Using lower bounds to approximate integrals

Thomas P. Minka

February 22, 2001

Abstract

Variational lower bounds are a simple and efficient way to approximate Bayesian integrals. By bounding the integrand at every point, we obtain a bound on the integral value. Variational Bayes (Attias, 1999) and convex duality methods (Jaakkola & Jordan 1999a; 1999b) are of this type, and—as shown in this paper—so is the approximation method of Cheeseman & Stutz (1996). In each case, EM is used to maximize the bound. The EM updates are derived for Gaussian mixtures and multinomial mixtures with conjugate priors. Experimental results show the lower bound method to be less accurate than Laplace's method, but often simpler.

1 Introduction

Bayesian inference often involves difficult integrals that must be evaluated numerically. For example, in classification we want to compute the predictive density $p(c|x, D)$, the distribution over class index c given new measurement x and training data D. This requires an integral over parameters θ: $p(c|x, D) = \int_\theta p(c|x, \theta)p(\theta|D)d\theta$. In regression, we want to compute $E[y|x, D] = \int_y y \int_\theta p(y|x, \theta)p(\theta|D)d\theta dy$. In density estimation, we want $p(x|D) = \int_\theta p(x|\theta)p(\theta|D)d\theta$, the predictive probability of a new measurement. In model selection, we want to compute $p(D|M) = \int_\theta p(D|\theta)p(\theta|M)d\theta$, the probability of the training data given the model M, also known as the evidence for M or likelihood for M.

The simplest and most common way to approximate such integrals is to evaluate the integrand at a single value of θ: the estimate of θ. Maximum likelihood, maximum a posteriori, and minimum bias/variance are typical criteria for constructing such estimates. Much effort has gone into designing optimization algorithms for such criteria. For example, the EM algorithm is useful for ML and MAP estimation in hidden variable models, such as mixture models and hidden Markov models. (For simple models like Gaussians and multinomials, most integrals are tractable so there is little excuse for using parameter estimates with such models.)

At the other extreme, we can approximate the integrals using general numerical techniques such as interpolative quadrature and Monte Carlo. This approach is attractive since the results are much closer to the true value of the integral. However, it is also very computationally intensive.

Between these two extremes, we can try approximating the integrand until the integral becomes tractable. This is the approach taken in this paper. A classical instance of this is Laplace's method, which is based on Taylor expansion of the logarithm of the integrand (see appendix A).

Unfortunately, Laplace's method is unweildy in high dimensions, requiring a large matrix of cross-derivatives. It also requires the integrand to be roughly log-quadratic, which may not be true for likelihoods with hidden variables (see section 6), especially when near the boundary of the parameter space.

Lower bound integration is simpler and more stable than Laplace's method. The idea is to bound the integrand from above or below, reducing the integration problem to an optimization problem over the bound: making the bound as tight as possible. No parameter estimate is needed; the quality of the integral is optimized directly. There are no problems at the boundary of the parameter space.

Variational bounds enjoy increasing attention in the statistical learning literature. Jordan et al. (1999) survey a variety of work on bounding posterior probabilities over hidden variables in a graphical model, with particular emphasis on missing data. MacKay (1996) extends this to include joint probabilities over parameters and missing data. Waterhouse et al. (1995) and Attias (1999) show how variational bounds can approximate other Bayesian quantities such as the evidence and predictive density. Penny & Roberts (2000) show that the performance for model selection exceeds BIC.

This paper corrects the Gaussian mixture equations given by Attias (1999) and parallels them with the multinomial case (MacKay, 1996). Second, it shows that the algorithms of MacKay and Attias are EM algorithms paralleling those of Jaakkola & Jordan (1999a; 1999b).

2 EM for lower bound integration

Schematically, the procedure is as follows. We want to evaluate some integral

$$F = \int_\theta f(\theta) d\theta \tag{1}$$

We find a lower (or upper) bound g such that

$$f(\theta) \geq g(\theta, \phi) \qquad \text{for all } \phi \tag{2}$$

The bound is chosen so that its integral

$$G(\phi) = \int_\theta g(\theta, \phi) d\theta \leq F \tag{3}$$

is tractable. The remaining task is to maximize $G(\phi)$ over ϕ, so that it is as close as possible to F. We could use any optimization routine for this, but a particularly convenient choice is EM. In this application of EM, θ is considered a hidden variable and ϕ the parameter. The algorithm consists of iteratively maximizing

$$
\begin{aligned}
& \int_\theta q(\theta) \log g(\theta, \phi^{new}) d\theta && (4) \\
\text{where } q(\theta) &= \frac{g(\theta, \phi^{old})}{\int_\theta g(\theta, \phi^{old}) d\theta} && (5)
\end{aligned}
$$

In many cases, the result looks like an EM algorithm for maximum likelihood (a maximization over θ), except that the E-step and M-step are swapped, since θ is now a hidden variable. EM was explicitly used to fit variational bounds by Jaakkola & Jordan (1999a; 1999b) and implicitly used by others such as Ghahramani (1995). The "EM-like" algorithm given by Waterhouse et al. (1995) and Attias (1999) is similar, but not identical because they reverse the E-step and M-step.

In the "variational Bayes" method (Waterhouse et al., 1995; MacKay, 1996; Attias, 1999), the bound $g(\theta, \phi)$ is defined implicitly through hidden variables. Start by writing $f(\theta)$ in terms of $h(\theta, \mathbf{y})$ ($\mathbf{y}$ is the hidden variable):

$$
f(\theta) = \int_{\mathbf{y}} h(\theta, \mathbf{y}) d\mathbf{y} \tag{6}
$$

Apply Jensen's inequality to get

$$
\begin{aligned}
F &= \int_{\theta,\mathbf{y}} h(\theta, \mathbf{y}) d\mathbf{y} d\theta && (7) \\
&\geq \exp\left(\int_{\theta,\mathbf{y}} q(\theta, \mathbf{y}) \log \frac{h(\theta, \mathbf{y})}{q(\theta, \mathbf{y})} d\mathbf{y} d\theta \right) && (8) \\
\text{as long as } \int_{\theta,\mathbf{y}} q(\theta, \mathbf{y}) d\mathbf{y} d\theta &= 1 && (9)
\end{aligned}
$$

The variational Bayes method constrains $q(\theta, \mathbf{y})$ to factor into separate functions for θ and for $\mathbf{y}$:

$$
q(\theta, \mathbf{y}) = q_\theta(\theta) q_y(\mathbf{y}) \tag{10}
$$

with no other constraints on functional form. The q_θ and q_y functions are iteratively optimized to maximize the value of the bound. To see that this is equivalent to (3), note that for any q_y we can solve analytically for the optimal q_θ, which is

$$
\begin{aligned}
q_\theta(\theta) &= \frac{g(\theta)}{\int_\theta g(\theta) d\theta} && (11) \\
\text{where } g(\theta) &= \exp\left(\int_{\mathbf{y}} q_y(\mathbf{y}) \log \frac{h(\theta, \mathbf{y})}{q_y(\mathbf{y})} d\mathbf{y} \right) && (12)
\end{aligned}
$$

When we substitute this q_θ, the bound becomes

$$
F \geq \int_\theta g(\theta) d\theta \tag{13}
$$

Here ϕ is the function q_y, i.e. the distribution over missing data.

Lower bound integration generalizes variational Bayes since the bound does not have to be a Jensen bound defined by hidden varables, in the same way that lower bound optimization generalizes EM (Minka, 1998). Variational Bayes is sometimes interpreted as minimizing the KL-divergence between an approximate and true distribution. Lower bound integration, in its general form, does not have this interpretation.

There are other approaches to bounding an integral that do not involve bounding the integrand. For example, we can apply Jensen's inequality without introducing hidden variables:

$$\int_\theta f(\theta)d\theta \geq \exp\left(\int_\theta q(\theta)\log\frac{f(\theta)}{q(\theta)}d\theta\right) \tag{14}$$

$$\text{as long as } \int_\theta q(\theta)d\theta = 1 \tag{15}$$

To optimize the bound, we optimize $q(\theta)$ over some convenient class of densities, e.g. Gaussian. See e.g. Barber & Bishop (1997) and Jaakkola & Jordan (1999c). This approach takes advantage of the fact that $\log(f(\theta))$ is often easy to integrate when $f(\theta)$ is not. But this is not always true. For the mixture problems considered in this paper, f is a product of sums, which does not simplify under a logarithm. Consequently, this method is not used in this paper.

3 Posterior expectations

For approximating posterior expectations, we have the same options with the lower method as we have with Laplace's method. The expectation of $f(\theta)$ over the posterior $p(\theta|D)$ is

$$E[f(\theta)|D] = \int_\theta f(\theta)p(\theta|D)d\theta = \frac{\int_\theta f(\theta)p(\theta, D)d\theta}{\int_\theta p(\theta, D)d\theta} \tag{16}$$

In the *ratio method*, we approximate each integral with a separate variational bound, as suggested by Attias (1999). When there are many expectations to evaluate, this may be too expensive. For example, if we want the predictive density

$$p(x|D) = \int_\theta p(x|\theta)p(\theta|D)d\theta = E[p(x|\theta)|D] = \frac{p(x, D)}{p(D)} \tag{17}$$

then the ratio method would require a separate integral approximation for every location x. A simpler approach is to bound $p(\theta, D)$ once and then use the same bound to evaluate the numerator integral for every x. This is the *expectation method*, since it uses expectations over a single bound (Waterhouse et al., 1995).

4 Bounding the evidence of a mixture

A finite mixture model has the form

$$p(x|\theta) = \sum_c p(x|c,\theta)p(c|\theta) \tag{18}$$

This section focuses on bounding the evidence

$$p(D) = \int_\theta p(D|\theta)p(\theta)d\theta \tag{19}$$

$$p(D|\theta) = \prod_i p(x_i|\theta) \tag{20}$$

$$= \prod_i \left(\sum_j p(x_i|c_i = j, \theta)p(c_i = j|\theta) \right) \tag{21}$$

The resulting bound can also be used to approximate posterior expectations via the expectation method (section 3).

The bound used in this section comes from Jensen's inequality and is the usual EM bound:

$$p(x_i|\theta) \geq \prod_j \left(\frac{p(x_i, c_i = j|\theta)}{q_{ij}} \right)^{q_{ij}} \tag{22}$$

$$\text{where } \sum_j q_{ij} = 1 \tag{23}$$

As long as the mixture components are simple, the value of the integral is now easy to compute. The EM bound is not necessarily the best bound to use, or even a good bound to use, but it is simple.

The running example will be a mixture of multinomial distributions. The Gaussian mixture case is handled in section 5.

4.1 Fixed components

We start by making the component densities $p(x|c)$ fixed so that the only parameters are the mixing weights $w_j = p(c = j)$. For simplicity, let the prior be Dirichlet:

$$p(\mathbf{w}) = \mathcal{D}(\alpha_1, ..., \alpha_J) \tag{24}$$

$$p(D) = \int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{25}$$

$$p(x|\mathbf{w}) = \sum_j p(x|c=j)p(c=j|\mathbf{w}) \quad (26)$$
$$p(c=j|\mathbf{w}) = w_j \quad (27)$$
$$p(\mathbf{w}) = \mathcal{D}(\alpha_j) \quad (28)$$

Applying the bound gives

$$p(D) = \int_{\mathbf{w}} p(\mathbf{w}) \prod_i \sum_j p(x_i|c_i=j) w_j d\mathbf{w} \geq \left(\int_{\mathbf{w}} p(\mathbf{w}) \prod_j w_j^{\sum_i q_{ij}} d\mathbf{w} \right) \prod_{ij} \left(\frac{p(x_i|c_i=j)}{q_{ij}} \right)^{q_{ij}} \quad (29)$$

The integral evaluates to

$$\int_{\mathbf{w}} p(\mathbf{w}) \prod_j w_j^{\sum_i q_{ij}} d\mathbf{w} = \frac{\Gamma(\sum_j \alpha_j)}{\Gamma(N + \sum_j \alpha_j)} \prod_j \frac{\Gamma(\alpha_j + \sum_i q_{ij})}{\Gamma(\alpha_j)} \quad (30)$$

There remains the issue of choosing q_{ij}. As described in section 2, we will use EM to maximize the integral of the bound. As a function of $\mathbf{w}$, the bound corresponds to a scaled Dirichlet distribution. So immediately we know that the E-step is

$$q(\mathbf{w}) = \mathcal{D}(\alpha_j + \sum_i q_{ij}) \quad (31)$$

and the M-step maximizes (using λ_i to enforce (23))

$$\sum_{ij} q_{ij} \log \frac{p(x_i|c_i=j)}{q_{ij}} + \sum_{ij} q_{ij} \left(\int_{\mathbf{w}} q(\mathbf{w}) \log w_j d\mathbf{w} \right) + \sum_i \lambda_i (\sum_j q_{ij} - 1) \quad (32)$$

whose maximum is

$$q_{ij} \propto p(x_i|c_i=j) \exp\left(\int_{\mathbf{w}} q(\mathbf{w}) \log w_j d\mathbf{w}\right) \quad (33)$$
$$\int_{\mathbf{w}} q(\mathbf{w}) \log w_j d\mathbf{w} = \Psi(\alpha_j + \sum_i q_{ij}) - \Psi(\sum_k \alpha_k + N) \quad (34)$$

The resulting EM algorithm is a simple iteration since $p(x_i|c_i=j)$ is fixed. If we were doing MAP estimation for $\mathbf{w}$, then the EM iteration would be

$$q_{ij} \propto p(x_i|c_i=j) \frac{\alpha_j - 1 + \sum_i q_{ij}}{\sum_k \alpha_k - K + N} \quad (35)$$

so the soft assignments q_{ij} in the variational bound will be slightly different than the ones used for MAP.

To better understand what this bound is doing, define the estimate at convergence

$$\hat{w}_j = \exp(\Psi(\alpha_j + \sum_i q_{ij}) - \Psi(\sum_k \alpha_k + N)) \quad (36)$$

The denominator of q_{ij} is $\sum_j p(x_i|c_i = j)\hat{w}_j = p(x_i|\mathbf{w} = \hat{\mathbf{w}})$. So the likelihood of $\mathbf{w} = \hat{\mathbf{w}}$ is

$$p(D|\mathbf{w} = \hat{\mathbf{w}}) = \prod_i p(x_i|\mathbf{w} = \hat{\mathbf{w}}) = \prod_{ij} \left(\frac{p(x_i|c_i = j)\hat{w}_j}{q_{ij}} \right)^{q_{ij}} \tag{37}$$

and we can rewrite the bound as the likelihood at $\mathbf{w} = \hat{\mathbf{w}}$ times a correction factor:

$$p(D) \geq p(D|\mathbf{w} = \hat{\mathbf{w}}) \frac{\int_{\mathbf{w}} p(\mathbf{w}) \prod_j w_j^{\sum_i q_{ij}} d\mathbf{w}}{\prod_j \hat{w}_j^{\sum_i q_{ij}}} \tag{38}$$

As the sample size gets large, the correction factor approaches 1, and $\hat{\mathbf{w}}$ approaches the MAP estimate of $\mathbf{w}$. The approximation (38) is the same as the heuristic proposal made by Cheeseman & Stutz (1996), when the mixture components are fixed and $\hat{\mathbf{w}}$ is used as the estimate of $\mathbf{w}$. The next section shows that this equivalence holds generally, even when the components are not fixed. This is an interesting result because while the Cheeseman-Stutz approximation is known to have good empirical performance (Chickering & Heckerman, 1996; Kontkanen et al., 1997), it was not known to be a bound on the exact model evidence. This equivalence also justifies using a smoothed estimate instead of a maximum-likelihood estimate in the Cheeseman-Stutz approximation (Chickering & Heckerman, 1996).

4.2 Fixed weights

Now let the mixing weights be fixed and the component parameters θ_j be unknown. Applying the bound gives

$$p(D) = \int_{\theta^1 \ldots \theta^J} \prod_i \sum_j p(x_i|c_i = j)p(c_i = j) \;\; \geq \;\; \prod_j E_j \prod_{ij} \left(\frac{p(c_i = j)}{q_{ij}} \right)^{q_{ij}} \tag{39}$$

$$E_j \;\; = \;\; \int_{\theta^j} p(\theta^j) \prod_i p(x_i|c_i = j, \theta^j)^{q_{ij}} d\theta^j \tag{40}$$

For simple component densities, the integral E_j is easy to compute. It is the usual evidence formula for that density but with weighted data. For a multinomial component where θ^j is a vector of probabilities, we would have

$$p(x_i|c_i = j, \theta^j) = \prod_k (\theta^j_k)^{N^i_k} \tag{41}$$

where N^i_k is the number of occurrences of symbol k in sample i. If the prior on θ is Dirichlet:

$$p(\theta^j) = \mathcal{D}(\alpha_1, ..., \alpha_K) \tag{42}$$

then

$$E_j = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + \sum_{ik} q_{ij} N^i_k)} \prod_k \frac{\Gamma(\alpha_k + \sum_i q_{ij} N^i_k)}{\Gamma(\alpha_k)} \tag{43}$$

Now we use EM to optimize the q_{ij}. Thanks to the bound, the parameter posterior decouples: $q(\theta_1, ..., \theta_J) = q(\theta_1) \cdots q(\theta_J)$. By inspection we see that the E-step is

$$q(\theta^j) = \mathcal{D}(M_k^j) \tag{44}$$

$$M_k^j = \alpha_k + \sum_i q_{ij} N_k^i \tag{45}$$

and the M-step maximizes

$$\sum_{ij} q_{ij} \left(\int_{\theta^j} q(\theta^j) \log p(x_i|c_i = j, \theta^j) d\theta^j \right) + \sum_{ij} q_{ij} \log \frac{p(c_i = j)}{q_{ij}} + \sum_i \lambda_i \left(\sum_j q_{ij} - 1 \right) \tag{46}$$

whose maximum is

$$q_{ij} \propto p(c_i = j) \exp\left(\int_{\theta^j} q(\theta^j) \log p(x_i|c_i = j, \theta^j) d\theta^j\right) \tag{47}$$

$$\int_{\theta^j} q(\theta^j) \log p(x_i|c_i = j, \theta^j) d\theta^j = \sum_k N_k^i \left(\Psi(M_k^j) - \Psi(\sum_k M_k^j) \right) \tag{48}$$

For MAP estimation, the EM algorithm would have been

$$q_{ij} \propto p(c_i = j) \prod_k \left(\frac{M_k^j - 1}{\sum_k M_k^j - K} \right)^{N_k^i} \tag{49}$$

This iteration uses a single estimate for θ^j, while (47) uses the full distribution over θ^j. Interestingly, some authors (Nigam et al., 1998; Monti & Cooper, 1999) have adopted the convention of using

$$q_{ij} \propto p(c_i = j) \prod_k \left(\frac{M_k^j}{\sum_k M_k^j} \right)^{N_k^i} \tag{50}$$

for the iteration, which corresponds to using a predictive estimate for θ^j instead of a MAP estimate (Minka, 1999). This can be justified in terms of (47) since

$$\frac{\exp(\Psi(M_k^j))}{\exp(\Psi(\sum_k M_k^j))} \approx \frac{M_k^j}{\sum_k M_k^j} \tag{51}$$

As in the last section, we can define the estimate at convergence

$$\hat{\theta}_k^j = \exp(\Psi(M_k^j) - \Psi(\sum_k M_k^j)) \tag{52}$$

The denominator of q_{ij} is $p(x_i|\theta = \hat{\theta})$. So the likelihood of $\theta = \hat{\theta}$ is

$$p(D|\theta = \hat{\theta}) = \prod_{ij} \left(\frac{p(x_i|c_i = j, \hat{\theta}_j) p(c_i = j)}{q_{ij}} \right)^{q_{ij}} \tag{53}$$

and we can rewrite the bound as

$$p(D) \geq p(D|\theta = \hat{\theta}) \prod_j \frac{E_j}{\prod_i p(x_i|c_i = j, \hat{\theta}_j)^{q_{ij}}} \tag{54}$$

which again is equivalent to the approximation of Cheeseman & Stutz (1996).

4.3 Unknown weights and components

This is a straightforward combination of the previous results. The bound is

$$p(D) \geq \left(\prod_j E_j\right)\left(\int_{\mathbf{w}} p(\mathbf{w}) \prod_j w_j^{\sum_i q_{ij}} d\mathbf{w}\right) \prod_{ij} \left(\frac{1}{q_{ij}}\right)^{q_{ij}} \tag{55}$$

The E-steps are (31) and (44). The M-step maximizes

$$\sum_{ij} q_{ij} \left(\int_{\theta^j} q(\theta^j) \log p(x_i|c_i = j, \theta^j) d\theta^j\right) + \sum_{ij} q_{ij} \left(\int_{\mathbf{w}} q(\mathbf{w}) \log w_j d\mathbf{w}\right) - \sum_{ij} q_{ij} \log q_{ij} + \sum_i \lambda_i \left(\sum_j q_{ij} - 1\right) \tag{56}$$

whose maximum is

$$q_{ij} \propto \exp\left(\int_{\theta^j} q(\theta^j) \log p(x_i|c_i = j, \theta^j) d\theta^j + \int_{\mathbf{w}} q(\mathbf{w}) \log w_j d\mathbf{w}\right) \tag{57}$$

5 Mixture of Gaussians

This section derives the EM algorithm for bounding the likelihood of a mixture of Gaussians. The handling of unknown mixture weights is identical to the previous section, so for simplicity of exposition let the weights be fixed. The data model is

$$p(\mathbf{x}|c = j, \mathbf{m}^j, \mathbf{V}^j) \sim \mathcal{N}(\mathbf{m}^j, \mathbf{V}^j) \tag{58}$$

$$= \frac{1}{|2\pi \mathbf{V}^j|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{m}^j)^{\mathrm{T}} (\mathbf{V}^j)^{-1} (\mathbf{x} - \mathbf{m}^j)) \tag{59}$$

First suppose that the variance parameter of each mixture component is known, so we only need to integrate over the means $\{\mathbf{m}^1, ..., \mathbf{m}^J\}$. Let's focus on a particular component j, so that superscript j is implicit. If the prior is conjugate:

$$p(\mathbf{m}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{V}_0) \tag{60}$$

$$\prod_i p(\mathbf{x}_i|c_i = j, \mathbf{m})^{q_{ij}} = \frac{|2\pi \mathbf{V}/K|^{1/2}}{|2\pi \mathbf{V}|^{K/2}} \mathcal{N}(\mathbf{m}; \bar{\mathbf{x}}, \mathbf{V}/K) \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) \tag{61}$$

$$\text{where } K = \sum_i q_{ij} \tag{62}$$

$$\bar{\mathbf{x}} = \frac{1}{K} \sum_i q_{ij} \mathbf{x}_i \tag{63}$$

$$\mathbf{S} = \sum_i q_{ij} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \tag{64}$$

The bound is (39) with

$$E_j = \int_{\mathbf{m}} p(\mathbf{m}) \prod_i p(\mathbf{x}_i | c_i = j, \mathbf{m})^{q_{ij}} d\mathbf{m} \tag{65}$$

$$= \frac{1}{|2\pi\mathbf{V}|^{(K-1)/2} K^{d/2}} \mathcal{N}(\bar{\mathbf{x}}; \mathbf{m}_0, \mathbf{V}/K + \mathbf{V}_0) \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{V}^{-1})) \tag{66}$$

where d is the dimensionality of $\mathbf{x}$. The terms in $\mathbf{m}$ are $\mathcal{N}(\bar{\mathbf{x}}, \mathbf{V}/K)\mathcal{N}(\mathbf{m}_0, \mathbf{V}_0)$, so E-step is

$$\mathbf{V}_m = ((\mathbf{V}/K)^{-1} + \mathbf{V}_0^{-1})^{-1} = (\mathbf{V}/K)(\mathbf{V}/K + \mathbf{V}_0)^{-1}\mathbf{V}_0 \tag{67}$$

$$\hat{\mathbf{m}} = \mathbf{V}_m((\mathbf{V}/K)^{-1}\bar{\mathbf{x}} + \mathbf{V}_0^{-1}\mathbf{m}_0) \tag{68}$$

$$q(\mathbf{m}) \sim \mathcal{N}(\hat{\mathbf{m}}, \mathbf{V}_m) \tag{69}$$

The M-step is (47) with

$$\int_{\mathbf{m}} q(\mathbf{m}) \log p(\mathbf{x}_i | c_i = j, \mathbf{m}) d\mathbf{m} = \tag{70}$$

$$-\frac{1}{2}\log|2\pi\mathbf{V}| - \frac{1}{2}E[(\mathbf{x}_i - \mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{m})] = \tag{71}$$

$$-\frac{1}{2}\log|2\pi\mathbf{V}| - \frac{1}{2}(\mathbf{x}_i - \hat{\mathbf{m}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}_i - \hat{\mathbf{m}}) - \frac{1}{2}\mathrm{tr}(\mathbf{V}_m\mathbf{V}^{-1}) \tag{72}$$

Again, this corresponds to a simple correction of the usual MAP iteration for $\mathbf{m}$. The correction is $-\frac{1}{2}\mathrm{tr}(\mathbf{V}_m\mathbf{V}^{-1})$, which for large $\mathbf{V}_0$ (a noninformative prior) is $-\frac{d}{2K}$. It gives extra weight to big clusters, whose means are well-defined. In doing so, it will reduce the total number of clusters used in the model. To understand why this should be so, note that there are two competing factors which define the area of a bound: the height of the bound and the width of the bound. For maximum height, we should use MAP. For maximum width, we should assign all points to one cluster, leaving the other clusters unconstrained by the data (less constraints imply a wider distribution). The bound with maximum area is therefore a compromise between these extremes. See section 6 for examples.

Now let the means be known and variances unknown. We need to integrate over $\{\mathbf{V}_1, ..., \mathbf{V}_J\}$. If the prior is conjugate:

$$p(\mathbf{V}) \sim \mathcal{W}^{-1}(\mathbf{S}_0, N_0) \tag{73}$$

$$= \frac{1}{\Gamma_d(N_0/2)\,|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{S}_0}{2}\right|^{N_0/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}_0)) \tag{74}$$

$$\Gamma_d(n/2) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((n+1-i)/2) \tag{75}$$

The bound is (39) with (omitting superscript j)

$$E_j = \int_{\mathbf{V}} p(\mathbf{V}) \prod_i p(\mathbf{x}_i | c_i = j, \mathbf{m})^{q_{ij}} d\mathbf{V} \tag{76}$$

$$
\begin{aligned}
&= \int_{\mathbf{V}} p(\mathbf{V}) \frac{1}{|2\pi\mathbf{V}|^{K/2}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{S}_m\mathbf{V}^{-1}))d\mathbf{V} && (77)\\
&= \frac{\Gamma_d((K+N_0)/2)}{\pi^{Kd/2}\Gamma_d(N_0/2)} |\mathbf{S}_0|^{N_0/2} |\mathbf{S}_m + \mathbf{S}_0|^{-(K+N_0)/2} && (78)\\
\mathbf{S}_m &= \sum_i q_{ij}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathrm{T}} && (79)
\end{aligned}
$$

The E-step is

$$
q(\mathbf{V}) \sim \mathcal{W}^{-1}(\mathbf{S}_m + \mathbf{S}_0, K + N_0) \quad (80)
$$

For the M-step, we use

$$
\begin{aligned}
\int_{\mathbf{V}} q(\mathbf{V}) \log p(\mathbf{x}_i|c_i = j, \mathbf{V})d\mathbf{V} &= -\frac{d}{2}\log(2\pi) - \frac{1}{2}E[\log|\mathbf{V}|] - \frac{1}{2}(\mathbf{x}_i - \mathbf{m})^{\mathrm{T}} E[\mathbf{V}^{-1}](\mathbf{x}_i - \mathbf{m}) && (81)\\
E[\log|\mathbf{V}|] &= \log|\mathbf{S}_m + \mathbf{S}_0| - d\log(2) - \sum_{i=1}^{d} \Psi((K + N_0 + 1 - i)/2) && (82)\\
E[\mathbf{V}^{-1}] &= (K + N_0)(\mathbf{S}_m + \mathbf{S}_0)^{-1} && (83)
\end{aligned}
$$

A MAP iteration would be the same except (82) is replaced by

$$
\log|\mathbf{S}_m + \mathbf{S}_0| - d\log(K + N_0 + d + 1) \quad (84)
$$

and (83) is replaced by $(K + N_0 + d + 1)(\mathbf{S}_m + \mathbf{S}_0)^{-1}$. The difference between (82) over (84) is a preference for bigger clusters, just as in the previous case.

Now let both means and variances be unknown. If the prior is conjugate:

$$
p(\mathbf{m}, \mathbf{V}) \sim \mathcal{N}(\mathbf{m}; \mathbf{m}_0, \mathbf{V}/K_0)\mathcal{W}^{-1}(\mathbf{V}; \mathbf{S}_0, N_0) \quad (85)
$$

The bound uses

$$
\begin{aligned}
E_j &= \int_{\mathbf{m},\mathbf{V}} p(\mathbf{m}, \mathbf{V}) \prod_i p(\mathbf{x}_i|c_i = j, \mathbf{m}, \mathbf{V})^{q_{ij}} d\mathbf{m}d\mathbf{V} && (86)\\
&= \int_{\mathbf{V}} \frac{p(\mathbf{V})}{|2\pi\mathbf{V}|^{K/2}} \left(\frac{K_0}{K + K_0}\right)^{d/2} \exp(-\frac{1}{2}\mathrm{tr}(\hat{\mathbf{S}}\mathbf{V}^{-1}))d\mathbf{V} && (87)\\
&= \frac{\Gamma_d((K+N_0)/2)}{\pi^{Kd/2}\Gamma_d(N_0/2)} \left(\frac{K_0}{K + K_0}\right)^{d/2} |\mathbf{S}_0|^{N_0/2} \left|\hat{\mathbf{S}}\right|^{-(K+N_0)/2} && (88)\\
\hat{\mathbf{S}} &= \mathbf{S} + \mathbf{S}_0 + \frac{K_0 K}{K_0 + K}(\bar{\mathbf{x}} - \mathbf{m}_0)(\bar{\mathbf{x}} - \mathbf{m}_0)^{\mathrm{T}} && (89)
\end{aligned}
$$

The E-step is

$$
\begin{aligned}
\hat{\mathbf{m}} &= (K + K_0)^{-1}(K\bar{\mathbf{x}} + K_0\mathbf{m}_0) && (90)\\
q(\mathbf{m}, \mathbf{V}) &\sim \mathcal{N}(\mathbf{m}; \hat{\mathbf{m}}, \mathbf{V}/(K_0 + K))\mathcal{W}^{-1}(\hat{\mathbf{S}}, K + N_0) && (91)
\end{aligned}
$$

The parameter posterior decouples across the classes. For the M-step, we use

$$\int_{\mathbf{m},\mathbf{V}} q(\mathbf{m},\mathbf{V}) \log p(\mathbf{x}_i|c_i=j,\mathbf{m},\mathbf{V}) d\mathbf{m} d\mathbf{V} = (\text{const.}) - \frac{1}{2}E[\log|\mathbf{V}|] - \frac{1}{2}E[(\mathbf{x}_i-\mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{m})] \quad (92)$$

$$E[\log|\mathbf{V}|] = \log\left|\hat{\mathbf{S}}\right| - d\log(2) - \sum_{i=1}^{d}\Psi((K+N_0+1-i)/2) \quad (93)$$

$$E[(\mathbf{x}_i-\mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{m})] = (K+N_0)(\mathbf{x}_i-\hat{\mathbf{m}})^{\mathrm{T}}\hat{\mathbf{S}}^{-1}(\mathbf{x}_i-\hat{\mathbf{m}}) + d/(K_0+K) \quad (94)$$

6 Results

This section tests the accuracy of these bounds and notes some general trends. In particular, while Chickering & Heckerman (1996) found that the Cheeseman-Stutz approximation (i.e. these bounds) is comparable to Laplace's method, the results here suggest that Variational Bayes is inferior except in special cases.

6.1 Gaussian mean

In the first experiment, we have a mixture of two Gaussians with all parameters fixed except m^1:

$$p(x|m^1) = \frac{1}{2}\mathcal{N}(x; m^1, 1) + \frac{1}{2}\mathcal{N}(x; 0, 1) \quad (95)$$

$$p(m^1) \sim \mathcal{N}(0, 100) \quad (96)$$

Ten points are sampled from this model, with $m^1 = 2$. For this data, figure 1 plots the exact posterior for m^1 versus the optimal variational bound. The exact evidence, computed numerically, is $p(D) = \exp(-19.57)$ while the bound gives $\exp(-19.767)$, loose by a factor of 1.2 (meaning $\frac{\text{exact}}{\text{bound}} = 1.2$). For large N like 1000, the absolute error of the bound goes to zero. However, the relative error remains around 1.2.

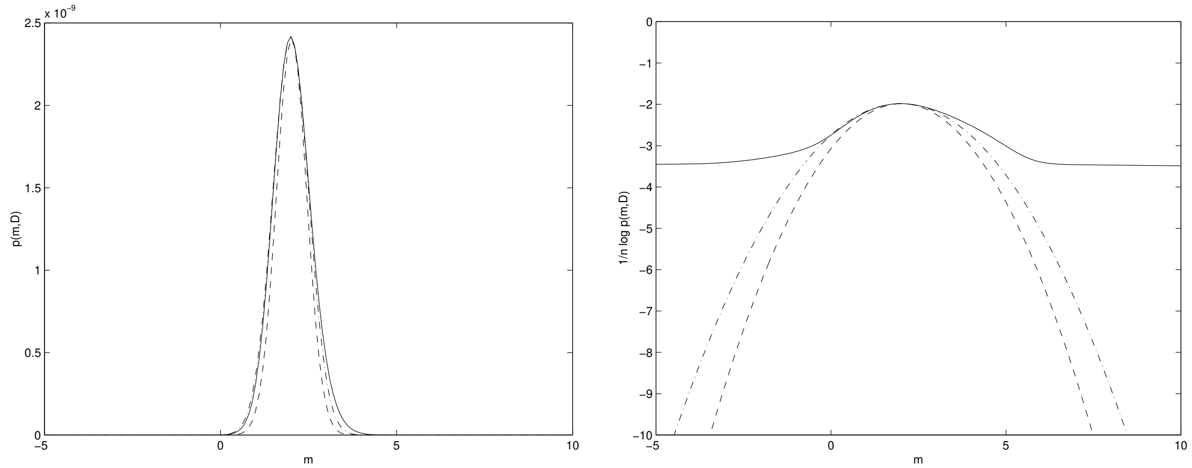

Figure 1: Exact joint $p(m^1, D)$ (solid) versus the optimal bound (dashed) and the Laplace approximation (dash-dot) for a Gaussian mean. On the right is the normalized logarithm of the joint, $\frac{1}{N}\log p(m^1, D)$, and the bound. The number of samples was $N = 10$ using $m^1 = 2$.

A useful point of comparison is Laplace's method. The relevant derivatives are (see appendix A)

$$\mathbf{g}_{i1} = x_i - m^1 \qquad \mathbf{g}_{i2} = 0 \quad (97)$$

$$\mathbf{H}_{i1} = -1 \qquad \mathbf{H}_{i2} = 0 \tag{98}$$

$$\frac{d^2 \log p(m^1)}{(dm^1)^2} = -1/100 \tag{99}$$

$$\mathbf{H} = -1/100 - \sum_i q_{i1} + \sum_i q_{i1}(1 - q_{i1})(x_i - m^1)^2 \tag{100}$$

The resulting approximation to the joint is shown in figure 1; it is not a bound. The resulting evidence approximation is $\exp(-19.592)$, which is quite good.

As mentioned earlier, the bound iteration (47) is similar to MAP iteration, especially when N is large. For this example, the soft assignments q_{ij} computed by MAP are very similar to those computed by (47). The posterior bound resulting from the MAP q's is indistinguishable from that in figure 1, and the corresponding evidence bound is $\exp(-19.773)$. This is a useful fact since algorithms for MAP EM are widely available.

There is a basic difference between the two methods, however, as illustrated in figure 2 for $N = 1$. Since there is only one data point, which is reasonably close to the second mixture component, integration EM chooses to give all weight to that component, that is $q_{i2} = 1$, leaving m^1 unconstrained by the data. This gives the bound maximum width, which in this case means maximum area. It also corresponds to a simpler model since only one component is used. The MAP bound, by contrast, always makes contact with the posterior at its mode. It achieves maximum height, but not maximum area. Both mixture components are used: the soft assignment is $q_{i2} = 0.3$. Laplace's method does the same thing, since it is equivalent to the MAP bound in this case.

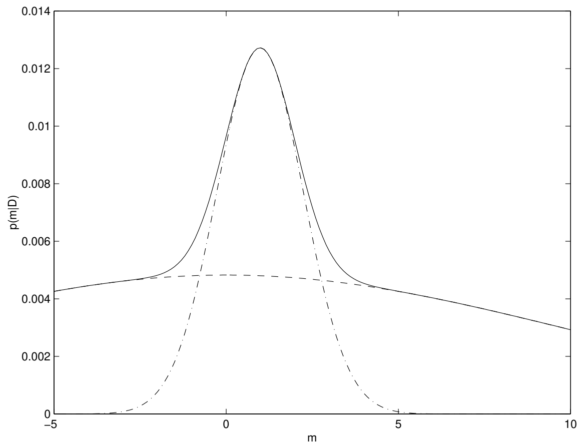

Figure 2: Exact joint $p(m^1, D)$ (solid) versus the optimal bound (dashed) and the MAP bound (dot-dash) for a Gaussian mean. The dataset contained a single sample at $x = 1$.

Another way to bound the evidence is to use the best set of hard assignments, instead of the soft assignments used by EM. This approach was taken by Vaithyanathan & Dom (1999). The hard assignments can be optimized by K-means, which is much faster than EM. However, the bound is much worse, as shown in figure 3. The evidence bound is $\exp(-21.78)$, which is loose by a factor of 9.

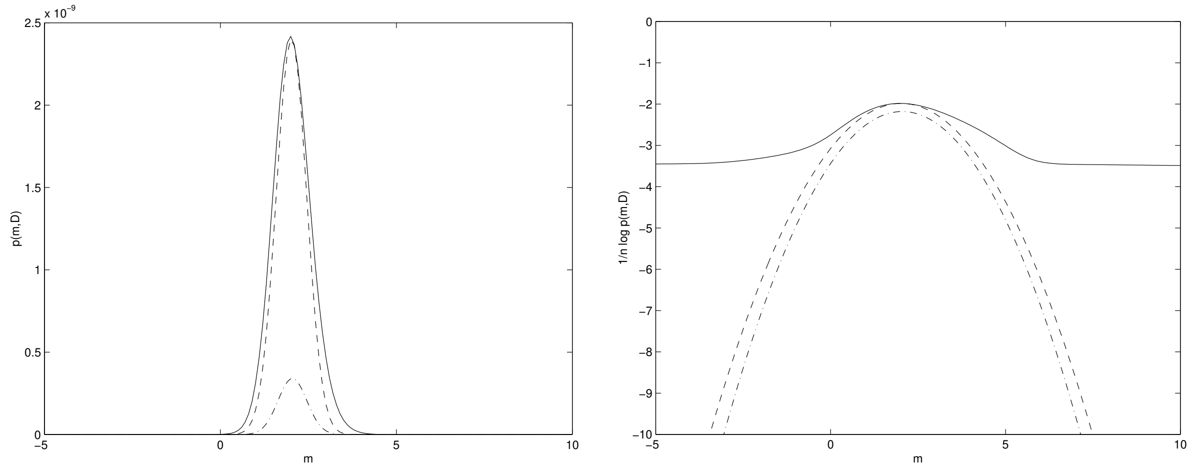

Figure 3: Same as figure 1 but including the best bound with hard assignments (dash-dot).

Now let both means be free parameters:

$$p(x|m^1, m^2) = \frac{1}{2}\mathcal{N}(x; m^1, 1) + \frac{1}{2}\mathcal{N}(x; m^2, 1) \tag{101}$$

$$p(m^1) \sim \mathcal{N}(-1, 100) \qquad p(m^2) \sim \mathcal{N}(1, 100) \tag{102}$$

One hundred points are sampled from this model, with $m^1 = -m^2$. Figure 4 plots the relative error of lower bound integration and Laplace's method, for various values of m^2. The asymmetric prior ensures a unique posterior mode, but there are still two distinct local maxima, corresponding to a swap of m^1 and m^2. Since the bound is unimodal, only one of these local maxima will be modeled. This is why both methods are consistently off by 2 when the means are well separated (m^2 is large). See figure 5.

When the means are not well separated, both methods degrade, but for different reasons. The variational bound does poorly because it doesn't capture the correlation between m^1 and m^2 induced by the data; the EM bound (22) always decouples the parameters for different components. (This is also the reason why EM is sometimes slow to converge.) Even though the posterior is elongated, the variational bound remains spherical and thus misses a lot of the area. Laplace's method doesn't have this problem because it captures correlations via the Hessian matrix. But Laplace's method suffers from near-singularity of the Hessian matrix as the two maxima get close together and merge. This is why it grossly overestimates the area. A possible fix for both methods is to change the parameterization from (m^1, m^2) to $(s = (m^1+m^2)/2, t = (m^1-m^2)/2)$, as suggested by Robert & Mengersen (1999) for sampling purposes.

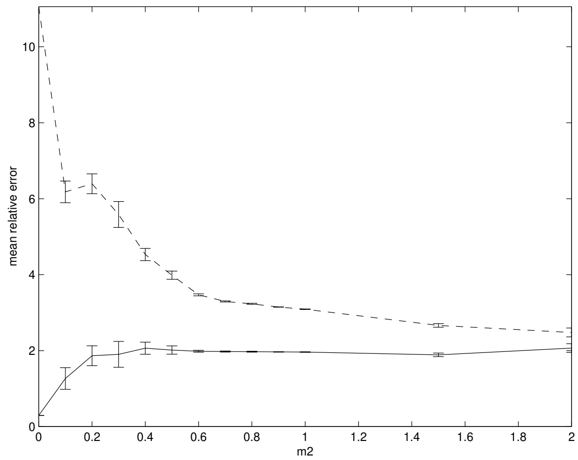

Figure 4: The relative error of the lower bound method (dashed) vs. Laplace's method (solid) for two Gaussian means. Each point represents an average over ten runs where one thousand data points were sampled from a mixture with $m^1 = -m^2$. The curve for the MAP bound is essentially identical to that for the optimal bound.

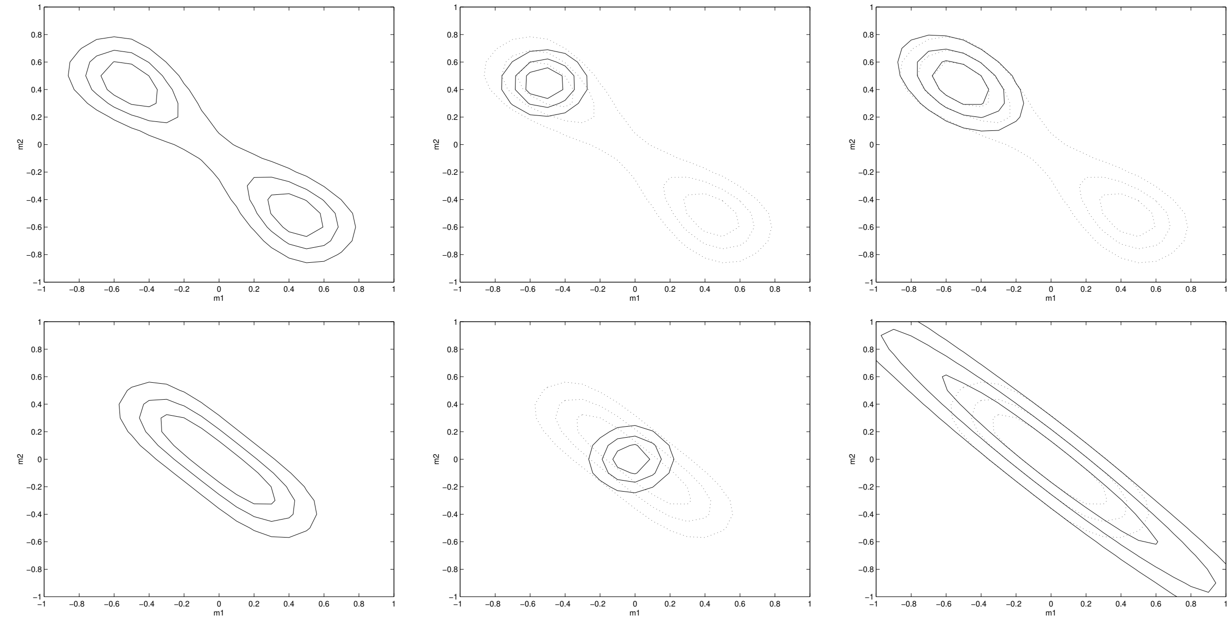

Figure 5: Contours of the exact joint $p(m^1, m^2, D)$ (left), the optimal bound (middle), and the Laplace approximation (right), with the exact joint dotted underneath. The top row uses one hundred data points from $(m^1 = -1/2, m^2 = 1/2)$ and the bottom row from $(m^1 = 0, m^2 = 0)$.

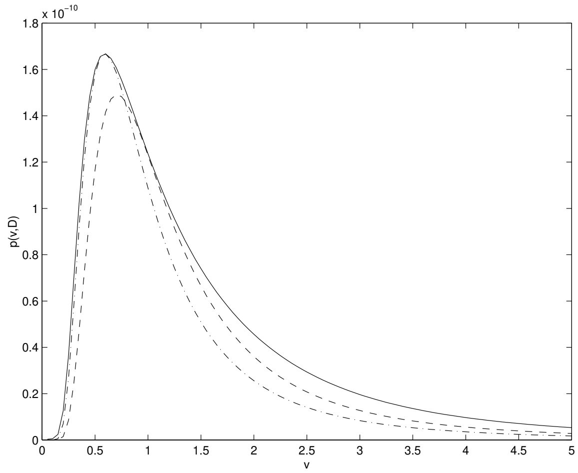

Figure 6: Exact joint $p(m^1, D)$ (solid) versus the optimal bound (dashed) and the MAP bound (dash-dot) for a Gaussian variance. The number of samples was $N = 10$ using $v^1 = 1$.

6.2 Gaussian variance

For the next experiment, let's use a mixture of two Gaussians with all parameters fixed except v^1:

$$\begin{aligned} p(x|v^1) &= \frac{1}{2}\mathcal{N}(x; 0, v^1) + \frac{1}{2}\mathcal{N}(x; 2, 1) & (103) \\ p(v^1) &\sim \mathcal{W}^{-1}(0.01, 0.01) & (104) \end{aligned}$$

Ten points are sampled from this model, with $v^1 = 1$. For this data, figure 6 plots the exact posterior for v^1 versus the optimal bound and the MAP bound. The MAP bound, which touches the peak, misses a lot of the area since the density is skewed. The exact evidence, computed numerically, is $p(D) = \exp(-22.175)$; the optimal bound gives $\exp(-22.35)$ (loose by a factor of 1.2) and the MAP bound gives $\exp(-22.42)$ (loose by a factor of 1.3). For $N = 1000$, the two bounds are the same, with a relative error of 1.3.

We can also use Laplace's method via the parameterization $v' = \log(v^1)$. The Jacobian is v^1, and the relevant derivatives wrt v' are:

$$\mathbf{g}_{i1} = -\frac{1}{2} + \frac{x_i^2}{2v^1} \qquad \mathbf{g}_{i2} = 0 \tag{105}$$

$$\mathbf{H}_{i1} = -\frac{x_i^2}{2v^1} \qquad \mathbf{H}_{i2} = 0 \tag{106}$$

$$\frac{d^2 \log p(v^1)}{(dv')^2} = -\frac{0.01}{2v^1} \tag{107}$$

$$\mathbf{H} = -\frac{0.01}{2v^1} - \sum_i q_{i1}\frac{x_i^2}{2v^1} + \sum_i q_{i1}(1 - q_{i1})\left(-\frac{1}{2} + \frac{x_i^2}{2v^1}\right)^2 \tag{108}$$

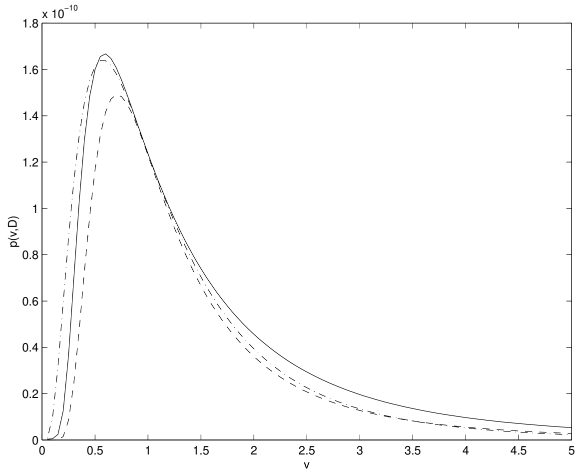

Figure 7: Exact joint $p(m^1, D)$ (solid) versus the optimal bound (dashed) and the Laplace approximation (dash-dot) for a Gaussian variance. The number of samples was $N = 10$ using $v^1 = 1$.

The location of the expansion is the MAP value of v', not v^1. It can be computed via the usual MAP iteration for v^1 but with the effective prior $p(v^1) \sim \mathcal{W}^{-1}(0.01, 0.01 - 2)$, which is (104) times the Jacobian. The lead term $p(D, \hat{\theta})$ in Laplace's method must also incorporate the Jacobian. Figure 7 plots the Laplace approximation versus the optimal bound. The resulting evidence approximation is $\exp(-22.2)$, which is very good.

6.3 Gaussian mean and variance

Now let both m^1 and v^1 be free:

$$
\begin{aligned}
p(x|m^1, v^1) &= \frac{1}{2}\mathcal{N}(x; m^1, v^1) + \frac{1}{2}\mathcal{N}(x; 2, 1) && (109)\\
p(m^1|v^1) &\sim \mathcal{N}(0, 100v^1) && (110)\\
p(v^1) &\sim \mathcal{W}^{-1}(0.01, 0.01) && (111)
\end{aligned}
$$

One hundred points are sampled from this model, with $(m^1 = 0, v^1 = 1)$. The exact joint and its approximations are shown in figure 8. For this data, the exact evidence is $p(D) = \exp(-179.35)$ and the optimal bound is $\exp(-180.03)$, loose by a factor of 2. As in the case of Gaussian means, it is loose because it doesn't capture the correlation between m^1 and v^1. The MAP bound is $\exp(-180.05)$, only slightly looser.

For Laplace's method, the relevant derivatives are (using $v' = log(v^1)$)

$$
\mathbf{g}_{i1} = \begin{bmatrix} \frac{x_i - m^1}{v^1} \\ -\frac{1}{2} + \frac{(x_i - m^1)^2}{2v^1} \end{bmatrix} \qquad \mathbf{g}_{i2} = 0 \tag{112}
$$

$$\mathbf{H}_{i1} = -\begin{bmatrix} \frac{1}{v^1} & \frac{x_i - m^1}{v^1} \\ \frac{x_i - m^1}{v^1} & \frac{(x_i - m^1)^2}{2v^1} \end{bmatrix} \qquad \mathbf{H}_{i2} = 0 \tag{113}$$

$$\frac{d^2 \log p(m^1, v^1)}{d[m^1 v']^{\mathrm{T}} d[m^1 v']} = -\begin{bmatrix} \frac{1}{100v^1} & \frac{m^1}{100v^1} \\ \frac{m^1}{100v^1} & \frac{(m^1)^2+1}{200v^1} \end{bmatrix} \tag{114}$$

The resulting evidence approximation is $\exp(-179.33)$, which is quite good.

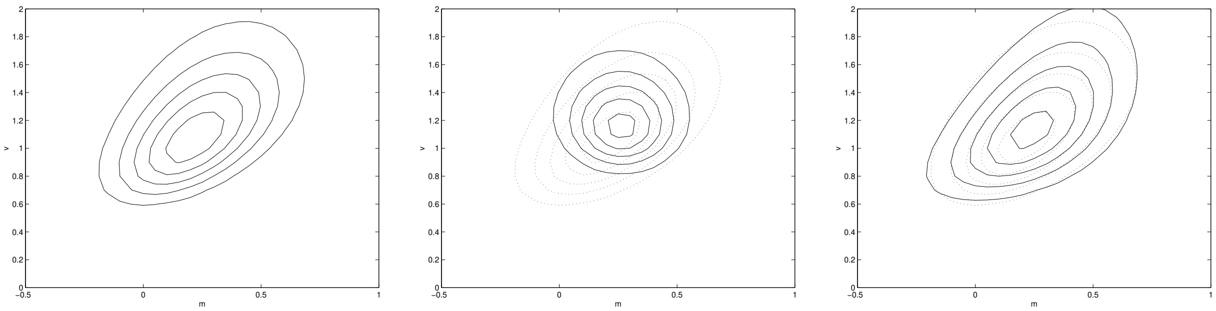

Figure 8: Contours of the exact joint $p(m^1, v^1, D)$ (left), the optimal bound (middle), and the Laplace approximation (right), with the exact joint dotted underneath. One hundred data points were sampled from $(m^1 = 0, v^1 = 1)$.

Acknowledgements

This work was supported by the MIT Media Lab Digital Life Consortium and by an internship at Just Research in summer 1999. I would like to thank Andrew McCallum for discussions on lower bound integration and for providing inspiration to develop these methods.

References

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Uncertainty in Artificial Intelligence.* `http://www.gatsby.ucl.ac.uk/~hagai/papers.html`.

Barber, D., & Bishop, C. (1997). Ensemble learning for multi-layer networks. *NIPS 10.* MIT Press. `http://www.mbfys.kun.nl/~davidb/papers/kl_mlp_nips10.html`.

Cheeseman, P., & Stutz, J. (1996). Bayesian classification (AutoClass): Theory and results. *Advances in Knowledge Discovery and Data Mining.* AAAI Press/MIT Press. `http://ic-www.arc.nasa.gov/ic/projects/bayes-group/images/kdd-95.ps`.

Chickering, D. M., & Heckerman, D. (1996). Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. *Uncertainty in Artificial Intelligence (UAI'96)* (pp. 158–168). `ftp://ftp.research.microsoft.com/pub/tr/tr-96-08.ps`.

Ghahramani, Z. (1995). Factorial learning and the EM algorithm. *NIPS* (pp. 617–624). MIT Press. `http://www.gatsby.ucl.ac.uk/~zoubin/zoubin/factorial.abstract.html`.

Jaakkola, T. S., & Jordan, M. I. (1999a). Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, *10*, 291–322. `http://www.cs.berkeley.edu/~jordan/papers/varqmr.ps.Z`.

Jaakkola, T. S., & Jordan, M. I. (1999b). Bayesian parameter estimation via variational methods. *Statistics and Computing, to appear.* `http://www.cs.berkeley.edu/~jordan/papers/variational-bayes.ps.Z`.

Jaakkola, T. S., & Jordan, M. I. (1999c). Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models.* MIT Press. `http://www.cs.berkeley.edu/~jordan/papers/mixture-mean-field.ps.Z`.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233. `http://www.cs.berkeley.edu/~jordan/papers/variational-intro.ps.Z`.

Kontkanen, P., Myllymaki, P., & Tirri, H. (1997). Experimenting with the Cheeseman-Stutz evidence approximation for predictive modeling and data mining. *Proceedings of the Tenth International FLAIRS Conference* (pp. 204–211). `http://www.cs.helsinki.fi/~tirri/flairs97.ps.gz`.

MacKay, D. J. C. (1996). Ensemble learning for Hidden Markov Models. `http://wol.ra.phy.cam.ac.uk/mackay/abstracts/ensemblePaper.html`.

Minka, T. P. (1998). Expectation-maximization as lower bound maximization. `http://vismod.www.media.mit.edu/~tpminka/papers/em.html`.

Minka, T. P. (1999). Bayesian inference, entropy, and the multinomial distribution. `http://vismod.www.media.mit.edu/~tpminka/papers/multinomial.html`.

Monti, S., & Cooper, G. F. (1999). A Bayesian network classifier that combines a finite-mixture model and a naive-Bayes model. *Uncertainty in Artificial Intelligence.* `http://www2.sis.pitt.edu/~dsl/UAI/UAI99/Monti.UAI99.html`.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (1998). Learning to classify text from labeled and unlabeled documents. *AAAI.* `http://www.cs.cmu.edu/~mccallum/papers/emcat-aaai98.ps.gz`.

Penny, W. D., & Roberts, S. J. (2000). *Variational Bayes for 1-dimensional mixture models* (Technical Report PARG-00-2). Department of Engineering Science, Oxford University. `http://www.robots.ox.ac.uk/~sjrob/Pubs/vbmog.ps.gz`.

Robert, C. P., & Mengersen, K. L. (1999). Reparameterisation issues in mixture modelling and their bearing on the Gibbs sampler. *Computational Statistics and Data Analysis*, *29*, 325–343. `ftp://ftp.ensae.fr/pub/labo_stat/CPRobert/Reparameterisation.ps.gz`.

Vaithyanathan, S., & Dom, B. (1999). Model selection in unsupervised learning with applications to document clustering. *ICML*. `http://www.almaden.ibm.com/cs/k53/ir.html`.

Waterhouse, S., MacKay, D., & Robinson, T. (1995). Bayesian methods for mixtures of experts. *NIPS*.

A Laplace's method

Laplace's method is based on the Taylor expansion

$$\log p(D,\theta) \approx \log p(D,\hat{\theta}) + \mathbf{g}^{\mathrm{T}}(\theta-\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})Tr\mathbf{H}(\theta-\hat{\theta}) + \cdots \tag{115}$$

$$\mathbf{g} = \left(\frac{d\log p(D,\theta)}{d\theta}\right)_{\theta=\hat{\theta}} \tag{116}$$

$$\mathbf{H} = \left.\frac{d^2\log p(D,\theta)}{d\theta d\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}} \tag{117}$$

If we choose $\hat{\theta}$ to maximize the joint density, or equivalently the posterior for θ, then the second term disappears and we get

$$p(D,\theta) \approx p(D,\hat{\theta})\exp(\frac{1}{2}(\theta-\hat{\theta})^{\mathrm{T}}\mathbf{H}(\theta-\hat{\theta})) \tag{118}$$

$$p(D) = \int_\theta p(D,\theta)d\theta \approx p(D,\hat{\theta})(2\pi)^{\mathrm{rows}(\mathbf{H})/2}\left|-\mathbf{H}\right|^{-1/2} \tag{119}$$

The Hessian matrix $\mathbf{H}$ is guaranteed to be negative semidefinite if $\hat{\theta}$ is a maximum. Singular $\mathbf{H}$ can usually be avoided by choosing the right parameterization. To compute $\mathbf{H}$ for a mixture density, define

$$\mathbf{g}_{ij} = \frac{d\log p(\mathbf{x}_i, c_i = j|\theta)}{d\theta} \tag{120}$$

$$\mathbf{H}_{ij} = \frac{d^2\log p(\mathbf{x}_i, c_i = j|\theta)}{d\theta d\theta^{\mathrm{T}}} \tag{121}$$

$$q_{ij} = p(c_i = j|\mathbf{x}_i,\theta) = \frac{p(\mathbf{x}_i, c_i = j|\theta)}{\sum_k p(\mathbf{x}_i, c_i = k|\theta)} \tag{122}$$

Then

$$\log p(D,\theta) = p(\theta) + \sum_i \log(\sum_j p(\mathbf{x}_i, c_i = j|\theta)) \tag{123}$$

$$\frac{d\log p(D,\theta)}{d\theta} = \frac{d\log p(\theta)}{d\theta} + \sum_{ij} q_{ij}\mathbf{g}_{ij} \tag{124}$$

$$\frac{d^2\log p(D,\theta)}{d\theta d\theta^{\mathrm{T}}} = \frac{d^2\log p(\theta)}{d\theta d\theta^{\mathrm{T}}} + \sum_{ij} q_{ij}\mathbf{H}_{ij} + \sum_{ij} q_{ij}\mathbf{g}_{ij}\mathbf{g}_{ij}^{\mathrm{T}} - \sum_i\left(\sum_j q_{ij}\mathbf{g}_{ij}\right)\left(\sum_j q_{ij}\mathbf{g}_{ij}\right)^{\mathrm{T}} \tag{125}$$

By comparison, the Hessian of the EM bound (22) is (Minka, 1998)

$$\frac{d^2\log g(\theta,q)}{d\theta d\theta^{\mathrm{T}}} = \frac{d^2\log p(\theta)}{d\theta d\theta^{\mathrm{T}}} + \sum_{ij} q_{ij}\mathbf{H}_{ij} \tag{126}$$

though in that case q is a parameter of the bound and need not correspond to (122).

B Moments of a Normal-Wishart density

The density is

$$\begin{aligned} p(\mathbf{m}, \mathbf{V}) &\sim \mathcal{NW}^{-1}(\mathbf{m}_0, K, \mathbf{S}, N) & (127) \\ &= \mathcal{N}(\mathbf{m}; \mathbf{m}_0, \mathbf{V}/K)\mathcal{W}^{-1}(\mathbf{V}; \mathbf{S}, N) & (128) \end{aligned}$$

The moments are

$$\begin{aligned} E[\mathbf{V}^{-1}] &= N\mathbf{S}^{-1} & (129) \\ E_{\mathbf{m}}[\mathbf{m}\mathbf{m}^{\mathrm{T}}] &= \mathbf{V}/K + \mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} & (130) \\ E[\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}] &= E[\mathrm{tr}(\mathbf{V}^{-1}\mathbf{m}\mathbf{m}^{\mathrm{T}})] & (131) \\ &= E[d/K + \mathbf{m}_0^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}_0] & (132) \\ &= d/K + N\mathbf{m}_0^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{m}_0 & (133) \\ E[\log|\mathbf{V}|] &= \log|\mathbf{S}/2| - \sum_{i=1}^{d}\Psi((N+1-i)/2) & (134) \end{aligned}$$

To prove (129), start with

$$\int_{\mathbf{V}\geq 0} \mathcal{W}^{-1}(\mathbf{V}; \mathbf{S}, N)d\mathbf{V} = \int_{\mathbf{V}\geq 0} \frac{1}{\Gamma_d(N/2)\,|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{S}}{2}\right|^{N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}))d\mathbf{V} = 1 \quad (135)$$

Take the derivative of both sides with respect to $\mathbf{S}$:

$$\begin{aligned} \int_{\mathbf{V}\geq 0} \left(\frac{N}{2}\mathbf{S}^{-1} - \frac{1}{2}\mathbf{V}^{-1}\right)\mathcal{W}^{-1}(\mathbf{V}; \mathbf{S}, N)d\mathbf{V} &= 0 & (136) \\ N\mathbf{S}^{-1} - E[\mathbf{V}^{-1}] &= 0 & (137) \end{aligned}$$

To prove (134), start with

$$\int_{\mathbf{V}\geq 0} \frac{1}{|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{S}}{2}\right|^{N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}))d\mathbf{V} = \Gamma_d(N/2) \quad (138)$$

Take the derivative of both sides with respect to N:

$$\begin{aligned} \frac{1}{2}\int_{\mathbf{V}\geq 0} \frac{\log|\mathbf{V}^{-1}\mathbf{S}/2|}{|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{S}}{2}\right|^{N/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{S}))d\mathbf{V} &= \frac{1}{2}\Gamma_d(N/2)\sum_{i=1}^{d}\Psi((N+1-i)/2) & (139) \\ E[\log\left|\mathbf{V}^{-1}\mathbf{S}/2\right|] &= \sum_{i=1}^{d}\Psi((N+1-i)/2) & (140) \end{aligned}$$