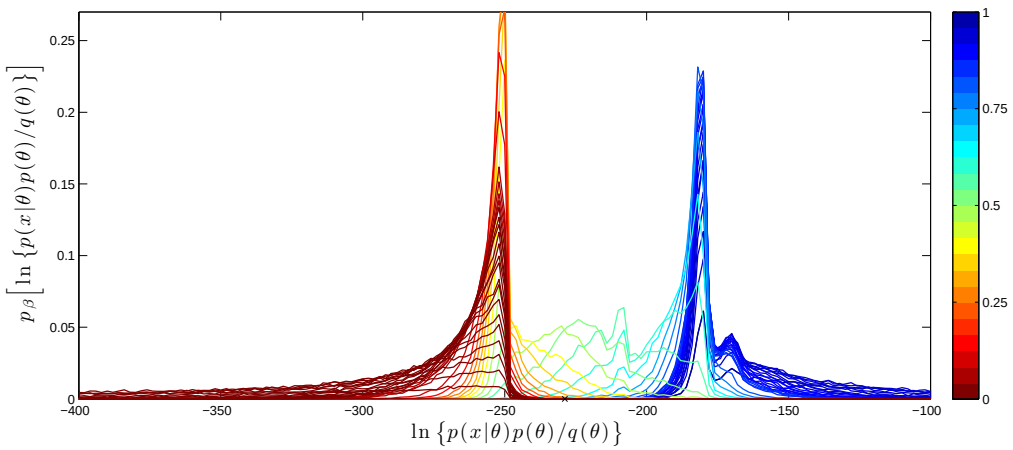
Another possibility[3] is to take a small subset of data points, and analytically evaluate the posterior using the chosen subset. This can in turn be taken as choice for $q(\boldsymbol{\theta})$. With more data points included, we can expect better performance of PT. There is a trade-off in including more data points into the surrogate prior, though, as the problem's original difficulty lay in the fact that the number of posterior terms grew as J^N. We do not give an explicit derivation of $q(\boldsymbol{\theta})$ here—it is a mixture of Dirichlet-Normal-Wishart products, which can again be effectively sampled from using Gibbs sampling and further latent variables. An energy histogram obtained by this improved scheme is shown in figure 4.4(b).

[3]Thanks to Zoubin Ghahramani for this idea.

The energy histograms for PT and generalized PT are shown in figures 4.4(a) and 4.4(b). It is possible to observe phase transitions, with largely disconnected energy densities under various inverse temperatures β. In figure 4.4(a) the maximum attainable energy would be the log maximum likelihood evaluation; notice that this is not true any more with the introduction of a surrogate prior. Notice also the 'better connection' between the two energy peaks in figure 4.4(b). A further step to achieve better sampling across the energy spectrum is multicanonical sampling (Berg, 2000). This leaves us with a starting point for venturing into the vast fields of statistical physics, in the hope of deriving even better algorithms.



(a) Energy histograms without the introduction of a surrogate prior.



(b) Energy histograms with the introduction of a surrogate prior $q(\boldsymbol{\theta})$. The surrogate prior was chosen as the analytic posterior $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x}_{n_1}, \mathbf{x}_{n_3}, \mathbf{x}_{n_3})$ on seeing three (out of 82) data points.

FIGURE 4.4: This figure follows figure 4.1(a), which shows an approximation to the distributions of $E(\boldsymbol{\theta}) = -\ln p(\mathbf{x}|\boldsymbol{\theta})$ under replicas at different temperatures, $p(\boldsymbol{\theta}|\mathbf{x}, \beta)$, using the means and standard deviations from samples obtained from different temperature chains. By rather binning evaluations of $\ln p(\mathbf{x}|\boldsymbol{\theta})$, or $\ln\{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})/q(\boldsymbol{\theta})\}$, we obtain more accurate energy histograms. These plots show the energy histograms for a selection of reciprocal temperatures β (on a coloured scale) for the **galaxy** data set with $J = 3$ components, a prior $v_{0j} = 0.01$, and all other prior parameters following section 3.7.

Chapter 5

Variational Transition Kernels

5.1 Introduction

In this chapter, variational methods are incorporated into the design of Markov chain Monte Carlo (MCMC) transition kernels. We introduce a new Monte Carlo algorithm, based on a variational approximation to the transition kernel of a Markov chain that has the parameter posterior as invariant distribution, for performing inference with latent variable models.

The idea of combining variational and Monte Carlo methods is by no means new. A mixture of two MCMC kernels—a sample from the (static) variational approximation to the posterior mixed with a random-walk Metropolis step—was used by de Freitas et al. (2001). The variational approximation is used as proposal distribution so that regions of high probability are efficiently located; as the approximation underestimates the true variance, a Metropolis kernel is used to sample from these regions. Ghahramani & Beal (2000) have previously used a variational approximation for mixtures of factor analyzers as the proposal density for an importance sampler. In this chapter the variational proposal is not static but *adaptive*, as it depends on the previous state of the chain. This also allows for greater exploration of the parameter space. No iterative method is needed to find the optimal variational distribution: a closed-form solution for the optimal kernel, based on the previous state of the chain, exists. The variational transition kernel will also allow us to circumvent the explicit latent variable sample that is common to Gibbs sampling for latent variable models.

This chapter started out as an attempt at a new idea, which finally proved to be less effective than expected. What we will illustrate here are the inherent shortcomings of variational methods in MCMC, giving further insight into why de Freitas et al. (2001), for example, found it necessary to mix random-walk steps into their MCMC algorithm. The shortcomings mostly pertain to problems with the variance of estimates obtained from the samples. (The cause of this behavior can be traced back to section 2.2 in chapter 2, where it was illustrated how VB underestimates the true variance of the density it approximates.) For Metropolis-Hastings, the resulting chain cannot be shown to be geometrically ergodic, and we cannot show the existence of a central limit theorem for our estimate. We will show how importance sampling can suffer from a similar problem of possibly infinite variance of the importance weights.

The rest of the chapter takes a short theoretical tour through the world of general state space Markov chains and Monte Carlo integration in section 5.2, presenting some theory needed to show different forms of convergence later in the chapter. The Metropolis Hastings algorithm is introduced as a practical method of constructing a Markov chain with a particular predefined distribution as its stationary distribution. We then introduce variational methods into MCMC transition kernels in section 5.3. We give a detailed analysis on a small toy example in section

5.4, to see what the shortcoming of such an approach is.

5.2 Monte Carlo methods

The principle of drawing samples from a distribution, and using sample averages to approximate expectations, lies at the heart of Monte Carlo integration. This forms a flexible method with great scope in statistical modeling. The exposition presented here assumes that some *posterior* distribution is the distribution of interest, but the underlying principles can readily be applied to any distribution.

5.2.1 Monte Carlo integration

Many tasks in Bayesian inference require the evaluation of an expectation of some function, say $\phi(\boldsymbol{\theta})$, over a distribution of interest. Our interest generally lies in the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})$, for which we implicitly condition on the model assumptions $\mathcal{M}$ to write the expectation as

$$\Phi = \langle \phi(\boldsymbol{\theta}) \rangle = \int \phi(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \ . \tag{5.1}$$

The principle behind Monte Carlo integration is to approximate such an integral with an empirical average; if we are able to draw *independent and identically distributed* samples $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$ from the target distribution $p(\boldsymbol{\theta}|\mathbf{x})$, then Φ can be approximated with the unbiased estimate

$$\hat{\Phi}_T = \frac{1}{T} \sum_{t=1}^{T} \phi(\boldsymbol{\theta}^{(t)}) \to \Phi \ . \tag{5.2}$$

By the law of large numbers the estimate will converge *almost surely* to Φ, $\mathbb{P}(\hat{\Phi}_T = \Phi) = 1$ as $T \to \infty$. Crucially, the variance of the estimate will be well behaved: If the variance of $\phi(\boldsymbol{\theta})$ is finite, $\sigma_\phi^2 \equiv \langle \phi^2(\boldsymbol{\theta}) \rangle - \langle \phi(\boldsymbol{\theta}) \rangle^2 < +\infty$, then the variance of the estimate is equal to $\mathsf{var}(\hat{\Phi}_T) = \sigma_\phi^2 / T$, and a central limit theorem yields convergence in distribution of the error $\sqrt{T}(\hat{\Phi}_T - \Phi) \to \mathcal{N}(0, \sigma_\phi^2)$ as $T \to \infty$ (Andrieu et al., 2003). The central limit theorem result holds if we are able to draw i.i.d. samples.

It is generally not possible to draw independent samples from $p(\boldsymbol{\theta}|\mathbf{x})$. However, the samples $\{\boldsymbol{\theta}^{(t)}\}$ need not be independent, but can loosely speaking be simulated using any process that draws samples from the support of $p(\boldsymbol{\theta}|\mathbf{x})$ in the correct proportions. One possible such process is a Markov chain that has $p(\boldsymbol{\theta}|\mathbf{x})$ as its stationary distribution; the chain can be simulated and the resulting series of states taken as samples. For brevity, we let $p^*(\boldsymbol{\theta})$ indicate the target distribution.

5.2.2 Markov chains

A Markov chain generates a sequence of random variables $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots\}$ where the next state of the chain $\boldsymbol{\theta}^{(t+1)}$ is sampled from a *transition kernel* $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$[1], a conditional probability density that only depends on the current state of the chain. The next state of the chain is therefore *independent* of the history of the chain. Markov chain Monte Carlo methods are constructed from a time homogeneous chain, that is, the transition kernel is independent of t.

[1]Instead of $\mathcal{K}(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta})$, we use the less common notation $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in the Bayesian sense, as the distribution of $\boldsymbol{\theta}^{(t+1)}$ is conditional on the present value of $\boldsymbol{\theta}^{(t)}$.

The distribution of a time-homogeneous Markov chain $\{\boldsymbol{\theta}^{(t)}\}$ is affected by the transition kernel and the initial state (or distribution of the initial state). Of key importance is that the distribution of the sampler's state converges to the correct invariant distribution, regardless of the starting state $\boldsymbol{\theta}^{(0)}$. The influence of the starting state $\boldsymbol{\theta}^{(0)}$ on $\boldsymbol{\theta}^{(t)}$ affects the ergodicity of the chain, and our hope is that the initial state will gradually be forgotten. If the chain is *ergodic* the invariant distribution should be reachable from any initial distribution. The distribution $\mathcal{K}^{(t)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$—the distribution of $\boldsymbol{\theta}^{(t)}$ given the starting value $\boldsymbol{\theta}^{(0)}$—should converge to the stationary distribution as $t \to \infty$. The tth iterate of transition kernel is recursively defined as

$$\mathcal{K}^{(t)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) = \int \mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')\mathcal{K}^{(t-1)}(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(0)})\, d\boldsymbol{\theta}' \ . \tag{5.3}$$

Under certain regularity conditions the initial state will have no effect on the long-term outcome of the chain, and $\mathcal{K}^{(t)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ will converge to a stationary or invariant distribution, independent of the initial state or t. A brief overview of these conditions is given below.

After a period of, say, m samples, the samples $\{\boldsymbol{\theta}^{(t)}\}_{t=m+1}^{T}$ will be a set of dependent samples that approximately come from the target density. This initial set of samples is called the *burn-in*, and is usually discarded when estimators are determined, i.e. $\hat{\Phi}_T = \frac{1}{T-m}\sum_{t=m+1}^{T}\phi(\boldsymbol{\theta}^{(t)})$.

Conditions guaranteeing convergence to a stationary distribution

A few conditions are sufficient to ensure that the distribution of the state of a Markov chain converges to the invariant distribution. These conditions are needed for the law of large numbers to hold for sample path averages, such that the empirical estimate $\hat{\Phi}$ converges to the true expectation Φ. As we will later see with a toy example in section 5.4.1, we may need *additional* conditions to ensure a central limit theorem and say something meaningful about the variance of the estimate.

We first define the 'distance' between two probability distributions p_1 and p_2 with the total variation norm,

$$\|p_1 - p_2\| = 2 \sup_{A\in\Theta} |p_1(A) - p_2(A)| \ , \tag{5.4}$$

where we use notation $p(A) = \int_A p(\boldsymbol{\theta})d\boldsymbol{\theta}$. We also need the concept of the first return time of the Markov chain to set A, indicated by τ_A, and properly defined as $\tau_A = \inf\{t \geq 1 : \boldsymbol{\theta}^{(t)} \in A\}$, with $\tau_A = \infty$ if the chain never returns to A.

For the Markov chain to converge to its stationary distribution the chain must be:

Irreducible. Given any initial state, the chain has to be able to reach any other state with a positive probability in a finite number of steps. For general state-space Markov chains, the definition of irreducibility is with respect to a distribution (Gilks et al., 1996):

Definition 1. *A Markov chain is φ-irreducible for a probability distribution φ on $\boldsymbol{\Theta}$ if $\varphi(A) > 0$ for a set $A \subset \boldsymbol{\Theta}$ implies that*

$$\mathbb{P}(\tau_A < \infty \mid \boldsymbol{\theta}^{(0)}) > 0 \tag{5.5}$$

for all $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$. A chain is irreducible if it is φ-irreducible for some probability distribution φ. If a chain is φ-irreducible, then φ is called an irreducibility distribution for the chain.

An irreducible chain has many irreducibility distributions, all of which are absolutely continuous with respect to some maximal irreducibility distribution ψ. (Saying that p_1 is absolutely continuous with respect to p_2 here means that if $p_2(A) > 0$, then $p_1(A) > 0$, for any $A \in \boldsymbol{\Theta}$.)

Aperiodic. The chain must not by cyclic, so that an oscillation between two states, or sets of states, is for example not possible.

Recurrent. If a chain is irreducible then all interesting sets can be reached. Recurrence implies that all such sets can be reached infinitely often, from all starting positions. A distinction is made between *positive* recurrent chains, where the average return time to all states is finite, and *null* recurrent chains, where the average time to return to some state can be infinite. For discrete chains positive recurrence is needed, and follows from the existence of an invariant distribution. For general state-space chains we have the following definition (Gilks et al., 1996):

Definition 2. *An irreducible Markov chain with maximal irreducibility distribution ψ is recurrent if for any set $A \subset \boldsymbol{\Theta}$ with $\psi(A) > 0$*

1. $\mathbb{P}(\boldsymbol{\theta} \in A$ *infinitely often* $\mid \boldsymbol{\theta}^{(0)}) > 0$ *for all* $\boldsymbol{\theta}^{(0)}$*, and*
2. $\mathbb{P}(\boldsymbol{\theta} \in A$ *infinitely often* $\mid \boldsymbol{\theta}^{(0)}) = 1$ *for ψ-almost all* $\boldsymbol{\theta}^{(0)}$*.*

An irreducible recurrent chain is positive recurrent if it has an invariant distribution, otherwise it is null recurrent.

As the chain is started from any initial $\boldsymbol{\theta}^{(0)}$, we need to be sure that the chain has the same limiting behaviour for *every* starting value instead of *almost* every starting value. This is ensured by *Harris recurrence*, a stronger condition which requires $\mathbb{P}(\boldsymbol{\theta} \in A$ infinitely often $\mid \boldsymbol{\theta}^{(0)}) = 1$ for *all* $\boldsymbol{\theta}^{(0)}$.

If we can show that the Markov chain $\{\boldsymbol{\theta}^{(t)}\}$ is irreducible and has invariant distribution $p^*(\boldsymbol{\theta})$, then the chain is p^*-irreducible, $p^*(\boldsymbol{\theta})$ is a maximal irreducibility distribution, $p^*(\boldsymbol{\theta})$ is the unique invariant distribution of the chain, and the chain is positive recurrent. Recurrence is sufficient to imply convergence of averages of probabilities, and to let a strong law of large numbers hold (Gilks et al., 1996):

Theorem 3. *If $\{\boldsymbol{\theta}^{(t)}\}$ is an irreducible Markov chain with transition kernel $\mathcal{K}$ and invariant distribution p^*, and $\phi(\boldsymbol{\theta})$ a real-valued function such that $\int |\phi(\boldsymbol{\theta})| p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$. Then $\mathbb{P}(\hat{\Phi}_T \to \Phi \mid \boldsymbol{\theta}^{(0)}) = 1$ for p^*-almost all $\boldsymbol{\theta}^{(0)}$, where $\hat{\Phi}_T$ and Φ are defined in (5.2) and (5.1).*

Whether a Markov chain will converge to its invariant distribution is given by the following result:

Theorem 4. *Suppose $\{\boldsymbol{\theta}^{(t)}\}$ is an irreducible, aperiodic Markov chain with transition kernel $\mathcal{K}$ and invariant distribution p^*. Then*

$$\|\mathcal{K}^{(t)}(\cdot|\boldsymbol{\theta}) - p^*(\cdot)\| \to 0 \tag{5.6}$$

for p^*-almost all $\boldsymbol{\theta}$.

This convergence result will hold for *all* $\boldsymbol{\theta}$ if and only if we can guarantee that the chain is positive Harris recurrent as well. (Harris recurrence will hold for the samplers that we are concerned with in this chaper.) For a positive Harris recurrent chain, asymptotic results, such as laws of large numbers and central limit theorems, that do not depend on any initial portion of a sample path can be shown to hold for all initial distributions if they hold for any.

Central Limit Theorems

Thus far we have seen that positive recurrence alone is sufficient to ensure that a law of large numbers holds for a Markov chain, and hence that ergodic averages converge to their expectations under the stationary distribution. Ergodic averages $\hat{\Phi}_T$, and the asymptotic properties of these averages, are clearly very important. The ergodic theorem 3 does however not specify (a) how long we need to run the chain for, and (b) it gives no estimate of the size of error that the estimate $\hat{\Phi}_T$ makes.

We need stronger conditions for theorem 4 besides recurrence to provide a central limit theorem. A condition that is often used, also in this chapter, is that of geometric convergence for *ergodic* chains, that is, chains that are irreducible, aperiodic and positive Harris recurrent.

Definition 3. *An ergodic Markov chain with invariant distribution $p^*(\boldsymbol{\theta})$ is geometrically ergodic if there exists a non-negative extended real-valued function M such that $\int M(\boldsymbol{\theta})p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$ and a positive constant $r < 1$ such that*

$$\|\mathcal{K}^{(t)}(\cdot|\boldsymbol{\theta}) - p^*(\cdot)\| \leq M(\boldsymbol{\theta})r^t \tag{5.7}$$

for all $\boldsymbol{\theta}$ and all t.

If function M does not depend on $\boldsymbol{\theta}$ in the above definition, i.e., it is constant, then the stronger condition gives *uniform convergence*. Importantly, a Markov chain that is geometrically or uniformly ergodic satisfies a central limit theorem (Gilks et al., 1996):

Theorem 5. *Suppose an ergodic Markov chain $\{\boldsymbol{\theta}^{(t)}\}$ with invariant distribution p^* and a real valued function ϕ satisfy one of the following conditions:*

1. *The chain is geometrically ergodic and $\int |\phi(\boldsymbol{\theta})|^{2+\epsilon}p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$ for some $\epsilon < 0$.*
2. *The chain is uniformly ergodic and $\int \phi(\boldsymbol{\theta})^2 p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$.*

Then

$$\sigma_\phi^2 = \mathbb{E}_{p^*}[(\phi(\boldsymbol{\theta}^{(0)}) - \Phi)^2] + 2\sum_{t=1}^{T} \mathbb{E}_{p^*}[(\phi(\boldsymbol{\theta}^{(0)}) - \Phi)(\phi(\boldsymbol{\theta}^{(t)}) - \Phi)] \tag{5.8}$$

is well defined, non-negative and finite, and $\sqrt{T}(\hat{\Phi}_T - \Phi)$ converges in distribution to a $\mathcal{N}(0, \sigma_\phi^2)$ random variable.

(In the above theorem the esimator $\hat{\Phi}_T = \frac{1}{T+1}\sum_{t=0}^{T} \phi(\boldsymbol{\theta}^{(t)})$ counts from zero).

The notion of geometric convergence will reoccur in section 5.4.1, when we observe that the variational kernel, that will later be introduced, gives 'sticky tails'. Geometric convergence will be the key to showing that simply sampling with a variational proposal cannot, in a simple case examined, give a Central Limit Theorem. Before considering variational approximations to transition kernels, a short overview of the Metropolis Hastings algorithm is given.

5.2.3 Metropolis-Hastings

Markov chain Monte Carlo methods are constructed such that the stationary condition of detailed balance holds,

$$p^*(\boldsymbol{\theta}^{(t)})\mathcal{K}(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = p^*(\boldsymbol{\theta}^{(t+1)})\mathcal{K}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)}) \ . \tag{5.9}$$

As usual we let $p^*(\boldsymbol{\theta})$ be a (possibly unnormalised) shorthand for $p(\boldsymbol{\theta}|\mathbf{x})$.

Constructing a Markov chain with $p^*(\boldsymbol{\theta})$ as stationary distribution is very easy with the method of Metropolis et al. (1953), later generalized by Hastings (1970). At each time t, a new state $\boldsymbol{\theta}^{\text{new}}$ is generated from a proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. This new state is a candidate that is being accepted with probability

$$\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{\text{new}}) = \min\left(1, \frac{p^*(\boldsymbol{\theta}^{\text{new}})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\text{new}})}{p^*(\boldsymbol{\theta}^{(t)})q(\boldsymbol{\theta}^{\text{new}}|\boldsymbol{\theta}^{(t)})}\right) . \tag{5.10}$$

If the proposed state is accepted the next state in the chain becomes $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{\text{new}}$, otherwise the state of the chain does not change with $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$. An overview is given in algorithm 3.

When the support of q includes the support of $p^*(\boldsymbol{\theta})$, the resulting transition kernel

$$\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + [1 - \mathsf{acc}(\boldsymbol{\theta}^{(t)})]\delta(\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}) \tag{5.11}$$

satisfies the detailed balance condition and the stationary distribution of the chain will be $p^*(\boldsymbol{\theta})$. The transition kernel consists of two terms, the first is the probability of generating a new point multiplied by the probability of accepting it, and the second is the probability of repeating the previous sample $\boldsymbol{\theta}^{(t)}$. Notation $\mathsf{acc}(\boldsymbol{\theta}^{(t)}) = \int \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})d\boldsymbol{\theta}$ indicates the probability of accepting a new point, while $\delta(\cdot = \boldsymbol{\theta}^{(t)})$ indicates the Dirac delta mass at $\boldsymbol{\theta}^{(t)}$.

From (5.10) we have

$$p^*(\boldsymbol{\theta}^{(t)})q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) = p^*(\boldsymbol{\theta}^{(t+1)})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)})\alpha(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) , \tag{5.12}$$

from which we obtain the detailed balance equation

$$p^*(\boldsymbol{\theta}^{(t)})\mathcal{K}(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = p^*(\boldsymbol{\theta}^{(t+1)})\mathcal{K}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)}) . \tag{5.13}$$

When we integrate both sides of this equation with respect to $\boldsymbol{\theta}^{(t)}$ we get the stationary distribution condition

$$\int p^*(\boldsymbol{\theta}^{(t)})\mathcal{K}(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})\, d\boldsymbol{\theta}^{(t)} = p^*(\boldsymbol{\theta}^{(t+1)}) , \tag{5.14}$$

and $\mathcal{K}$ is the correct transition kernel. Therefore, the equation says that if $\boldsymbol{\theta}^{(t)}$ comes from $p^*(\boldsymbol{\theta}^{(t)})$, then $\boldsymbol{\theta}^{(t+1)}$ will come from the same stationary distribution. Once we have a sample from the stationary distribution, all subsequent samples will also be from that distribution. To fully justify the MH algorithm we need more than just a stationary distribution, we also need a guarantee that $\mathcal{K}^{(t)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ will converge to the stationary distribution. From construction of the MH method, we already have a stationary distribution, while aperiodicity follows from the fact that the MH algorithm allows for rejection. Finally, for irreducibility we only need to ensure that $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) > 0$ over the entire space.

Whether the algorithm succeeds in exploring all modes of parameter space often depends on the choice of proposal distribution. If the proposal is too narrow, the chain may only explore one mode, while if on the other hand it is too wide, the rejection rate can be very high. We have seen in chapter 4 how parallel tempering, with fast relaxation at high temperature simulations, can be used to ensure that the chain mixes well and all modes be visited with high acceptance probability.

We will now explore a new route, leading from the cross roads of approximate inference from chapters 2 and 3, and MCMC methods discussed earlier in this chapter. In the spirit of variational methods, an approximation to a 'good' transition kernel will be made, and a detailed discussion and proof given illustrating why such an approximation may lead to failure of a MCMC method.

Algorithm 3 Metropolis-Hastings

initialize: $\boldsymbol{\theta}^{(0)}$, $t = 0$.
repeat
  sample $\boldsymbol{\theta}^{\text{new}}$ from $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.
  sample a uniform random variable $u \sim \mathcal{U}(0,1)$.
  if $u \leq \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{\text{new}})$ **then**
    set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{\text{new}}$
  else
    set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$
  end if
  $t \leftarrow t + 1$
until $t = t_{\max}$

5.3 Variational transition kernel

A natural setting for variational transition kernel sampling arises when the posterior can be completed with latent variables $\mathbf{z}$ into a joint distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$, such that the completed posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$ becomes easy to evaluate (Robert & Casella, 2004).

For completeness the familiar example of chapters 2, 3 and 4 is repeated here. In mixture of distributions, we assume that the examples in $\mathbf{x} = \{\mathbf{x}_n\}_{n=1}^N$ are independent and identically drawn from $p(\mathbf{x}_n|\boldsymbol{\theta}) = \sum_{j=1}^J \pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j)$. Let $\boldsymbol{\theta}$ encompass all unknown parameters in the model: the parameters of the J component distributions $p(\cdot|\boldsymbol{\theta}_j)$ and the component weights π_j that sum to one. Hidden latent variables $\mathbf{z} = \{z_{nj}\}$, where z_{nj} is equal to 1 if $\mathbf{x}_n$ was generated from component j in the mixture, and zero otherwise, naturally augment the data. The likelihood, which considers all possible partitions of the sample and expands into J^N terms, and the 'easier' complete-data likelihood are

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{n=1}^N \Big[\sum_{j=1}^J \pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j)\Big] \quad \text{and} \quad p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \prod_{n=1}^N \prod_{j=1}^J \Big[\pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j)\Big]^{z_{nj}} . \tag{5.15}$$

5.3.1 An exact transition kernel

We can draw samples from the posterior distribution over $\boldsymbol{\theta}$ if we can construct a transition kernel $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ that has the parameter posterior as correct invariant distribution,

$$\int \mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}') p(\boldsymbol{\theta}'|\mathbf{x})\, d\boldsymbol{\theta}' = p(\boldsymbol{\theta}|\mathbf{x}) , \tag{5.16}$$

and that marginalizes out the latent variables. Samples from $\mathcal{K}$ are then coming from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$, and we can create such a sample $\{\boldsymbol{\theta}^{(t)}\}$ by defining $\mathcal{K}$ with:

1. Given $\boldsymbol{\theta}'$ (i.e. $\boldsymbol{\theta}^{(t-1)}$), determine $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}')$.
2. Given $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}')$, sample $\boldsymbol{\theta}$ (i.e. $\boldsymbol{\theta}^{(t)}$) from

$$\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}')\, d\mathbf{z} . \tag{5.17}$$

This kernel is exact as we can show that $p(\boldsymbol{\theta}|\mathbf{x})$ is the invariant distribution with a rearrangement,

$$\begin{aligned}\int \mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')p(\boldsymbol{\theta}'|\mathbf{x})\,d\boldsymbol{\theta}' &= \int \left\{ \int p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')\,d\mathbf{z} \right\} p(\boldsymbol{\theta}'|\mathbf{x})\,d\boldsymbol{\theta}' \\ &= \int \left\{ \int \frac{p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x},\mathbf{z})} \frac{p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta}')}{p(\mathbf{x}|\boldsymbol{\theta}')}\,d\mathbf{z} \right\} \frac{p(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{x})}\,d\boldsymbol{\theta}' \\ &= \int \frac{p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x},\mathbf{z})} \int \frac{p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta}')}{p(\mathbf{x}|\boldsymbol{\theta}')} \frac{p(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{x})}\,d\boldsymbol{\theta}'\,d\mathbf{z} \\ &= \frac{1}{p(\mathbf{x})} \int p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\mathbf{z} = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} = p(\boldsymbol{\theta}|\mathbf{x})\ , \end{aligned} \tag{5.18}$$

and therefore if the kernel is used as proposal density in a MH algorithm, the acceptance ratio in (5.10) will always be unity.

This construction is similar to doing Gibbs sampling with $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')$ and $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})$. We arrive at an *alternative* sampling scheme, as choosing $\mathcal{K}$ in this way to average over the latent variable space allows a derivation of a transition kernel that replaces discrete or categorical model changes with continuous ones.

5.3.2 A tractable approximation

The required transition kernel is, however, in general not analytically tractable and impossible to sample from directly. The kernel can be approximated by finding a proposal density $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ that is close to $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ (remembering that sampling with $\mathcal{K}$ automatically gives the required invariant distribution and can in theory be used as a proposal density, although in practice it is not analytically tractable) by minimizing the Kullback-Leibler divergence between the defined proposal and the kernel (exact proposal),

$$\mathsf{KL}(q\|\mathcal{K}) = \int q(\boldsymbol{\theta}|\boldsymbol{\theta}') \ln \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')}\,d\boldsymbol{\theta}\ . \tag{5.19}$$

The problem that we face is that $\mathcal{K}$ is not in a product form (unlike $p(\boldsymbol{\theta},\mathbf{z}|\mathbf{x})$, for example). This means that even if a specific factorization of q is assumed, a tractable solution will not be found. However, Jensen's inequality can be used to construct an upper bound to $\mathsf{KL}(q\|\mathcal{K})$, and this bound can in turn be minimized to give an analytical solution:

$$\ln \mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}') \geq \int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}') \ln p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})\,d\mathbf{z}\ . \tag{5.20}$$

It is important to note that this is a bound for the distribution in $\mathbf{z}$, so the equality only holds in the extreme cases of deterministic relation or independence: $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}') = \delta(\mathbf{z}-\mathbf{z}_0)$ or $p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z}) = p(\boldsymbol{\theta}|\mathbf{x})$. We now have an upper bound:

$$\mathsf{KL}(q\|\mathcal{K}) \leq \int q(\boldsymbol{\theta}|\boldsymbol{\theta}') \Big[\ln q(\boldsymbol{\theta}|\boldsymbol{\theta}') - \int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}') \ln p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})\,d\mathbf{z} \Big]\,d\boldsymbol{\theta} \equiv \mathcal{F}[q(\boldsymbol{\theta}|\boldsymbol{\theta}')]\ . \tag{5.21}$$

The minimum of $\mathsf{KL}(q\|\mathcal{K})$, which is $q = \mathcal{K}$, is of course not attained by minimizing the upper bound $\mathcal{F}$ (apart from the trivial case discussed above). The solution can, as we see in the following, still serve as a useful approximation to the transition kernel.

The usual free form optimization, typical of variational methods, can be used to find an appropriate q. A Lagrange multiplier ℓ is added to the functional such that q is constrained to integrate to one, with

$$\tilde{\mathcal{F}}[q(\boldsymbol{\theta}|\boldsymbol{\theta}')] = \mathcal{F}[q(\boldsymbol{\theta}|\boldsymbol{\theta}')] + \ell\Big[\int q(\boldsymbol{\theta}|\boldsymbol{\theta}')\, d\boldsymbol{\theta} - 1\Big] \ . \tag{5.22}$$

By using elementary calculus of variations, we take the functional derivative of $\mathcal{F}[q]$ with respect to q under the integral constraint to make q a density,

$$\begin{aligned}\frac{\partial\tilde{\mathcal{F}}[q(\boldsymbol{\theta}|\boldsymbol{\theta}')]}{\partial q(\boldsymbol{\theta}|\boldsymbol{\theta}')} &= \int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')\Big[\frac{\partial}{\partial q(\boldsymbol{\theta}|\boldsymbol{\theta}')}\int q(\boldsymbol{\theta}|\boldsymbol{\theta}')\ln\frac{p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})}{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}\, d\boldsymbol{\theta}\Big]\, d\mathbf{z} \\ &\quad \cdots + \frac{\partial}{\partial q(\boldsymbol{\theta}|\boldsymbol{\theta}')}\ell\Big[\int q(\boldsymbol{\theta}|\boldsymbol{\theta}')\, d\boldsymbol{\theta} - 1\Big] \\ &= \int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')[\ln p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z}) - \ln q(\boldsymbol{\theta}|\boldsymbol{\theta}') - 1]\, d\mathbf{z} + \ell \ ,\end{aligned} \tag{5.23}$$

and set it to zero:

$$\ln q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')\ln p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})\, d\mathbf{z} - 1 + \ell \ . \tag{5.24}$$

By exponentiating and integrating over $\boldsymbol{\theta}$ on both sides, keeping the integral constraint in mind, we can solve for ℓ to correctly normalize the distribution. The upper bound on the divergence is therefore minimized with

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \exp\Big\{\int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')\ln p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})\, d\mathbf{z}\Big\}\Big/\int \exp\Big\{\int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')\ln p(\boldsymbol{\theta}|\mathbf{x},\mathbf{z})\, d\mathbf{z}\Big\} d\boldsymbol{\theta} \ . \tag{5.25}$$

Using Bayes' theorem we get the variational approximation to the transition kernel,

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}') \propto p(\boldsymbol{\theta})\exp\Big\{\int p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')\ln p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})\, d\mathbf{z}\Big\} \ . \tag{5.26}$$

This form is instantly recognizable from the Variational Bayesian EM algorithm, as discussed in section 1.4.3.

5.3.3 Illustrative example: Mixture of distributions

As an illustrative example, consider the mixture of distributions discussion from the start of Section 5.3. Let $p(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\theta}') = \boldsymbol{\gamma}_n = (\gamma_{n1},\ldots,\gamma_{nJ})$ be the probability that $\mathbf{x}_n$ was generated by each of the J components (the component's *responsibility*), thus defining a distribution over each binary vector $\mathbf{z}_n$. Then the responsibilities are

$$\gamma_{nj} = p(z_{nj} = 1|\mathbf{x}_n,\boldsymbol{\theta}') = \frac{\pi_j p(\mathbf{x}_n|\boldsymbol{\theta}'_j)}{\sum_{k=1}^{J}\pi_k p(\mathbf{x}_n|\boldsymbol{\theta}'_k)} \ . \tag{5.27}$$

Using the log of the complete-data likelihood from equation (5.15) and $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}')$ determined above, we get $\sum_{\mathbf{z}}\ln p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}') = \sum_{n=1}^{N}\sum_{j=1}^{J}\gamma_{nj}\ln[\pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j)]$, and from equation (5.26),

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}') \propto p(\boldsymbol{\theta})\prod_{j=1}^{J}\prod_{n=1}^{N}\Big[\pi_j p(\mathbf{x}_n|\boldsymbol{\theta}_j)\Big]^{\gamma_{nj}}. \tag{5.28}$$

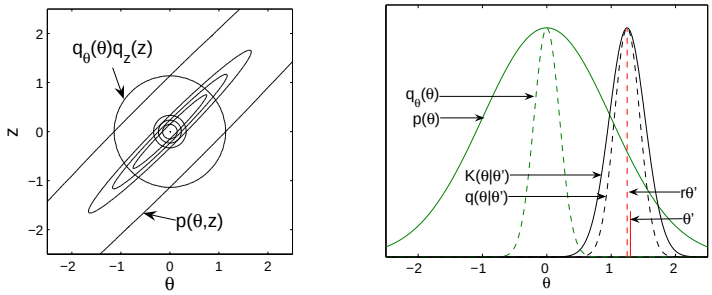


FIGURE 5.1: A toy example of a two-dimensional Gaussian over θ and z, centered at zero, with covariance matrix $\Sigma = [1, 0.98; 0.98, 1]$. The variational Bayes approximation $q_\theta(\theta)q_z(z)$ to $p(\theta, z)$ is shown on the *left*. The marginal $p(\theta)$ and the factor $q_\theta(\theta)$ from the variational approximation to $p(\theta, z)$ is on the *right*. With $\theta' = 1.3$, $\mathcal{K}(\theta|\theta')$ and $q(\theta|\theta')$, both with mean $r\theta'$, are also shown. (The densities are not normalised in the figure.)

The kernel simply raises the likelihood terms to its respective responsibility powers, and if $p(\cdot|\boldsymbol{\theta}_j)$ comes from an exponential family of distributions the result will be a weighted contribution of data points to each mixture component. Notice that if γ_{nj} is discrete, we fall back directly to two-stage Gibbs sampling, where a discrete sample $z_{nj} \in \{0, 1\}$ is used instead of $\gamma_{nj} \in [0, 1]$. Effectively we have replaced categorical labels with continuous variables. When γ is strictly binary the kernel approximation is of course exact.

A potential difficulty surfacing here is that the true kernel may be *multimodal*, but we attempt to approximate it with a distribution of single mode, as was seen in examples in chapter 2. As is true for variational methods where a factorization $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})q_{\mathbf{z}}(\mathbf{z})$ is assumed, the variance of the approximation—similar to the variance of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ from a variational lower bound—is an underestimation of the true variance of the kernel. In approximate methods this is acceptable, but it may be ruinous when we are concerned with creating a practical MCMC method (see section 5.4.1 for a formal discussion).

5.4 Using the proposal in Monte Carlo methods

Although the main idea of taking a Markov chain with an intractable transition kernel $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ and correct invariant distribution, and approximating $\mathcal{K}$ with $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ as a MH proposal, sounds appealing, some questions on convergence can be raised. It turns out that simply using q as proposal is not enough, and the method needs to be supplemented with additional standard moves (symmetric proposal densities, for instance), or dropped altogether in favour of more orthodox samplers (the Gibbs sampler of chapter 4 being a good example). The following toy example is insightful.

5.4.1 Toy example

As an informal illustration to compare $\mathcal{K}(\theta|\theta')$, the approximation $q(\theta|\theta')$ as MH proposal density, and the variational approximation $q_\theta(\theta)$, consider a two-dimensional Gaussian, where ρ_{ij} indicate the elements of the inverse covariance matrix,

$$p(\theta, z) \propto \exp\left\{ -\tfrac{1}{2}[\rho_{11}\theta^2 + (\rho_{12} + \rho_{21})\theta z + \rho_{22}z^2]\right\} . \tag{5.29}$$

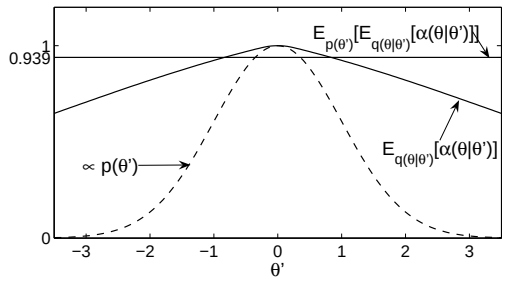


FIGURE 5.2: Continuing Figure 5.1's toy example. The average acceptance probability $\mathbb{E}[\alpha(\theta|\theta')]$ is shown as a function of θ'. Averaging $\mathbb{E}[\alpha(\theta|\theta')]$ numerically over $p(\theta')$ gives 0.939 as average acceptance over the target distribution.

Let

$$r \equiv \frac{(\rho_{12}+\rho_{21})^2}{4\rho_{11}\rho_{22}} \ , \tag{5.30}$$

with $r \in [0,1)$. The example is chosen such that $\mathcal{K}(\theta|\theta')$ from (5.17) is analytically tractable, and also such that the required marginal $p(\theta)$—that we would like to draw samples from—can be analytically normalized. The marginal and conditional distributions for θ are

$$p(\theta) = \mathcal{N}\big(\theta \mid 0, [\rho_{11}(1-r)]^{-1}\big) \tag{5.31}$$

$$\text{and} \quad p(\theta|z) = \mathcal{N}\Big(\theta \,\Big|\, -\frac{(\rho_{12}+\rho_{21})}{2\rho_{11}} z, \rho_{11}^{-1}\Big) \ . \tag{5.32}$$

The marginal $p(z)$ and conditional $p(z|\theta)$ are the same as above, except that occurrences of ρ_{11}, ρ_{22} and z are respectively swapped with ρ_{22}, ρ_{11} and θ. Here both $\mathcal{K}(\theta|\theta') = \int p(\theta|z)p(z|\theta')\,dz$ and the approximation $q(\theta|\theta') = \frac{1}{Z}\exp\{\ln p(\theta|z)p(z|\theta')\,dz\}$ are analytically tractable,

$$\mathcal{K}(\theta|\theta') = \mathcal{N}\big(\theta \mid r\theta', \rho_{11}^{-1}(1+r)\big) \tag{5.33}$$

$$q(\theta|\theta') = \mathcal{N}\big(\theta \mid r\theta',\ \rho_{11}^{-1}\big) \ . \tag{5.34}$$

A further possible approximation to note is that given by Variational Bayes, where $p(\theta, z)$ is approximated with a factorized Gaussian $q_\theta(\theta)q_z(z)$, the minimizer of $\mathsf{KL}(q_\theta(\theta)q_z(z) \parallel p(\theta, z))$. The factor involving θ is

$$q_\theta(\theta) = \mathcal{N}\big(\theta \mid 0, \rho_{11}^{-1}\big) \ , \tag{5.35}$$

which can naively be taken as a proposal density. Figure 5.1 plots all of these densities, showing that an adaptive proposal $q(\theta|\theta')$ appears more sensible (in terms of exploration) than using $q_\theta(\theta)$ as proposal. The variational (approximate) kernel $q(\theta|\theta')$ illustrates the potential underestimation of the variance of $\mathcal{K}(\theta|\theta')$: here by a factor of $(1+r)$. How much the variance is underestimated depends on the strength of coupling between θ and z, with $q = \mathcal{K}$ when $r = 0$. When the $q(\theta|\theta')$ is used as a MH proposal density, $\mathbb{E}[\alpha(\theta|\theta')]$ is in practice close to one in areas of high density and good approximation (see Figure 5.2 with regard to this example). However, due to the underestimation of the variance by the variational approximation, the variational proposal may fail to produce adequate results at the tails of a distribution. A similar effect occurs when the exact kernel is multimodal, and is approximated with a distribution with single mode. This problem is discussed in more detail in the following section.

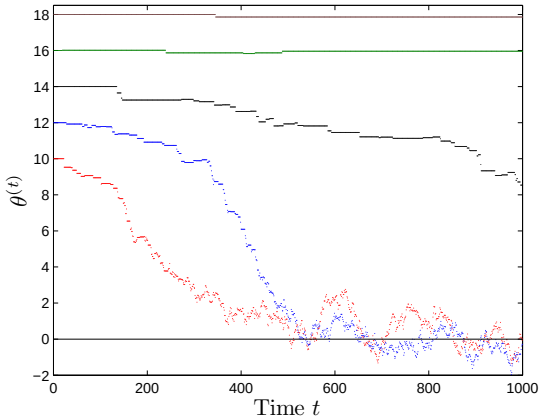


FIGURE 5.3: An illustration of why the approximation $q(\theta|\theta')$ may not give a geometrically ergodic chain. Different chains are started further down the tails of the target distribution $p(\theta)$, and the average acceptance probability can be made *arbitrarily* small by choosing a $\theta^{(t)}$ far enough down the tail of the distribution. In this example the covariance matrix $\Sigma = [1, 0.98; 0.98, 1]$ was used, giving $\rho_{11} = \rho_{22} = 25.2525$ and $\rho_{12} = \rho_{21} = -24.7475$. The standard deviation of the target density $p(\theta)$ is 1; the chains were started at least 10 standard deviations away from $p(\theta)$'s mean.

Geometric ergodicity

To provide for a central limit theorem on the estimate $\hat{\Phi}_T$, i.e. to have $\sqrt{T}(\hat{\Phi}_T - \Phi)$ converge in distribution to a Gaussian zero-mean random variable with *finite*, well-defined variance, a condition on the convergence rate of $\{\boldsymbol{\theta}^{(t)}\}$ is typically needed. An often-considered convergence rate condition is *geometric ergodicity*.

Continuing the toy example, figure 5.3 illustrates the use of $q(\theta|\theta')$ as proposal, with different chains starting further down the tails of the target distribution $p(\theta)$. It will be shown later that the tails become 'sticky', with arbitrarily small average acceptance probability—this forms the core of showing that the chain is not geometrically ergodic. Section 5.4.3 presents a simple method for addressing this problem.

Preliminary theory

We want to be able to show that a chain is geometrically ergodic, or not, and here only the necessary theory for the proof is given. For general state-space chains, we need the notion of a small set. A small set will play a role similar to individual states in a discrete chain, and in many problems under consideration compact sets are small.

Definition 4. *A set $C \subset \boldsymbol{\Theta}$ is small for an irreducible transition kernel $\mathcal{K}$ with maximal irreducibility distribution ψ if $\psi(C) > 0$ and there exists a probability distribution ν on $\boldsymbol{\Theta}$, a non-negative integer t, and a constant $\beta > 0$ such that $\mathcal{K}^{(t)}(A|\boldsymbol{\theta}) \geq \beta\nu(A)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $A \subset \boldsymbol{\Theta}$.*

The following theorem is adapted from (Mengersen & Tweedie, 1996) and will be used in showing geometric ergodicity.

Theorem 6. *Suppose that $\{\boldsymbol{\theta}^{(t)}\}$ is φ-irreducible and aperiodic. Then the following is an equivalent definition for a geometrically ergodic chain:*

- *For some small set C with $\varphi(C) > 0$, there exists a $\kappa > 1$ such that*

$$\sup_{\boldsymbol{\theta} \in C} \mathbb{E}_{\boldsymbol{\theta}}[\kappa^{\tau_C}] = \sup_{\boldsymbol{\theta} \in C} \sum_{k=0}^{\infty} \mathbb{P}[\tau_C \geq k]\kappa^k < \infty \ . \tag{5.36}$$

(The notation $\mathbb{E}_{\boldsymbol{\theta}}$ indicates the expectation for a Markov chain started in state $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta} \in C$.)

We are now in the position to pinpoint where the use of the variational approximation to the kernel may fail, and the proof merely gives a formal description of figure 5.3.

Proposition 1. *Provided that $r > 0$, the toy example's proposal density $q(\theta|\theta')$ does not give rise to a geometrically ergodic Metropolis Hastings chain.*

Proof. Consider some θ'. The acceptance ratio, as a function of θ', is an exponential that is symmetric around zero, the mean of $p(\theta)$,

$$A = \exp\{\tfrac{1}{2}\rho_{11} r(1-r)(\theta^2 - \theta'^2)\} \ , \tag{5.37}$$

with $\rho_{11} r(1-r)$ being nonnegative. Given θ', a proposed θ is accepted with probability $\alpha(\theta', \theta) = \min(1, A)$. The average acceptance probability, as a function of θ', is

$$a(\theta') \equiv \mathbb{E}[\alpha(\theta', \theta)] = \int_{|\theta|<|\theta'|} A\mathcal{N}(\theta \mid r\theta',\, \rho_{11}^{-1})\, d\theta + \int_{|\theta|\geq|\theta'|} 1\mathcal{N}(\theta \mid r\theta',\, \rho_{11}^{-1})\, d\theta \ , \tag{5.38}$$

and we can show that it can be made arbitrarily small by choosing θ' further away from zero. Evaluating $a(\theta')$ for $\theta' > 0$ gives

$$\begin{aligned} a(\theta') = {} & \Big(\frac{\rho_{11}}{2\pi}\Big)^{1/2} \exp\Big\{ -\frac{1}{2}\rho_{11} r\big(1 - \frac{r}{r^2 - r + 1}\big)\theta'^2 \Big\} \\ & \cdots \times \underbrace{\int_{|\theta|<|\theta'|} \exp\Big\{ -\frac{1}{2}\rho_{11}[r^2 - r + 1]\big(\theta - \frac{r\theta'}{r^2 - r + 1}\big)^2 \Big\}\, d\theta}_{\leq (2\pi/\rho_{11}[r^2 - r + 1])^{1/2}} \\ & \cdots + \Phi\Big(-\theta'(1+r)\sqrt{\rho_{11}}\Big) + \Big[1 - \Phi\Big(\theta'(1-r)\sqrt{\rho_{11}}\Big)\Big] \ , \end{aligned} \tag{5.39}$$

where $\Phi(\cdot)$ is the $\mathcal{N}(0,1)$ cumulative density function, and arises from integrating the two tails given by the second term in (5.38). Simplifying further, we get

$$\begin{aligned} a(\theta') \leq {} & \underbrace{[r^2 - r + 1]^{-1/2}}_{1 \leq \cdot \leq \sqrt{4/3}} \exp\Big\{ -\frac{1}{2}\rho_{11} r\big(1 - \frac{r}{r^2 - r + 1}\big)\theta'^2 \Big\} \\ & \cdots + \Phi\Big(-\theta'(1+r)\sqrt{\rho_{11}}\Big) + \Big[1 - \Phi\Big(\theta'(1-r)\sqrt{\rho_{11}}\Big)\Big] \to 0 \quad \text{as} \quad \theta' \to \infty \ . \end{aligned} \tag{5.40}$$

We can construct a similar upper bound on $a(\theta')$ when $\theta' < 0$, and show that $a(\theta') \to 0$ as $\theta' \to -\infty$.

Now suppose that $\{\theta^{(t)}\}$ is geometrically ergodic. Then for some small set C, there exists a $\kappa > 1$ such that

$$\sup_{\theta^{(0)} \in C} \mathbb{E}_{\theta^{(0)}}[\kappa^{\tau_C}] = \sup_{\theta^{(0)} \in C} \sum_{k=0}^{\infty} \mathbb{P}(\tau_C \geq k)\kappa^k < \infty \ , \tag{5.41}$$

where τ_C denotes the return time to C. Define a set of states from which the probability of accepting a proposal is smaller than $1 - 1/\kappa$,

$$D_\kappa = \{\theta' \,:\, a(\theta') < 1 - 1/\kappa\} \ , \tag{5.42}$$

which has positive measure for any given $\kappa > 1$ because $a(\theta') \to 0$ as $\theta' \to \infty$. The proof hinges on this: we can choose θ' such that the average chance of accepting a proposed θ is arbitrarily small.

Define θ_κ as the element in D_κ that gives rise to the largest probability of accepting a proposed sample, $\theta_\kappa = \arg\max_{\theta'}\{a(\theta') \,:\, \theta' \in D_\kappa\}$. Let C be a small set such that (5.41) holds. Because the chain is irreducible under p by construction, we can find for any D_κ at least one $\theta^{(0)} \in C$ and some m such that

$$\mathbb{P}(\theta^{(m)} \in D_\kappa, \tau_C > m) > 0 \ . \tag{5.43}$$

As $\theta^{(m)} \in D_\kappa$, we have $\mathbb{P}(\theta^{(m+1)} = \theta^{(m)}) \geq 1 - a(\theta_\kappa)$, and therefore

$$\mathbb{P}(\tau_C \geq m + k \mid \theta^{(m)} \in D_\kappa) \geq (1 - a(\theta_\kappa))^k \ . \tag{5.44}$$

For a given $\kappa > 1$, expectation (5.41) can be lower bounded in the following way:

$$\begin{aligned}
\mathbb{E}_{\theta^{(0)}}[\kappa^{\tau_C}] &= \sum_{k=0}^{\infty} \mathbb{P}(\tau_C \geq k,\, \theta^{(m)} \notin D_\kappa)\kappa^k + \sum_{k=0}^{\infty} \mathbb{P}(\tau_C \geq k,\, \theta^{(m)} \in D_\kappa)\kappa^k \\
&\geq \sum_{k=m}^{\infty} \mathbb{P}(\tau_C \geq k \mid \theta^{(m)} \in D_\kappa)\mathbb{P}(\theta^{(m)} \in D_\kappa)\kappa^k \\
&= \mathbb{P}(\theta^{(m)} \in D_\kappa)\kappa^m \sum_{k=0}^{\infty} \mathbb{P}(\tau_C \geq m + k \mid \theta^{(m)} \in D_\kappa)\kappa^k \\
&\geq \mathbb{P}(\theta^{(m)} \in D_\kappa)\kappa^m \sum_{k=0}^{\infty} \Big[(1 - a(\theta_\kappa))\kappa\Big]^k \ .
\end{aligned} \tag{5.45}$$

From (5.42) we have $(1 - a(\theta_\kappa))\kappa > 1$, and therefore the sum $\sum_{k=0}^{\infty}[(1 - a(\theta_\kappa))\kappa]^k$ will diverge to infinity with our choice of D_κ, thus giving a contradiction to our assumption in (5.41). □

We have now seen through a toy example why MH moves with the variational proposal density $q(\theta|\theta')$ can fail. Another attempt is to use so-called importance weights in conjunction with samples drawn directly from the chain $q(\theta|\theta')$, which we discuss next.

5.4.2 Importance sampling

Another possibility to consider might be to run a Markov chain with $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ as transition kernel, as it is an approximation to a kernel with target density $p^*(\boldsymbol{\theta})$ as stationary distribution. As the former chain doesn't exactly do a random walk proportional to $p^*(\boldsymbol{\theta})$, the samples from the chain $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ can be reweighted to the correct proportions with importance weights. As we will soon see, the practical problems arising are intimately related to the fact that using q as MH proposal does not allow for a central limit theorem. Although ergodic estimates (averages) should converge to true averages, not much can again be said about the variance of the estimate, as the variance of the importance weights used can be infinite. In this sense, using the chain with kernel q as instrumental densities in importance sampling is an impractical idea. For completeness, a full justification is given below.

Importance weights

We are sampling from $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$, which is an approximation of the required kernel that has the posterior as invariant distribution. Say we want to determine the expectation of some function $\phi(\boldsymbol{\theta})$ under the posterior distribution, $\Phi = \langle\phi(\boldsymbol{\theta})\rangle$.

Let $p^*(\boldsymbol{\theta})$ again be a shorthand for the (possibly unnormalized) target density, and assume that it, or the posterior $p(\boldsymbol{\theta}|\mathbf{x})$, can be evaluated up to a normalizing constant. To estimate Φ, an importance weight $\omega^{(t)}$ is added to each sample $\boldsymbol{\theta}^{(t)}$ coming from the kernel $q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})$,

$$\omega^{(t)} \equiv \omega(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}) = \frac{p^*(\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})} \ . \tag{5.46}$$

We require that $q(\boldsymbol{\theta}|\boldsymbol{\theta}') > 0$ wherever $p^*(\boldsymbol{\theta}) > 0$; hence the approximate kernel must at least have the same support as the required posterior distribution. Notice that in normal importance sampling we need not add the restriction that q is normalized, as the sampling distribution q (also known as an instrumental distribution) does not change and hence the normalizing constants will cancel out when we take the expectation of a function using our weights and sample. But now the sampling distribution *changes* and we can expect different normalizing factors for each $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$. We use

$$\hat{\Phi}_T^{\mathcal{I}} = \frac{\sum_{t=1}^T \omega^{(t)}\phi(\boldsymbol{\theta}^{(t)})}{\sum_{t=1}^T \omega^{(t)}} \tag{5.47}$$

as an estimate for Φ. Superscript $\mathcal{I}$, for 'importance', is merely to differentiate between (5.2) and (5.47).

The correct estimate

In a similar argument used for a fixed instrumental distribution in MacKay (2003, chapter 29), it can be shown that $\hat{\Phi}_T^{\mathcal{I}}$ should converge to Φ, provided that the ratio $\omega^{(t)}$ does not give rise to infinite variance.

Say the random variables $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ have a joint distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')$, such that the transition kernel $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ has *unknown* invariant distribution

$$\pi(\boldsymbol{\theta}) = \int q(\boldsymbol{\theta}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' \ . \tag{5.48}$$

(We hope that this invariant distribution will be close to $p^*(\boldsymbol{\theta})$, but have no guarantee.) A sample $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T$ is taken using $q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})$, and we need to determine the average weights under it. We average over $\pi(\boldsymbol{\theta}')$, the probability of seeing $\boldsymbol{\theta}'$ as a conditional argument; we then average over $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$, the probability of $\boldsymbol{\theta}$ *given* $\boldsymbol{\theta}'$.

$$\begin{aligned}\langle\omega\rangle &= \iint \omega(\boldsymbol{\theta}|\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' d\boldsymbol{\theta} \\ &= \iint \frac{p^*(\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\theta}')} q(\boldsymbol{\theta}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' d\boldsymbol{\theta} \\ &= \iint p^*(\boldsymbol{\theta})\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' d\boldsymbol{\theta} = \int p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathcal{Z}_p \ .\end{aligned} \tag{5.49}$$

Hence $\langle\sum_{t=1}^T \omega^{(t)}\rangle = T\mathcal{Z}_p$. We similarly show that $\langle\sum_{t=1}^T \omega^{(t)}\phi(\boldsymbol{\theta}^{(t)})\rangle = T\mathcal{Z}_p\Phi$:

$$\big\langle\omega\phi(\boldsymbol{\theta})\big\rangle = \iint \frac{p^*(\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}\phi(\boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'\, d\boldsymbol{\theta}$$

$$= \int \phi(\boldsymbol{\theta}) p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathcal{Z}_p \Phi \ . \tag{5.50}$$

Finally $\langle \hat{\Phi}_T^{\mathcal{I}} \rangle = T\mathcal{Z}_p\Phi / T\mathcal{Z}_p = \Phi$, and $\hat{\Phi}_T^{\mathcal{I}} \to \Phi$ as $T \to \infty$. As only p^* is not normalized, the estimator is consistent and unbiased.

Where importance sampling fails

Consider again the toy example of section 5.4.1. We shall assume here without loss of generality that $p^*(\boldsymbol{\theta})$ is normalized. The average importance weight is one,

$$\langle \omega \rangle = \iint \Big[\frac{p^*(\theta)}{q(\theta|\theta')}\Big] q(\theta|\theta')\pi(\theta)\, d\theta' d\theta = \iint p^*(\theta)\pi(\theta')\, d\theta' d\theta = 1 \ . \tag{5.51}$$

The variance of the importance weight is determined as

$$\begin{aligned}
\mathsf{var}(\omega) &= \iint \Big[\frac{p^*(\theta)}{q(\theta|\theta')} - 1\Big]^2 q(\theta|\theta')\pi(\theta')\, d\theta' d\theta \\
&= \iint \frac{p^*(\theta)^2}{q(\theta|\theta')}\pi(\theta)\, d\theta' d\theta - 2\iint p^*(\theta)\pi(\theta')\, d\theta' d\theta + \iint q(\theta|\theta')\pi(\theta')\, d\theta' d\theta \\
&= \iint \frac{p^*(\theta)^2}{q(\theta|\theta')}\pi(\theta')\, d\theta' d\theta - 1 \ .
\end{aligned} \tag{5.52}$$

We return to the toy example. To keep the notation simple, let $\rho \equiv \rho_{11}$. As $q(\theta|\theta')$ is Gaussian, we can analytically solve for the stationary distribution $\pi(\theta')$. Substituting the toy example's distributions into (5.52) (with Z indicating the appropriate normalizing constants) we have

$$\begin{aligned}
\mathsf{var}(\omega) &= \frac{Z_q}{Z_p^2 Z_\pi} \iint \Big(e^{-\frac{1}{2}\rho(1-r)\theta^2}\Big)^2 \Big(e^{\frac{1}{2}\rho(\theta - r\theta')^2}\Big)\Big(e^{-\frac{1}{2}\rho(1-r^2)\theta'^2}\Big)\, d\theta' d\theta - 1 \\
&= \frac{Z_q}{Z_p^2 Z_\pi} \int \exp\Big\{ -\frac{\rho}{2}\Big[1 - 2r - \frac{r^2}{1-2r^2}\Big]\theta^2 \Big\} \ldots \\
&\qquad \int \exp\Big\{ -\tfrac{\rho}{2}(1-2r^2)\big[\theta' + \frac{r\theta}{1-2r^2}\big]^2 \Big\}\, d\theta' d\theta - 1 \ .
\end{aligned} \tag{5.53}$$

The integral will *diverge* if $1 - 2r^2 \leq 0$, hence if $r \geq \frac{1}{\sqrt{2}}$. Assume that $r < \frac{1}{\sqrt{2}}$, so that the inner integral over θ' is finite:

$$\mathsf{var}(\omega) = \frac{Z_q}{Z_p^2 Z_\pi} \sqrt{\frac{2\pi}{\rho(1-2r^2)}} \int \exp\Big\{ -\frac{\rho}{2}\Big[1 - 2r - \frac{r^2}{1-2r^2}\Big]\theta^2 \Big\}\, d\theta - 1 \ . \tag{5.54}$$

The last integral over θ is finite if and only if (remember the assumption that $r < \frac{1}{\sqrt{2}}$)

$$\begin{aligned}
1 - 2r - \frac{r^2}{1-2r^2} &> 0 \\
\Rightarrow (r-1)(4r^2 + r - 1) &> 0 \ ,
\end{aligned} \tag{5.55}$$

and this is true when $r < 1$ (which is trivially true already) and $4r^2 + r - 1 < 0$. The variance of the importance weights will only be *finite* for values of r (recalling that r is nonnegative) that satisfy[2]

$$r < \frac{\sqrt{17} - 1}{8} \ . \tag{5.56}$$

[2]Changing the order of integration gives the same result.

Even though the chain with $q(\theta|\theta')$ explores a greater part of parameter space, the failure of importance sampling can be ascribed to the variance of q being too small; this is also a clear motivation for using heavy-tailed distributions (e.g. the Student t-distribution) as instrumental densities.

For the sake of interest, using $\mathcal{K}(\theta|\theta')$ instead of $q(\theta|\theta')$ in the above derivation gives a finite weight variance only if $r < \frac{1}{2}$. Although the kernel $\mathcal{K}(\theta|\theta')$ is the transition kernel of a Markov chain with $p^*(\theta)$ as invariant distribution (*no* importance weights needed), the addition of importance weights (if we did not know that $\mathcal{K}$ was correct) can still give an infinite variance!

Having shown a lack of geometric ergodicity, and also potential failure of importance sampling, in the toy example, we turn our attention to a practical fix.

5.4.3 Mixing kernels

When using q in practice, we often have an acceptance ratio A (in $\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}') = \min(1, A)$) close to one in areas of high density (e.g. a posterior mode), as q approximates $\mathcal{K}$ well. Despite this, we have no guarantee of geometric ergodicity, as we have just seen in section 5.4.1. This is because we have minimized a *bound* on the KL-divergence between the true kernel and q, and not the KL-divergence itself. The bound is not necessarily good, and q typically underestimates the variance of $\mathcal{K}$. For for $\boldsymbol{\theta}'$ in the extreme tails of the posterior we often find that A is small, and in fact can be made arbitrarily small by choosing $\boldsymbol{\theta}'$ to be far enough out in the tail.

This does not pose a serious difficulty, and a path similar to (de Freitas et al., 2001) can be taken. We can proceed by allowing a mixture between a proposal that is biased towards and performs well in high density areas, and a symmetric random walk that improves acceptance in the tails. An elegant property of detailed balance is that a set of Markov kernels K_1 and K_2, each with invariant distribution $p^*(\boldsymbol{\theta})$, can be combined in a *cycle* K_1K_2 or *mixture* $\alpha K_1 + (1-\alpha)K_2$, with $0 < \alpha < 1$, in order to improve convergence. We can therefore adopt a *mixture* by sampling $\boldsymbol{\theta}$ from $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ with probability α, and use a *symmetric* proposal density with probability $1-\alpha$.

5.5 Concluding remarks

We have seen how a variational kernel, despite its elegant derivation and possible good performance in areas of high density, has a potential weakness, as we have no guarantee that the variance of estimates $\hat{\Phi}_T$ or $\hat{\Phi}_T^{\mathcal{T}}$ are finite. The kernel can be combined with other methods with well-established convergence results. As an example, Gibbs sampling from $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$ in a mixture modelling problem comes with a central limit theorem and geometric convergence guarantee (Tierney, 1994). Such a combination therefore seems akin to flogging a dead horse; from the other side, we may argue that if we have a method with proven convergence properties, it is hard to justify augmenting it with a variational kernel that comes without the same guarantees.

In section 2.2 we illustrated that other divergence measures may not suffer from the same problem of underestimating the variance of a target density. The variational kernel that was derived in this chapter didn't rely on any optimization routines, but came in a closed-form solution. On the other hand, the expectation propagation algorithm relies on continually updating individual factors. If we keep in mind that a transition kernel should ideally be fast to compute, it therefore remains to be seen how these deterministic approximate methods can be successfully incorporated into MCMC samplers.

Chapter 6

Conclusion

6.1 Summary of contributions

This thesis presented practical methods for evaluating the large sums (often with an exponential number of terms) or intractable integrals that are often required in Bayesian inference.

Chapter 2 introduced Minka (2005)'s generic α-divergence message passing scheme, which allowed us to interpolate between VB and EP and beyond. A simple mixture of Gaussians with unknown means was taken as a running toy example, for which a new generic algorithm was derived. The main purpose of chapter 2 was to give a broad overview of the behavior of such algorithms. New intuition was given on the effect of the width of the prior distribution to model pruning and local minima in VB, and why EP is not prone to the same behavior. These findings were used to increase the robustness of the VB message passing algorithm for multivariate mixtures. We also showed under which conditions the VB message passing algorithm over a factor graph behaves like the standard VBEM algorithm, where a lower bound on the marginal likelihood is always increased.

Two new approaches to inference for a multivariate mixture of Gaussians, namely EP and the more general α-divergence message passing scheme, were contributed by chapter 3. A benchmark comparison with parallel tempering and thermodynamic integration showed that VB, EP, and $\alpha = \frac{1}{2}$ message passing are suitable for model selection, and approximating the predictive distribution with high accuracy. It was practically shown that EP need not have a unique fixed point, and if the fixed points are not unique, they depend on both the initialization and the random order in which factor refinements take place. A number of other points were empirically illustrated: the log marginal likelihood estimates increase with α; the number of local solutions depends on the prior width; the discrepancy between the approximate and true log marginal likelihoods increase with model size; the marginal likelihoods give a characteristic 'Ockham hill' as the model size increases, providing a useful tool for model selection.

Parallel tempering and thermodynamic integration, with a new parallel tempered approach to sampling from a mixture of Gaussians posterior through Gibbs sampling, were introduced in chapter 4. We have also made thermodynamic integration numerically stabler with a principled method of interpolation over high-temperature averages. The numerical stability was further addressed by generalizing parallel tempering to include a surrogate prior. With a carefully chosen surrogate prior the variance of the samples in the chain, especially at high temperatures, can be meaningfully reduced.

In chapter 5 we have also attempted to introduce variational methods into the design of MCMC transition densities. A detailed proof was given why such methods may not give a geometrically ergodic chain. In essence the mean of our MCMC estimate converges to the

correct mean by the law of large numbers, but unfortunately we cannot say anything useful about the variance of the estimate, as a central limit theorem cannot be shown to hold. This rendered the use of the variational kernel impractical as a stand-alone MCMC algorithm, and therefore further layers of MH proposals need to be introduced.

6.2 Future work

There are many directions for future research, both fundamental and practical.

Convergence of EP. EP is an effective method for minimizing the EC/EP free energy, but does not come with a convergence guarantee. Minka (2001a) made the initial empirical observation that EP converged on unimodal posteriors, but failed to converge on strongly multimodal posteriors. In this thesis we have seen where EP converges even when the posterior is multimodal, provided the modes are 'well enough' separated. Furthermore, L. Csató has conjectured (but not proven) that EP is guaranteed to converge if the likelihood is log concave (Rasmussen & Williams, 2006). A formal study on these questions will be a valuable contribution to the field of machine learning.

Double loop algorithms. When EP does not converge, we may switch to double loop algorithms. None of the algorithms in this thesis were extended to double loop algorithms, and it provides a basis for further improvements.

Perturbative corrections. Opper (2006) showed how perturbative corrections can be used to improve expectation consistent (EC) approximations. A convincing example, showing the improvement on a toy example, was given as a conclusion to chapter 3 (section 3.8.3). This is clearly a promising direction for future work.

Approximate moments. We have also seen in section 3.8 that intractable partition functions often arise for the product of a prior and one likelihood term, even for simple models. For EP/C ($\alpha = 1$), we will not be able to do moment matching, and it begs the question whether it is worthwhile to find approximations to the moments, and how accurate such approximations in an approximate algorithm is. The models mentioned in section 3.8 rely on inner products between two random vectors, where the elements are neither independent nor identically distributed. Under certain conditions we can rely on Lyapunov's central limit theorem to show that the inner product is still approximately Gaussian, therefore both approximating and simplifying the problem to an integral over the inner product. Then again, we may bypass all these difficulties by choosing another divergence measure, namely $\alpha = 0$ (VB).

Extensions. In section 3.8 it was mentioned that extensions of EP/C to higher-order mixtures, namely HMMs, are also possible. How will these extensions compare to existing algorithms for treating HMMs?

Infinite models. If we prefer a non-parametric approach to mixture modeling, we can always implement an infinite mixture. Minka & Ghahramani (2003) implemented an infinite mixture of Gaussians (with fixed variance) through a Dirichlet process, but concluded that Gibbs sampling remains, in that case, the method of choice. It remains an open question whether the accuracy of EP can be improved in this particular case.

Surrogate distributions in parallel tempering. As argued in section 4.3, the nature of parallel tempering can be radically changed with the introduction of a distribution other than

the prior at $\beta = 0$. Can we use the fact that we can find approximations to the posterior to our advantage, i.e. can we choose a surrogate prior such that the variance of the MCMC estimate is minimized? This is a fruitful direction of questioning, as we know that when we choose the posterior as surrogate prior, the log likelihood expectations have a zero variance.

Can this technique be applicable elsewhere? For example, parallel tempering cannot treat first order phase transitions, where there is a point of no energy overlap in a figure like figure 4.1(a). May a clever introduction of an additional distribution change that?

Approximations and MH kernels. We have seen the dangers of incorporating variational methods in the construction of MH proposal densities. However, we would be wrong to conclude that methods from approximate inference cannot greatly enhance Monte Carlo methods: One way may be their introduction into parallel tempering algorithms. Effective geometrically ergodic kernels, built around fast approximations, are still to be found.

Appendix A

Useful results

A.1 Kullback-Leibler as special cases of an α-divergence

In section 2.2 an α-divergence was introduced for unnormalized distributions. We shall formally show here that the two Kullback-Leibler (KL) divergences are special cases of an α-divergence. Take the limit $\alpha \to 1$,

$$\begin{aligned}
\lim_{\alpha\to 1} D_\alpha\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) &= \lim_{\alpha\to 1} \frac{\int \alpha p(\mathbf{x},\boldsymbol{\theta}) + (1-\alpha) sq(\boldsymbol{\theta}) - p(\mathbf{x},\boldsymbol{\theta})^\alpha [sq(\boldsymbol{\theta})]^{1-\alpha} d\boldsymbol{\theta}}{\alpha(1-\alpha)} \\
&= \lim_{\alpha\to 1} \Bigg(\int p(\mathbf{x},\boldsymbol{\theta}) - sq(\boldsymbol{\theta}) - \ln[p(\mathbf{x},\boldsymbol{\theta})] p(\mathbf{x},\boldsymbol{\theta})^\alpha [sq(\boldsymbol{\theta})]^{1-\alpha} \\
&\qquad \cdots + \ln[sq(\boldsymbol{\theta})] p(\mathbf{x},\boldsymbol{\theta})^\alpha [sq(\boldsymbol{\theta})]^{1-\alpha} d\boldsymbol{\theta} \Bigg) / (1-2\alpha) \\
&= \int p(\mathbf{x},\boldsymbol{\theta}) \ln \frac{p(\mathbf{x},\boldsymbol{\theta})}{sq(\boldsymbol{\theta})} \, d\boldsymbol{\theta} + \int \Big(sq(\boldsymbol{\theta}) - p(\mathbf{x},\boldsymbol{\theta}) \Big) \, d\boldsymbol{\theta} \\
&= \mathsf{KL}\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) \ , \qquad \text{(A.1)}
\end{aligned}$$

where the second line follows from l'Hôpital's rule (taking the derivative of the numerator and denumerator with respect to α). With a very similar argument we can show that

$$\lim_{\alpha\to 0} D_\alpha\big(p(\mathbf{x},\boldsymbol{\theta}) \,\big\|\, sq(\boldsymbol{\theta})\big) = \mathsf{KL}\big(sq(\boldsymbol{\theta}) \,\big\|\, p(\mathbf{x},\boldsymbol{\theta})\big) \ . \qquad \text{(A.2)}$$

A.2 Responsibility-weighted moment matching: two derivations

This section aims to give two example derivations of weighed moment-matching equations. The first derivation comes from chapter 2 forms a skeleton for finding moments for more complex distributions in a mixture model, most notably for the Normal-Wishart distributions following in chapter 3. The second derivation is for the moments of a Dirichlet distribution.

A.2.1 A Gaussian derivation

In section 2.4 we had a choice of approximating distribution $q(\mu_j) = \mathcal{N}(\mu_j \,|\, m_j, v_j^{-1})$, and would like to solve for m_j in

$$\partial \mathsf{KL}\big(p(x_n,\boldsymbol{\mu}) \,\big\|\, sq(\boldsymbol{\mu})\big) / \partial m_j = 0 \ . \qquad \text{(A.3)}$$

To update the parameter m_j of each approximate distribution $q(\mu_j)$, write the KL divergence as a function of m_j. This gives

$$\begin{aligned}
\mathsf{KL}\big(p(x_n, \boldsymbol{\mu}) \,\big\|\, sq(\boldsymbol{\mu})\big) &= -\int p(x_n, \boldsymbol{\mu}) \ln q(\boldsymbol{\mu})\, d\boldsymbol{\mu} + \mathsf{const} \\
&= -\int \Big[\sum_{i=1}^{J} \ln q(\mu_j)\Big] \Big[p(\boldsymbol{\mu}) \sum_{k=1}^{J} \pi_k p(x_n|\mu_k)\Big]\, d\boldsymbol{\mu} + \mathsf{const} \\
&= -\sum_{k=1}^{J} \pi_k \int p(\boldsymbol{\mu}) p(x_n|\mu_k) \ln q(\mu_j)\, d\boldsymbol{\mu} + \mathsf{const}\ . \qquad \text{(A.4)}
\end{aligned}$$

The absence of components $i \neq j$ in the last line follows merely from the independence of $q(\mu_i)$ and $q(\mu_j)$, and are included in the constant. Take the derivative with respect to m_j:

$$\begin{aligned}
\frac{\partial \mathsf{KL}}{\partial m_j} &= -\sum_{k=1}^{J} \pi_k \int \frac{\partial}{\partial m_j} \ln q(\mu_j|m_j, v_j^{-1}) \prod_{i=1}^{J} p(\mu_i) \times p(x_n|\mu_k)\, d\boldsymbol{\mu} \\
&= -\sum_{k\neq j} \pi_k \int v_j(\mu_j - m_j) p(\mu_j) p(\mu_k) p(x_n|\mu_k)\, d\mu_j\, d\mu_k \\
&\qquad \cdots - \pi_j \int v_j(\mu_j - m_j) p(\mu_j) p(x_n|\mu_j)\, d\mu_j \\
&= -v_j \Big[\Big(\int \mu_j p(\mu_j)\, d\mu_j\Big) \sum_{k\neq j} \pi_k \int p(\mu_k) p(x_n|\mu_k)\, d\mu_k \\
&\qquad \cdots + \pi_j \int \mu_j p(\mu_j) p(x_n|\mu_j)\, d\mu_j - s m_j\Big]\ . \qquad \text{(A.5)}
\end{aligned}$$

Define the responsibilities as

$$r_{nj} = \frac{\pi_j \int p(\mu_j) p(x_n|\mu_j)\, d\mu_j}{\sum_k \pi_k \int p(\mu_k) p(x_n|\mu_k)\, d\mu_k} = \frac{\pi_j \mathcal{N}(x_n|m_{0j}, \lambda_j^{-1} + v_{0j}^{-1})}{\sum_k \pi_k \mathcal{N}(x_n|m_{0k}, \lambda_k^{-1} + v_{0k}^{-1})}\ , \qquad \text{(A.6)}$$

so that when we equate the above expression in (A.5) to zero, we get

$$\begin{aligned}
m_j &= (1 - r_{nj}) \int \mu_j p(\mu_j)\, d\mu_j + r_{nj} \int \mu_j p(\mu_j|x_n)\, d\mu_j \\
&= (1 - r_{nj}) \langle \mu_j \rangle + r_{nj} \langle \mu_j | x_n \rangle\ . \qquad \text{(A.7)}
\end{aligned}$$

We can use a similar derivation to derive an update equation for each v_j, or more generally the parameters in other mixtures of exponential distributions (see for example chapter 3).

A.2.2 A Dirichlet derivation

In section 3.3.3 we had a choice of approximating distribution $q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta})$, and would like to solve for δ_j in

$$\partial \mathsf{KL}\big(p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) \| sq(\boldsymbol{\mu}, \boldsymbol{\Lambda}) q(\boldsymbol{\pi})\big) \big/ \partial \delta_j = 0\ . \qquad \text{(A.8)}$$

This gives $\langle \ln \pi_j \rangle_q = \langle \ln \pi_j | \mathbf{x}_n \rangle$, where the second expectation is taken given that we observed $\mathbf{x}_n$, i.e. over the posterior distribution. As $p_k(\mathbf{x}_n)$, given by (3.9), already defines the likelihood

integrated over $\boldsymbol{\mu}_j$ and $\boldsymbol{\Lambda}_j$ of each respective component prior, the expectation of $\ln \pi_j$ under the posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}|\mathbf{x}_n)$ is shortened with

$$\begin{aligned}
\langle \ln \pi_j | \mathbf{x}_n \rangle &= \frac{1}{s} \int \ln \pi_j \, p(\boldsymbol{\pi}) p(\mathbf{x}_n|\boldsymbol{\pi}) \, d\boldsymbol{\pi} \qquad \text{(A.9)} \\
&= \frac{1}{s} \sum_{k=1}^{J} p_k(\mathbf{x}_n) \frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0, \delta_{0k}+1)}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)} \int \ln \pi_j \, \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}_0, \delta_{0k}+1) \, d\boldsymbol{\pi} \\
&= \frac{1}{s} \sum_{k \neq j} \frac{\delta_{0k}}{\sum_{i=1}^{J} \delta_{0i}} p_k(\mathbf{x}_n) \left[\Psi(\delta_{0j}) - \Psi\Big(\sum_{i=1}^{J} \delta_{0i} + 1 \Big) \right] \\
&\quad + \frac{1}{s} \frac{\delta_{0j}}{\sum_{i=1}^{J} \delta_{0i}} p_j(\mathbf{x}_n) \left[\Psi(\delta_{0j}+1) - \Psi\Big(\sum_{i=1}^{J} \delta_{0i} + 1 \Big) \right] \\
&= -\Psi\Big(\sum_{i=1}^{J} \delta_{0i} + 1 \Big) + (1 - r_{nj}) \Psi(\delta_{0j}) + r_{nj} \Psi(\delta_{0j}+1) \\
&= \Psi(\delta_{0j}) - \Psi\Big(\sum_{i=1}^{J} \delta_{0i} \Big) - \frac{1}{\sum_{i=1}^{J} \delta_{0i}} + \frac{r_{nj}}{\delta_{0j}} \, . \qquad \text{(A.10)}
\end{aligned}$$

In the second last line we again see a *responsibility-weighed* sum, in this case given by r_{nj} defined in (3.14).

A.3 The scale for multivariate mixtures

A short derivation of the scale (or partition function) needed in section 3.3 is presented here. The scale is determined with $s = \int p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) \, d\boldsymbol{\mu} \, d\boldsymbol{\Lambda} \, d\boldsymbol{\pi}$, and as $p_j(\mathbf{x}_n)$ from equation (3.9) already summarizes the integral over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, we have

$$\begin{aligned}
s &= \int \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\delta}) \sum_{k=1}^{J} \pi_k p_k(\mathbf{x}_n) \, d\boldsymbol{\pi} \\
&= \sum_{k=1}^{J} p_k(\mathbf{x}_n) \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)} \int \pi_k^{\delta_{0k}} \prod_{j \neq k} \pi_j^{\delta_{0j}-1} \, d\boldsymbol{\pi} \\
&= \sum_{k=1}^{J} p_k(\mathbf{x}_n) \frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0, \delta_{0k}+1)}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)} \\
&= \frac{1}{\sum_{j=1}^{J} \delta_{0j}} \sum_{k=1}^{J} \delta_{0k} p_k(\mathbf{x}_n) \, , \qquad \text{(A.11)}
\end{aligned}$$

where the notation in $\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0, \delta_{0k}+1)$ implies using parameter vector $\boldsymbol{\delta}_0$, except that the value of component δ_{0k} is incremented by one.

A.4 α-divergence scales

The fixed point iterations that are used to minimize an α-diverge require the evaluation the partition function (or scale) of a distribution. The scales needed for the fixed point algorithms in sections 2.6 and 3.5 are presented here.

A.4.1 For section 2.6: a simple mixture

For the fixed point scheme in section 2.6.1, the following integral needs to be evaluated:

$$\begin{aligned} s_{(t')} &= \int \sum_{\mathbf{z}_n} p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)^\alpha [s_{(t)} q_{(t)}(\mathbf{z}_n) q_{(t)}(\boldsymbol{\mu})]^{1-\alpha} d\boldsymbol{\mu} \\ &= s_{(t)}^{1-\alpha} \sum_{k=1}^{J} \pi_k^\alpha \gamma_{nk(t)}^{1-\alpha} \int p(\boldsymbol{\mu})^\alpha q_{(t)}(\boldsymbol{\mu})^{1-\alpha} p(x_n|\mu_k)^\alpha d\boldsymbol{\mu} \\ &= s_{(t)}^{1-\alpha} \sum_{k=1}^{J} \pi_k^\alpha \gamma_{nk(t)}^{1-\alpha} \mathcal{I}_k \ . \end{aligned} \tag{A.12}$$

We shall set about evaluating integral $\mathcal{I}_k$. For a shorthand define

$$\hat{v}_i = \alpha v_{0i} + (1-\alpha) v_{i(t)} \tag{A.13}$$

$$\hat{m}_i = \frac{\alpha v_{0i} m_{0i} + (1-\alpha) v_{i(t)} m_{i(t)}}{\alpha v_{0i} + (1-\alpha) v_{i(t)}} \ , \tag{A.14}$$

so that a specific component i from $p(\boldsymbol{\mu})^\alpha q_{(t)}(\boldsymbol{\mu})^{1-\alpha}$ can be rearranged as

$$\begin{aligned} p(\mu_i)^\alpha q_{(t)}(\mu_i)^{1-\alpha} &= \mathcal{N}(\mu_i|m_{0i}, v_{0i}^{-1})^\alpha \mathcal{N}(\mu_i|m_{i(t)}, v_{i(t)}^{-1})^{1-\alpha} \\ &= \frac{\mathcal{Z}_{\mathcal{N}}(\hat{v}_i)}{\mathcal{Z}_{\mathcal{N}}(v_{0i})^\alpha \mathcal{Z}_{\mathcal{N}}(v_{i(t)})^{1-\alpha}} \exp\Big\{ -\frac{1}{2} \frac{\alpha v_{0i}(1-\alpha) v_{i(t)}}{\hat{v}_i} \big[m_{0i} - m_{i(t)}\big]^2 \Big\} \\ &\quad \times \mathcal{N}(\mu_i \,|\hat{m}_i,\ \hat{v}_i^{-1}) \ . \end{aligned} \tag{A.15}$$

In the above rearrangement we only have one $\mathcal{N}(\mu_i \,|\hat{m}_i,\ \hat{v}_i^{-1})$. When we evaluate integral $\mathcal{I}_k$ all of these component distributions $i \neq k$ integrate to one. The last integration needed is to determine the evidence, with a likelihood raised to the power α, for component k:

$$\begin{aligned} \int \mathcal{N}(\mu_k|\hat{m}_k, \hat{v}_k^{-1}) \mathcal{N}(x_n|\mu_k, \lambda_k^{-1})^\alpha \, d\mu_k &= \frac{1}{\mathcal{Z}_{\mathcal{N}}(\hat{v}_k)} \frac{1}{\mathcal{Z}_{\mathcal{N}}(\lambda_k)^\alpha} \exp\Big\{ -\frac{1}{2} \frac{\hat{v}_k \alpha \lambda_k}{\hat{v}_k + \alpha\lambda_k} (x_n - \hat{m}_k)^2 \Big\} \\ &\quad \times \int \underbrace{\exp\Big\{ -\frac{1}{2} (\hat{v}_k + \alpha\lambda_k) \Big(\mu_k - \frac{\hat{v}_k \hat{m}_k + \alpha\lambda_k x_n}{\hat{v}_k + \alpha\lambda_k} \Big)^2 \Big\}}_{\propto \mathcal{N}(\mu_k|x_n)} d\mu_k \\ &= \alpha^{-1/2} \mathcal{Z}_{\mathcal{N}}(\lambda_k)^{1-\alpha} \mathcal{N}\Big(x_n \,\Big|\, \hat{m}_k,\ \frac{1}{\alpha\lambda_k} + \frac{1}{\hat{v}_k} \Big) \ . \end{aligned} \tag{A.16}$$

We therefore get a scale evaluation

$$\begin{aligned} s_{(t')} = s_{(t)}^{1-\alpha} \prod_{i=1}^{J} &\frac{\mathcal{Z}_{\mathcal{N}}(\hat{v}_i)}{\mathcal{Z}_{\mathcal{N}}(v_{0i})^\alpha \mathcal{Z}_{\mathcal{N}}(v_{i(t)})^{1-\alpha}} \exp\Big\{ -\frac{1}{2} \frac{\alpha v_{0i}(1-\alpha) v_{i(t)}}{\hat{v}_i} \big[m_{0i} - m_{i(t)}\big]^2 \Big\} \\ &\times \alpha^{-1/2} \sum_{k=1}^{J} \pi_k^\alpha \gamma_{nk(t)}^{1-\alpha} \mathcal{Z}_{\mathcal{N}}(\lambda_k)^{1-\alpha} \mathcal{N}\Big(x_n \,\Big|\, \hat{m}_k,\ \frac{1}{\alpha\lambda_k} + \frac{1}{\hat{v}_k} \Big) \ . \end{aligned} \tag{A.17}$$

A.4.2 For section 3.5: a multivariate mixture

For the fixed point scheme in section 3.5, the following integral needs to be evaluated:

$$s_{(t')} = \int \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}, \mathbf{z}_n)^\alpha [s_{(t)} q_{(t)}(\mathbf{z}_n) q_{(t)}(\boldsymbol{\pi}) q_{(t)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})]^{1-\alpha} \, d\boldsymbol{\mu} \, d\boldsymbol{\Lambda} \, d\boldsymbol{\pi}$$

$$
\begin{aligned}
&= s_{(t)}^{1-\alpha} \int p(\boldsymbol{\pi})^{\alpha} q_{(t)}(\boldsymbol{\pi})^{1-\alpha} p(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{\alpha} q_{(t)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{1-\alpha} \sum_{\mathbf{z}_n} \prod_{k=1}^{J} [\pi_k^{\alpha} p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)^{\alpha} \gamma_{nk(t)}^{1-\alpha}]^{z_{nk}} d\boldsymbol{\mu}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\pi} \\
&= s_{(t)}^{1-\alpha} \sum_{k=1}^{J} \gamma_{nk(t)}^{1-\alpha} \int \pi_k^{\alpha} p(\boldsymbol{\pi})^{\alpha} q_{(t)}(\boldsymbol{\pi})^{1-\alpha} p(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{\alpha} q_{(t)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{1-\alpha} p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)^{\alpha} d\boldsymbol{\mu}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\pi} \\
&= s_{(t)}^{1-\alpha} \sum_{k=1}^{J} \gamma_{nk(t)}^{1-\alpha} \mathcal{I}_k \ . \qquad \text{(A.18)}
\end{aligned}
$$

The integral $\mathcal{I}_k$ contains $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{\alpha} q_{(t)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{1-\alpha} = \prod_{i=1}^{J} p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)^{\alpha} q_{(t)}(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)^{1-\alpha}$, and we shall first evaluate $\mathcal{I}_k$ over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. Define as a shorthand,

$$
\hat{v}_i = \alpha v_{0i} + (1-\alpha) v_{i(t)} \qquad \text{(A.19)}
$$

$$
\hat{\mathbf{m}}_i = \frac{\alpha v_{0i} \mathbf{m}_{0i} + (1-\alpha) v_{i(t)} \mathbf{m}_{i(t)}}{\alpha v_{0i} + (1-\alpha) v_{i(t)}} \qquad \text{(A.20)}
$$

$$
\hat{\mathbf{B}}_i = \alpha \mathbf{B}_{0i} + (1-\alpha) \mathbf{B}_{i(t)} + \frac{1}{2} \frac{\alpha(1-\alpha) v_{0i} v_{i(t)}}{\alpha v_{0i} + (1-\alpha) v_{i(t)}} (\mathbf{m}_{0i} - \mathbf{m}_{i(t)})(\mathbf{m}_{0i} - \mathbf{m}_{i(t)})^{\top} \qquad \text{(A.21)}
$$

$$
\hat{a}_i = \alpha a_{0i} + (1-\alpha) a_{i(t)} \ , \qquad \text{(A.22)}
$$

so that a specific component i from $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{\alpha} q_{(t)}(\boldsymbol{\mu}, \boldsymbol{\Lambda})^{1-\alpha}$ can be rearranged as

$$
\begin{aligned}
p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)^{\alpha} q_{(t)}(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)^{1-\alpha} = &\frac{\mathcal{Z}_{\mathcal{NW}}(\hat{v}_i, \hat{a}_i, \hat{\mathbf{B}}_i)}{\mathcal{Z}_{\mathcal{NW}}(v_{0i}, a_{0i}, \mathbf{B}_{0i})^{\alpha} \mathcal{Z}_{\mathcal{NW}}(v_{i(t)}, a_{i(t)}, \mathbf{B}_{i(t)})^{1-\alpha}} \\
&\times \mathcal{NW}\big(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i \,\big|\, \hat{\mathbf{m}}_i, \hat{v}^{-1}, \hat{a}_i, \hat{\mathbf{B}}_i\big) \ . \qquad \text{(A.23)}
\end{aligned}
$$

This rearrangement now contains one distribution over $\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i$, namely $\mathcal{NW}(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i|\hat{\mathbf{m}}_i, \hat{v}_i, \hat{a}_i, \hat{\mathbf{B}}_i)$, and we shall treat this Normal-Wishart as a prior distribution.

The integral $\mathcal{I}_k$ in (A.18) contains an exponentiated likelihood term corresponding to component k, so that all the component distributions $i \neq k$—where a component distribution i is the Normal-Wishart given in (A.23)—will integrate to one. Component k remains, and the next integration needed (fully evaluated in appendix A.6) is to determine the evidence, with a likelihood raised to the power α, for component k:

$$
\begin{aligned}
&\int \mathcal{NW}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\hat{m}_k, \hat{v}_k, \hat{a}_k, \hat{\mathbf{B}}_k) \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{\alpha}\, d\boldsymbol{\mu}_k\, d\boldsymbol{\Lambda}_k \\
&= (2\pi)^{(1-\alpha)d/2} \alpha^{-d/2} |\hat{\mathbf{B}}_k|^{(1-\alpha)/2} \frac{\Gamma(\frac{[2\hat{a}_k+\alpha-d]}{2})}{\Gamma(\frac{[2\hat{a}_k+\alpha-d]+d}{2})} \prod_{l=1}^{d} \frac{\Gamma(\frac{2\hat{a}_k+\alpha+1-l}{2})}{\Gamma(\frac{2\hat{a}_k+1-l}{2})} \\
&\quad \times \mathcal{T}\Big(\mathbf{x}_n \,\Big|\, \hat{\mathbf{m}}_k, \frac{\hat{v}_k+\alpha}{\hat{v}_k \alpha} \frac{2\hat{\mathbf{B}}_k}{2\hat{a}_k+\alpha-d}, 2\hat{a}_k+\alpha-d\Big) \ . \qquad \text{(A.24)}
\end{aligned}
$$

After integrating $\mathcal{I}_k$ over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, we also integrate over $\boldsymbol{\pi}$. Under the definition

$$
\hat{\delta}_j = \alpha \delta_{0j} + (1-\alpha) \delta_{j(t)} \qquad \text{(A.25)}
$$

we have

$$
\int \pi_k^{\alpha} p(\boldsymbol{\pi})^{\alpha} q_{(t)}(\boldsymbol{\pi})^{1-\alpha}\, d\boldsymbol{\pi} = \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)^{\alpha}} \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t)})^{1-\alpha}} \int \pi_k^{\alpha} \prod_{j=1}^{J} \pi_j^{\alpha\delta_{0j}+(1-\alpha)\delta_{j(t)}-1} d\boldsymbol{\pi}
$$

$$= \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)^{\alpha}} \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t)})^{1-\alpha}} \left(\frac{\prod_{j=1}^{J} \Gamma(\hat{\delta}_j)}{\Gamma(\alpha + \sum_{j=1}^{J} \hat{\delta}_j)} \right) \frac{\Gamma(\hat{\delta}_k + \alpha)}{\Gamma(\hat{\delta}_k)} \ . \tag{A.26}$$

By collating the results from (A.23), (A.24), and (A.26), $\mathcal{I}_k$ from (A.18) is evaluated to give the scale $s_{(t')}$ as

$$\begin{aligned} s_{(t')} = s_{(t)}^{1-\alpha} & \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_0)^{\alpha}} \frac{1}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{\delta}_{(t)})^{1-\alpha}} \left(\frac{\prod_{j} \Gamma(\hat{\delta}_j)}{\Gamma(\alpha + \sum_{j=1}^{J} \hat{\delta}_j)} \right) \\ & \times \prod_{j=1}^{J} \frac{\mathcal{Z}_{\mathcal{NW}}(\hat{v}_j, \hat{a}_j, \hat{\mathbf{B}}_j)}{\mathcal{Z}_{\mathcal{NW}}(v_{0j}, a_{0j}, \mathbf{B}_{0j})^{\alpha} \mathcal{Z}_{\mathcal{NW}}(v_{j(t)}, a_{j(t)}, \mathbf{B}_{j(t)})^{1-\alpha}} \\ & \times (2\pi)^{(1-\alpha)d/2} \alpha^{-d/2} \sum_{k=1}^{J} R_{nk} \ . \end{aligned} \tag{A.27}$$

The unscaled responsibilities used in (A.27) are:

$$\begin{aligned} R_{nj} = \gamma_{nj(t)}^{1-\alpha} & \frac{\Gamma(\hat{\delta}_j + \alpha)}{\Gamma(\hat{\delta}_j)} |\hat{\mathbf{B}}_j|^{(1-\alpha)/2} \frac{\Gamma(\frac{[2\hat{a}_j + \alpha - d]}{2})}{\Gamma(\frac{[2\hat{a}_j + \alpha - d] + d}{2})} \prod_{l=1}^{d} \frac{\Gamma(\frac{2\hat{a}_j + \alpha + 1 - l}{2})}{\Gamma(\frac{2\hat{a}_j + 1 - l}{2})} \\ & \times \mathcal{T}\Big(\mathbf{x}_n \,\Big|\, \hat{\mathbf{m}}_j, \ \frac{\hat{v}_j + \alpha}{\hat{v}_j \alpha} \frac{2\hat{\mathbf{B}}_j}{2\hat{a}_j + \alpha - d}, \ 2\hat{a}_j + \alpha - d \Big) \ . \end{aligned} \tag{A.28}$$

Note that the scale may not be finite, and we suddenly find ourselves with a set of practical constrains when α is outside the interval $[0, 1]$. These constrains are discussed in greater detail in section 3.5.

A.5 Multinomial updates for a fixed point scheme

In the fixed point scheme described in section 2.6.1, we had to continually update the parameters of the approximating multinomial distribution $q_{(t')}(\mathbf{z}_n)$. As described in section 2.6, this can be done by minimizing the KL divergence

$$\mathsf{KL}\Big(p(x_n, \boldsymbol{\mu}, \mathbf{x}_n)^{\alpha} [q_{(t)}(\mathbf{z}_n) q_{(t)}(\boldsymbol{\mu})]^{1-\alpha} \,\Big\|\, q_{(t')}(\mathbf{z}_n) q_{(t')}(\boldsymbol{\mu}) \Big) \tag{A.29}$$

with respect to $\gamma_{nj(t')}$, the parameters of $q_{(t')}(\mathbf{z}_n)$. Here it is necessary to add a Lagrange multiplier ℓ to enforce $\sum_j \gamma_{nj(t')} = 1$. Taking the partial derivative with respect to $\gamma_{nj(t')}$ gives

$$\begin{aligned} \frac{\partial \mathsf{KL}}{\partial \gamma_{nj(t')}} &= - \int \sum_{\mathbf{z}_n} \frac{\partial}{\partial \gamma_{nj(t')}} \Big[\sum_{i=1}^{J} z_{ni} \ln \gamma_{ni(t')} + \ln q_{(t')}(\boldsymbol{\mu}) \Big] \\ & \quad \times p(x_n, \boldsymbol{\mu}, \mathbf{z}_n)^{\alpha} [s_{(t)} q_{(t)}(\mathbf{z}_n) q_{(t)}(\boldsymbol{\mu})]^{1-\alpha} d\boldsymbol{\mu} + \frac{\partial}{\partial \gamma_{nj(t')}} \ell \Big[\sum_{i=1}^{J} \gamma_{ni(t')} - 1 \Big] \\ &= -s_{(t)}^{1-\alpha} \int p(\boldsymbol{\mu})^{\alpha} q_{(t)}(\boldsymbol{\mu})^{1-\alpha} \sum_{\mathbf{z}_n} \frac{z_{nj}}{\gamma_{nj(t')}} \prod_{k=1}^{J} [\pi_k^{\alpha} \gamma_{nk(t)}^{1-\alpha} p(x_n|\mu_k)^{\alpha}]^{z_{nk}} \, d\boldsymbol{\mu} + \ell \\ &= -s_{(t)}^{1-\alpha} \frac{1}{\gamma_{nj(t')}} \pi_j^{\alpha} \gamma_{nj(t)}^{1-\alpha} \int p(\boldsymbol{\mu})^{\alpha} q_{(t)}(\boldsymbol{\mu})^{1-\alpha} p(x_n|\mu_j)^{\alpha} d\boldsymbol{\mu} + \ell \ , \end{aligned} \tag{A.30}$$

where the last line follows as all terms in the sum over $\mathbf{z}_n$ are zero, except for when $j = k$. When the above derivative is set to zero, we can solve to find a unique expression for $\gamma_{nj(t')}$,

$$\gamma_{nj(t')} = \left[s_{(t)}^{1-\alpha} \pi_j^{\alpha} \gamma_{nj(t)}^{1-\alpha} \int p(\boldsymbol{\mu})^{\alpha} q_{(t)}(\boldsymbol{\mu})^{1-\alpha} p(x_n|\mu_j)^{\alpha} d\boldsymbol{\mu} \right] \Big/ \ell \ , \tag{A.31}$$

and we do that for every j. When all these unique expressions are added, and the constraint $\sum_j \gamma_{nj(t')} = 1$ kept in mind, the Lagrange multiplier is solved for as

$$\ell = s_{(t)}^{1-\alpha} \sum_{k=1}^{J} \pi_k^{\alpha} \gamma_{nk(t)}^{1-\alpha} \int p(\boldsymbol{\mu})^{\alpha} q_{(t)}(\boldsymbol{\mu})^{1-\alpha} p(x_n|\mu_k)^{\alpha} d\boldsymbol{\mu} \ . \tag{A.32}$$

Substituting ℓ back into equation (A.31) gives

$$\gamma_{nj(t')} = r_{nj} \ , \tag{A.33}$$

as was used in the iterative method in (2.63). Notice that this derivation also works for $\alpha = 1$, and hence we can determine an approximate $q(\mathbf{z}_n)$ for EP as well, if we so wish. The approximation for $q(\mathbf{z}_n)$ can be directly read from the responsibilities.

A.6 Normal-Wishart integrals

In chapter 3 a number of integrals, incorporating a Normal-Wishart distribution, need to be evaluated. For brevity, subscripts j and n are dropped, so that $\mathbf{x}$ implies a single observation $\mathbf{x}_n$. The Normal and Wishart distributions are

$$\mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, (v\boldsymbol{\Lambda})^{-1}) = \Big(\frac{v}{2\pi}\Big)^{\frac{d}{2}} |\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\mathsf{tr}[(\boldsymbol{\mu}-\mathbf{m})(\boldsymbol{\mu}-\mathbf{m})^{\top} v\boldsymbol{\Lambda}]\Big\} \tag{A.34}$$

$$\mathcal{W}(\boldsymbol{\Lambda}|a, \mathbf{B}) = \frac{|\mathbf{B}|^a}{\prod_{i=1}^d \Gamma(a + \frac{1-i}{2})} \pi^{\frac{-d(d-1)}{4}} |\boldsymbol{\Lambda}|^{a-\frac{d+1}{2}} \exp\Big\{ -\mathsf{tr}[\mathbf{B}\boldsymbol{\Lambda}]\Big\} \ . \tag{A.35}$$

The Normal-Wishart is the product of the above two distributions, where the Normal distribution is *conditional* on $\boldsymbol{\Lambda}$, the random variable on which a Wishart distribution is placed.

For dealing with the more general α-divergence, we evaluate moments under a 'power-posterior', i.e. a posterior with the likelihood term raised to some power, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})^{\alpha}$. Setting $\alpha = 1$ gives the true posterior. Let Z be the normalizer of the power posterior,

$$\begin{aligned}
Z &= \int \mathcal{W}(\boldsymbol{\Lambda}|a, \mathbf{B}) \int \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, (v\boldsymbol{\Lambda})^{-1}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})^{\alpha} \, d\boldsymbol{\mu} \, d\boldsymbol{\Lambda} \\
&= \int \mathcal{W}(\boldsymbol{\Lambda}|a, \mathbf{B}) \Big(\frac{v}{v+\alpha}\Big)^{d/2} |\boldsymbol{\Lambda}|^{\alpha/2} (2\pi)^{-\alpha d/2} e^{-\mathsf{tr}[\frac{1}{2}\frac{\alpha v}{v+\alpha}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\top}\boldsymbol{\Lambda}]} \\
&\quad \times \int \mathcal{N}\Big(\boldsymbol{\mu}\Big|\frac{v\mathbf{m}+\alpha\mathbf{x}}{v+\alpha}, \boldsymbol{\Lambda}^{-1}(v+\alpha)^{-1}\Big) \, d\boldsymbol{\mu} \, d\boldsymbol{\Lambda} \qquad \text{(A.36)} \\
&= \Big(\frac{v}{v+\alpha}\Big)^{d/2} (2\pi)^{-\alpha d/2} \frac{|\mathbf{B}|^a}{\prod_{i=1}^d \Gamma(a+\frac{1-i}{2})} \frac{\prod_{i=1}^d \Gamma(a+\frac{\alpha}{2}+\frac{1-i}{2})}{|\mathbf{B}+\frac{1}{2}\frac{\alpha v}{\alpha+v}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\top}|^{a+\alpha/2}} \\
&\quad \times \int \mathcal{W}\Big(\boldsymbol{\Lambda}\Big|a+\frac{\alpha}{2}, \mathbf{B}+\frac{1}{2}\frac{\alpha v}{\alpha+v}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\top}\Big) \, d\boldsymbol{\Lambda} \ . \qquad \text{(A.37)}
\end{aligned}$$

Equations (A.36) and (A.37) rely on factorizing the integrand such that factors dependent on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ occur in known distributions. Both integrals evaluate to one, but will be crucial in determining the power-posterior moments. Continuing,

$$
\begin{aligned}
Z = &\Big(\frac{v}{v+\alpha}\Big)^{d/2}(2\pi)^{-\alpha d/2}\frac{\prod_{i=1}^d \Gamma(a+\frac{\alpha}{2}+\frac{1-i}{2})}{\prod_{i=1}^d \Gamma(a+\frac{1-i}{2})} \\
&\times |\mathbf{B}|^a\left[|\mathbf{B}|\Big(1+\frac{1}{2}\frac{\alpha v}{\alpha+v}(\mathbf{x}-\mathbf{m})^\top\mathbf{B}^{-1}(\mathbf{x}-\mathbf{m})\Big)\right]^{-a-\alpha/2},
\end{aligned}
$$

where $|\mathbf{A}+\mathbf{x}\mathbf{y}^\top| = |\mathbf{A}|(1+\mathbf{y}^\top\mathbf{A}^{-1}\mathbf{x})$ was used. Now rearrange the exponent to be in the right form for a Student-t distribution.

$$
\begin{aligned}
Z = &\Big(\frac{\alpha v}{v+\alpha}\Big)^{d/2}\alpha^{-d/2}(2\pi)^{-\alpha d/2}\frac{\prod_{i=1}^d \Gamma(a+\frac{\alpha}{2}+\frac{1-i}{2})}{\prod_{i=1}^d \Gamma(a+\frac{1-i}{2})}|\mathbf{B}|^{-\alpha/2} \\
&\times\left(1+\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top\Big[\frac{v+\alpha}{\alpha v}\mathbf{B}\Big]^{-1}(\mathbf{x}-\mathbf{m})\right)^{-(2a-d+\alpha+d)/2} \\
= &\,\alpha^{-d/2}(2\pi)^{-\alpha d/2}\frac{\prod_{i=1}^d \Gamma(a+\frac{\alpha}{2}+\frac{1-i}{2})}{\prod_{i=1}^d \Gamma(a+\frac{1-i}{2})}\Big(\frac{2a-d+\alpha}{2}\Big)^{-d/2}|\mathbf{B}|^{(1-\alpha)/2}\Big|\frac{v+\alpha}{\alpha v}\frac{2\mathbf{B}}{2a-d+\alpha}\Big|^{-1/2} \\
&\times\left(1+\frac{1}{2a-d+\alpha}(\mathbf{x}-\mathbf{m})^\top\Big[\frac{v+\alpha}{\alpha v}\frac{2\mathbf{B}}{2a-d+\alpha}\Big]^{-1}(\mathbf{x}-\mathbf{m})\right)^{-(2a-d+\alpha+d)/2} \\
= &\,\alpha^{-d/2}(2\pi)^{(1-\alpha)d/2}|\mathbf{B}|^{(1-\alpha)/2}\frac{\prod_{i=1}^d \Gamma(a+\frac{\alpha}{2}+\frac{1-i}{2})}{\prod_{i=1}^d \Gamma(a+\frac{1-i}{2})}\frac{\Gamma(\frac{2a-d+\alpha}{2})}{\Gamma(\frac{[2a-d+\alpha]+d}{2})} \\
&\times[(2a-d+\alpha)\pi]^{-d/2}\Big|\frac{v+\alpha}{\alpha v}\frac{2\mathbf{B}}{2a-d+\alpha}\Big|^{-1/2}\frac{\Gamma(\frac{[2a-d+\alpha]+d}{2})}{\Gamma(\frac{2a-d+\alpha}{2})} \\
&\times\left(1+\frac{1}{2a-d+\alpha}(\mathbf{x}-\mathbf{m})^\top\Big[\frac{v+\alpha}{\alpha v}\frac{2\mathbf{B}}{2a-d+\alpha}\Big]^{-1}(\mathbf{x}-\mathbf{m})\right)^{-(2a-d+\alpha+d)/2} \\
= &\,\alpha^{-d/2}(2\pi)^{(1-\alpha)d/2}|\mathbf{B}|^{(1-\alpha)/2}\frac{\prod_{i=1}^d \Gamma(a+\frac{\alpha}{2}+\frac{1-i}{2})}{\prod_{i=1}^d \Gamma(a+\frac{1-i}{2})}\frac{\Gamma(\frac{2a-d+\alpha}{2})}{\Gamma(\frac{[2a-d+\alpha]+d}{2})} \\
&\times\mathcal{T}\Big(\mathbf{x}\,\Big|\,\mathbf{m},\,\frac{v+\alpha}{\alpha v}\frac{2\mathbf{B}}{2a-d+\alpha},\,2a-d+\alpha\Big)\,. \qquad \text{(A.38)}
\end{aligned}
$$

Substituting $\alpha = 1$ will give us an unscaled Student-t distribution, as the prefactor is one.

Moments

The moments of the power-posterior can be read from the Normal and Wishart distributions that occur respectively in equations (A.36) and (A.37). Notation $\langle\cdot|\mathbf{x}\rangle$ is used to indicate the moment under the 'power posterior' distribution, i.e. on including some likelihood term raised to the power α.

$$
\langle\boldsymbol{\Lambda}|\mathbf{x}\rangle = \frac{1}{Z}\int\boldsymbol{\Lambda}\,\mathcal{W}(\boldsymbol{\Lambda}|a,\mathbf{B})\int\mathcal{N}(\boldsymbol{\mu}|\mathbf{m},(v\boldsymbol{\Lambda})^{-1})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})^\alpha\,d\boldsymbol{\mu}\,d\boldsymbol{\Lambda}
$$

$$= \int \mathbf{\Lambda}\, \mathcal{W}\Big(\mathbf{\Lambda}\Big| a + \frac{\alpha}{2}, \mathbf{B} + \frac{1}{2}\frac{\alpha v}{\alpha + v}(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top\Big)\, d\mathbf{\Lambda}$$
$$= (a + \frac{\alpha}{2})\Big[\mathbf{B} + \frac{1}{2}\frac{\alpha v}{\alpha + v}(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top\Big]^{-1}, \tag{A.39}$$

similarly,

$$\langle \ln|\mathbf{\Lambda}|\, |\mathbf{x}\rangle = \int \ln|\mathbf{\Lambda}|\, \mathcal{W}\Big(\mathbf{\Lambda}\Big| a + \frac{\alpha}{2}, \mathbf{B} + \frac{1}{2}\frac{\alpha v}{\alpha + v}(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top\Big)\, d\mathbf{\Lambda}$$
$$= \sum_{i=1}^{d} \Psi\Big(a + \frac{\alpha}{2} + \frac{1-i}{2}\Big) - \ln\Big|\mathbf{B} + \frac{1}{2}\frac{\alpha v}{\alpha + v}(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top\Big|\,, \tag{A.40}$$

$$\langle \mathbf{\Lambda}\boldsymbol{\mu}|\mathbf{x}\rangle = \frac{1}{Z}\int \mathbf{\Lambda}\, \mathcal{W}(\mathbf{\Lambda}|a, \mathbf{B})\Big(\frac{v}{v+\alpha}\Big)^{d/2} |\mathbf{\Lambda}|^{\alpha/2}(2\pi)^{-\alpha d/2} e^{-\mathsf{tr}[\frac{1}{2}\frac{\alpha v}{v+\alpha}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^\top \mathbf{\Lambda}]}$$
$$\times \int \boldsymbol{\mu}\, \mathcal{N}\Big(\boldsymbol{\mu}\Big|\frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}, \mathbf{\Lambda}^{-1}(v+\alpha)^{-1}\Big)\, d\boldsymbol{\mu}\, d\mathbf{\Lambda}$$
$$= \langle \mathbf{\Lambda}|\mathbf{x}\rangle \frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}\,, \tag{A.41}$$

$$\langle \boldsymbol{\mu}^\top \mathbf{\Lambda}\boldsymbol{\mu}|\mathbf{x}\rangle = \frac{1}{Z}\iint \mathsf{tr}[\mathbf{\Lambda}\boldsymbol{\mu}\boldsymbol{\mu}^\top]\, \mathcal{W}(\mathbf{\Lambda}|a, \mathbf{B})\Big(\frac{v}{v+\alpha}\Big)^{d/2} |\mathbf{\Lambda}|^{\alpha/2}(2\pi)^{-\alpha d/2} e^{-\mathsf{tr}[\frac{1}{2}\frac{\alpha v}{v+\alpha}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^\top \mathbf{\Lambda}]}$$
$$\times \mathcal{N}\Big(\boldsymbol{\mu}\Big|\frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}, \mathbf{\Lambda}^{-1}(v+\alpha)^{-1}\Big)\, d\boldsymbol{\mu}\, d\mathbf{\Lambda}$$
$$= \int \mathsf{tr}\Bigg[\mathbf{\Lambda}\Big[\mathbf{\Lambda}^{-1}(v+\alpha)^{-1} + \Big(\frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}\Big)\Big(\frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}\Big)^\top\Big]\Bigg]$$
$$\times \mathcal{W}\Big(\mathbf{\Lambda}\Big| a + \frac{\alpha}{2}, \mathbf{B} + \frac{1}{2}\frac{\alpha v}{\alpha + v}(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top\Big)\, d\mathbf{\Lambda}$$
$$= \frac{d}{v+\alpha} + \Big(\frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}\Big)^\top \langle \mathbf{\Lambda}|\mathbf{x}\rangle \Big(\frac{v\mathbf{m} + \alpha\mathbf{x}}{v+\alpha}\Big)\,. \tag{A.42}$$

A.7 The matrix inversion lemma

The matrix inversion lemma is also known as the Woodbury-Sherman-Morrison formula, and states that

$$(\mathbf{Z} + \mathbf{U}\mathbf{W}\mathbf{V}^\top)^{-1} = \mathbf{Z}^{-1} - \mathbf{Z}^{-1}\mathbf{U}(\mathbf{W}^{-1} + \mathbf{V}^\top\mathbf{Z}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top\mathbf{Z}^{-1}\,, \tag{A.43}$$

if all the relevant inverses exist. Here $\mathbf{Z}$ is an $n \times n$ matrix, $\mathbf{W}$ has size $m \times m$, and both $\mathbf{U}$ and $\mathbf{V}$ are $n \times m$. If $m < n$ and a *low rank* perturbation is made to $\mathbf{Z}$, as we see in the left hand size of (A.43), then the computation of the inverse can be accelerated if $\mathbf{Z}^{-1}$ is known.

Bibliography

Amari, S. (1985). *Differential-geometrical Methods in Statistics*. Lecture Notes in Statistics. Berlin: Springer-Verlag.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, (pp. 21–30)., San Francisco, CA. Morgan Kaufmann Publishers.

Attias, H. (2000). A variational Bayesian framework for graphical models. In Solla, S. A., Leen, T. K., & Müller, K.-R. (Eds.), *Advances in Neural Information Processing Systems 12*, (pp. 209–215). MIT Press.

Beal, M. J. & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, *7*, 453–464.

Berg, B. A. (2000). Introduction to multicanonical Monte Carlo simulations. *Fields Institute Communications*, *26*, 1–24.

Bishop, C. M. (1999). Latent variable models. In Jordan, M. (Ed.), *Learning in Graphical Models*, (pp. 371–403). MIT Press.

Bishop, C. M. & Svensén, M. (2003). Bayesian hierarchical mixtures of experts. In Kjaerulff, U. & Meek, C. (Eds.), *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, (pp. 57–64). Morgan Kaufmann.

Chang, S., Dasgupta, N., & Carin, L. (2005). A Bayesian approach to unsupervised feature selection and density estimation using expectation propagation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, (pp. 1043–1050).

Corduneanu, A. & Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. In Richardson, T. & Jaakkola, T. (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, (pp. 27–34). Morgan Kaufmann.

de Freitas, N., Højen-Sørensen, P., Jordan, M. I., & Russell, S. (2001). Variational MCMC. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, (pp. 120–127).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*(1), 1–38.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, *56*(2), 363–375.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222.

Earl, D. J. & Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, *7*, 3910–3916.

Ferkinghoff-Borg, J. (2002). *Monte Carlo Methods in Complex Systems*. PhD thesis, University of Copenhagen.

Feynman, R. P. (1972). *Statistical Mechanics: a set of lectures*. Benjamin.

Frey, B., Patrascu, R., Jaakkola, T., & Moran, J. (2000). Sequentially fitting inclusive trees for inference in noisy-or networks. In *Advances in Neural Information Processing Systems 13*, (pp. 493–499). MIT Press.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Ghahramani, Z. & Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In Solla, S. A., Leen, T. K., & Müller, K.-R. (Eds.), *Advances in Neural Information Processing Systems 12*, (pp. 449–455). MIT Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Gregory, P. (2005). *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Heskes, T. & Zoeter, O. (2002). Expectation Propogation for approximate inference in dynamic Bayesian networks. In *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*, (pp. 216–223).

Hinton, G. E. & van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, (pp. 5–13). ACM Press.

Iba, Y. (2001). Extended ensemble Monte Carlo. *International Journal of Modern Physics C*, *12*(5), 623–656.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Kofke, D. A. (2002). On the acceptance probability of replica-exchange Monte Carlo trials. *Journal of Chemical Physics*, *117*(15), 6911–6914.

Kofke, D. A. (2004). Erratum: "On the acceptance probability of replica-exchange Monte Carlo trials" [J. Chem. Phys. 117, 6911 (2002)]. *Journal of Chemical Physics*, *120*(22), 10852.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447.

MacKay, D. J. C. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

MacKay, D. J. C. (1998). Choice of basis for Laplace approximation. *Machine Learning*, *33*(1), 77–86.

MacKay, D. J. C. (2001). A problem with variational free energy minimization. Technical report, Department of Physics, University of Cambridge.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Mengersen, K. L. & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, *24*(1), 101–121.

Metropolis, N., Rosenbluth, A. W., Teller, M. N., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Minka, T. (2000). Estimating a Dirichlet distribution. Available online at http://research.microsoft.com/ minka/papers/.

Minka, T. (2001a). *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology.

Minka, T. (2001b). Using lower bounds to approximate integrals. Available online at http://research.microsoft.com/ minka/papers/.

Minka, T. (2005). Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, Cambridge, UK.

Minka, T. & Ghahramani, Z. (2003). Expectation propagation for infinite mixtures. NIPS'03 Workshop on Nonparametric Bayesian Methods and Infinite Models.

Minka, T. P. (2001c). Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, (pp. 362–369).

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*(2), 125–139.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*, 705–767.

Neal, R. M. & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I. (Ed.), *Learning in Graphical Models*, (pp. 355–368). Kluwer Academic Publishers.

Opper, M. (2006). An approximate inference approach for the PCA reconstruction error. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1035–1042). Cambridge, MA: MIT Press.

Opper, M. & Winther, O. (2001a). Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Physical Review E*, *64*(5), 056131.

Opper, M. & Winther, O. (2001b). Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Physical Review Letters*, *86*(17), 3695–3699.

Opper, M. & Winther, O. (2005a). Expectation consistent approximate inference. *Journal of Machine Learning Research*, *6*, 2177–2204.

Opper, M. & Winther, O. (2005b). Expectation consistent free energies for approximate inference. In Thrun, S., Saul, L., & Schölkopf, B. (Eds.), *Advances in Neural Information Processing Systems 17*, (pp. 1001–1008). MIT Press.

Rabiner, L. & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16.

Rasmussen, C. E. & Ghahramani, Z. (2001). Occam's razor. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society (B)*, *59*(4), 731–792.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods* (Second ed.). Springer.

Roberts, S. J., Husmeier, D., Rezek, I., & Penny, W. (1998). Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1133–1142.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, *85*(–), 617–624.

Saul, L., Jaakkola, T., & Jordan, M. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, *4*, 61–76.

Skilling, J. (1998). Probabilistic data analysis: an introductory guide. *Journal of Microscopy*, *190*(1), 28–36.

Swendsen, R. H. & Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, *57*(21), 2607–2609.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*(11), 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, *22*(4), 1701–1762.

Waterhouse, S. R., MacKay, D. J. C., & Robinson, A. J. (1996). Bayesian methods for mixtures of experts. In Touretzky, D. S., Mozer, M. C., & Hasselmo, M. E. (Eds.), *Neural Information Processing Systems*, (pp. 351–357). MIT Press.

Winther, O. (2007). Personal communication.

Ypma, A. & Heskes, T. (2003). Iterated extended Kalman smoothing with Expectation-Propagation. In *Proceedings NNSP*, (pp. 219–228).

Yuille, A. (2002). CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, *14*, 1691–1722.