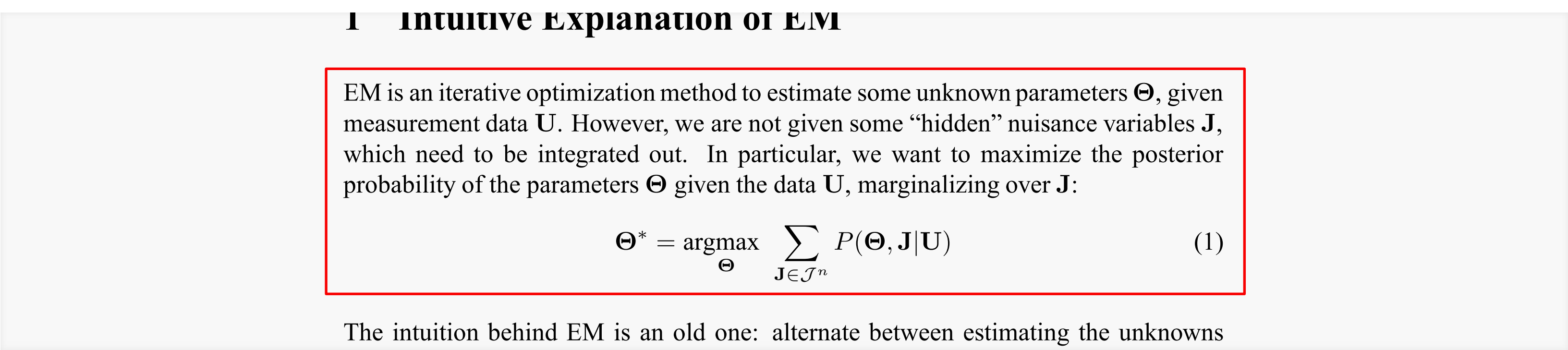
Summary

3 annotations on 2 pages by jmichaelbrunner

x3



1 Intuitive Explanation of EM

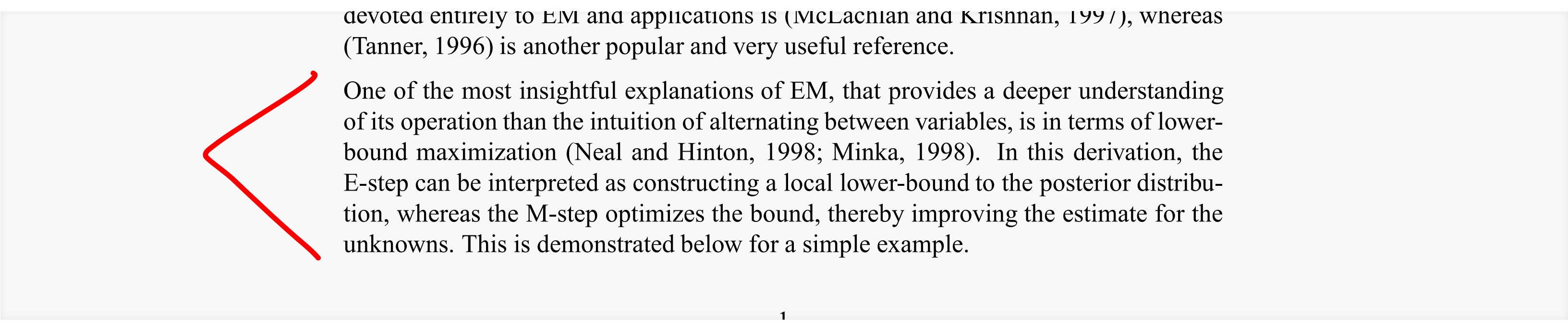
EM is an iterative optimization method to estimate some unknown parameters $\mathbf{\Theta}$, given measurement data $\mathbf{U}$. However, we are not given some "hidden" nuisance variables $\mathbf{J}$, which need to be integrated out. In particular, we want to maximize the posterior probability of the parameters $\mathbf{\Theta}$ given the data $\mathbf{U}$, marginalizing over $\mathbf{J}$:

$$\mathbf{\Theta}^* = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \sum_{\mathbf{J} \in \mathcal{J}^n} P(\mathbf{\Theta}, \mathbf{J} | \mathbf{U}) \tag{1}$$

The intuition behind EM is an old one: alternate between estimating the unknowns

#1

p.1

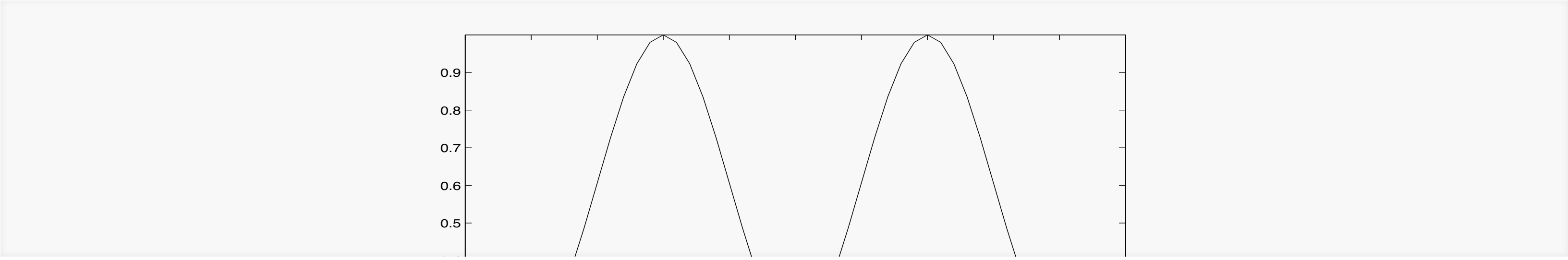


devoted entirely to EM and applications is (McLachlan and Krishnan, 1997), whereas (Tanner, 1996) is another popular and very useful reference.

One of the most insightful explanations of EM, that provides a deeper understanding of its operation than the intuition of alternating between variables, is in terms of lower-bound maximization (Neal and Hinton, 1998; Minka, 1998). In this derivation, the E-step can be interpreted as constructing a local lower-bound to the posterior distribution, whereas the M-step optimizes the bound, thereby improving the estimate for the the unknowns. This is demonstrated below for a simple example.

#2

p.1



#3

p.2

The Expectation Maximization Algorithm

Frank Dellaert

College of Computing, Georgia Institute of Technology

Technical Report number GIT-GVU-02-20
February 2002

Abstract

This note represents my attempt at explaining the EM algorithm (Hartley, 1958; Dempster et al., 1977; McLachlan and Krishnan, 1997). This is just a slight variation on Tom Minka's tutorial (Minka, 1998), perhaps a little easier (or perhaps not). It includes a graphical example to provide some intuition.

1 Intuitive Explanation of EM

EM is an iterative optimization method to estimate some unknown parameters $\mathbf{\Theta}$, given measurement data $\mathbf{U}$. However, we are not given some "hidden" nuisance variables $\mathbf{J}$, which need to be integrated out. In particular, we want to maximize the posterior probability of the parameters $\mathbf{\Theta}$ given the data $\mathbf{U}$, marginalizing over $\mathbf{J}$:

$$\mathbf{\Theta}^* = \operatorname*{argmax}_{\mathbf{\Theta}} \sum_{\mathbf{J}\in\mathcal{J}^n} P(\mathbf{\Theta},\mathbf{J}|\mathbf{U}) \tag{1}$$

The intuition behind EM is an old one: alternate between estimating the unknowns $\mathbf{\Theta}$ and the hidden variables $\mathbf{J}$. This idea has been around for a long time. However, instead of finding the best $\mathbf{J} \in \mathcal{J}$ given an estimate $\mathbf{\Theta}$ at each iteration, EM computes a *distribution* over the space $\mathcal{J}$. One of the earliest papers on EM is (Hartley, 1958), but the seminal reference that formalized EM and provided a proof of convergence is the "DLR" paper by Dempster, Laird, and Rubin (Dempster et al., 1977). A recent book devoted entirely to EM and applications is (McLachlan and Krishnan, 1997), whereas (Tanner, 1996) is another popular and very useful reference.

One of the most insightful explanations of EM, that provides a deeper understanding of its operation than the intuition of alternating between variables, is in terms of lower-bound maximization (Neal and Hinton, 1998; Minka, 1998). In this derivation, the E-step can be interpreted as constructing a local lower-bound to the posterior distribution, whereas the M-step optimizes the bound, thereby improving the estimate for the unknowns. This is demonstrated below for a simple example.

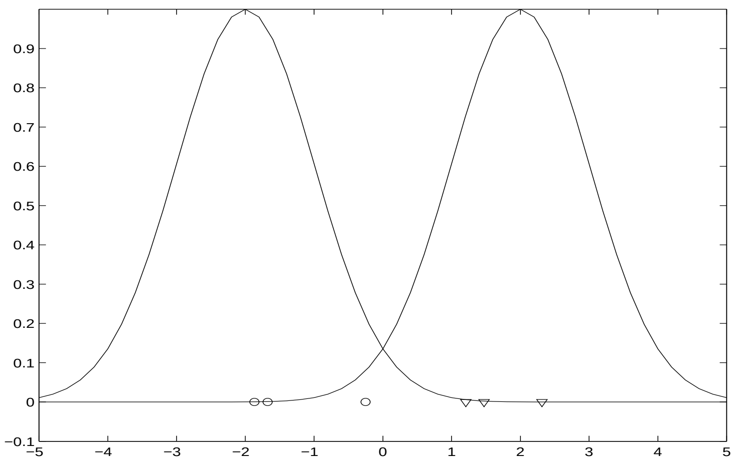


Figure 1: EM example: Mixture components and data. The data consists of three samples drawn from each mixture component, shown above as circles and triangles. The means of the mixture components are -2 and 2, respectively.

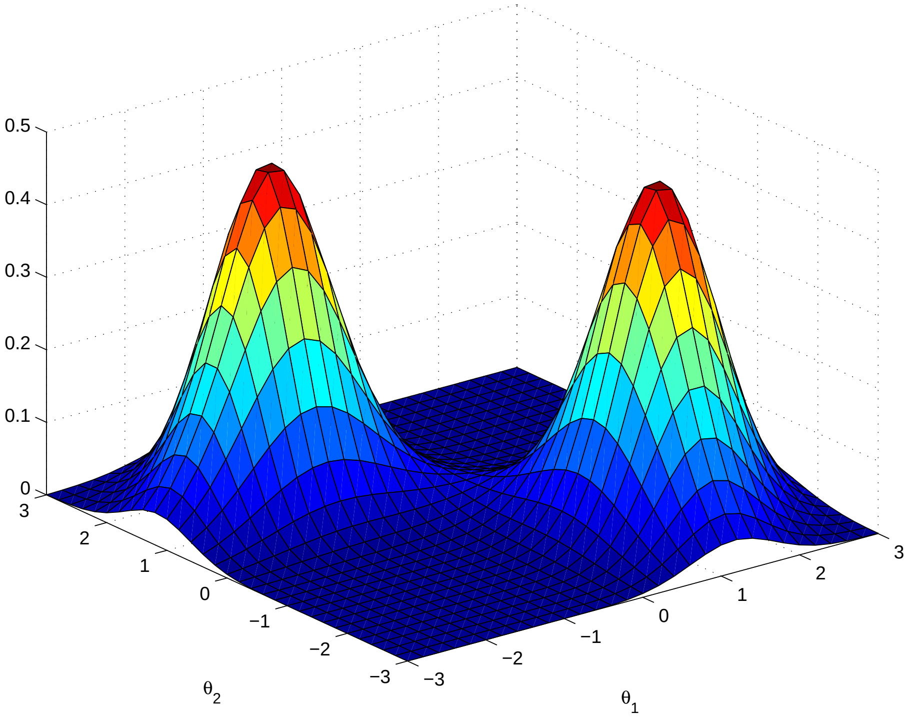


Figure 2: The true likelihood function of the two component means θ_1 and θ_2, given the data in Figure 1.

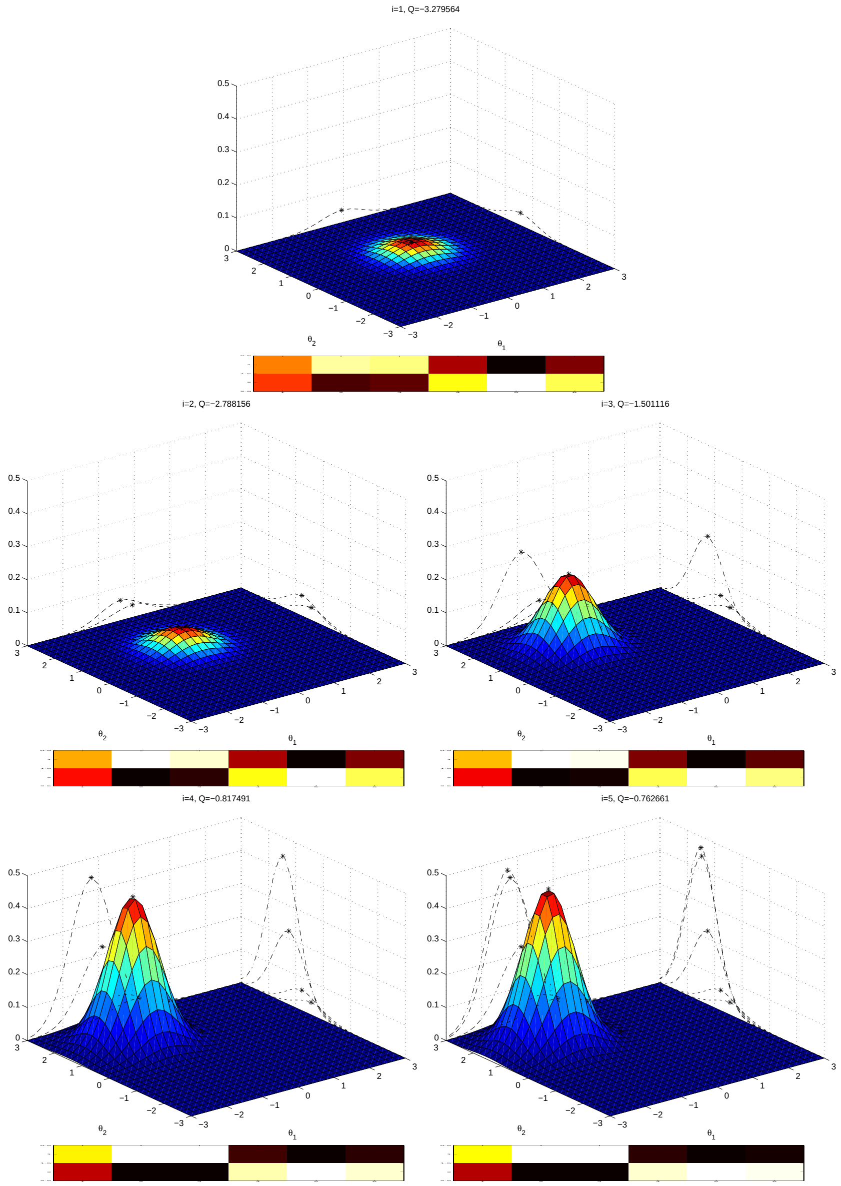


Figure 3: Lower Bounds

Consider the mixture estimation problem shown in Figure 1, where the goal is to estimate the two component means θ_1 and θ_2 given 6 samples drawn from the mixture, but *without knowing from which mixture each sample was drawn*. The state space is two-dimensional, and the true likelihood function is shown in Figure 2. Note that there are two modes, located respectively at $(-2, 2)$ and $(2, -2)$. This makes perfect sense, as we can switch the mixture components without affecting the quality of the solution. Note also that the true likelihood is computed by integrating over all possible data associations, and hence we can find a maximum likelihood solution without solving a correspondence problem. However, even for only 6 samples, this requires summing over the space of 64 possible data-associations.

EM proceeds as follows in this example. In the E-step, a "soft" assignment is computed that assigns a posterior probability to each possible association of each individual sample. In the current example, there are 2 mixtures and 6 samples, so the computed probabilities can be represented in a 2×6 table. Given these probabilities, EM computes a tight lower bound to the true likelihood function of Figure 2. The bound is constructed such that it touches the likelihood function at the current estimate, and it is only close to the true likelihood in the neighborhood of this estimate. The bound and its corresponding probability table is computed in each iteration, as shown in Figure 3. In this case, EM was run for 5 iterations. In the M-step, the lower bound is maximized (shown by a black asterisk in the figure), and the corresponding new estimate (θ_1, θ_2) is guaranteed to lie closer to the location of the nearest local maximum of the likelihood. Each next bound is an increasingly better approximation to the mode of the likelihood, until at convergence the bound touches the likelihood at the local maximum, and progress can no longer be made. This is shown in the last panel of Figure 3.

2 EM as Lower Bound Maximization

EM can be derived in many different ways, one of the most insightful being in terms of lower bound maximization (Neal and Hinton, 1998; Minka, 1998), as illustrated with the example from Section 1. In this section, we derive the EM algorithm on that basis, closely following (Minka, 1998).

The goal is to maximize the posterior probability (1) of the parameters $\mathbf{\Theta}$ given the data $\mathbf{U}$, in the presence of hidden data $\mathbf{J}$. Equivalently, we can maximize the logarithm of the joint distribution (which is proportional to the posterior):

$$\mathbf{\Theta}^* = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \; \log P(\mathbf{U}, \mathbf{\Theta}) = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \; \log \sum_{\mathbf{J} \in \mathcal{J}^n} P(\mathbf{U}, \mathbf{J}, \mathbf{\Theta}) \tag{2}$$

The idea behind EM is to start with a guess $\mathbf{\Theta}^t$ for the parameters $\mathbf{\Theta}$, compute an easily computed lower bound $B(\mathbf{\Theta}; \mathbf{\Theta}^t)$ to the function $\log P(\mathbf{\Theta}|\mathbf{U})$, and maximize that bound instead. If iterated, this procedure will converge to a local maximizer $\mathbf{\Theta}^*$ of the objective function, provided the bound improves at each iteration.

To motivate this, note that the key problem with maximizing (2) is that it involves the logarithm of a (big) sum, which is difficult to deal with. Fortunately, we can construct a tractable lower bound $B(\mathbf{\Theta};\mathbf{\Theta}^t)$ that instead contains a sum of logarithms. To derive the bound, first trivially rewrite $\log P(\mathbf{U},\mathbf{\Theta})$ as

$$\log P(\mathbf{U},\mathbf{\Theta}) = \log \sum_{\mathbf{J}\in\mathcal{J}^n} P(\mathbf{U},\mathbf{J},\mathbf{\Theta}) = \log \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J})\frac{P(\mathbf{U},\mathbf{J},\mathbf{\Theta})}{f^t(\mathbf{J})}$$

where $f^t(\mathbf{J})$ is an arbitrary probability distribution over the space $\mathcal{J}^n$ of hidden variables $\mathbf{J}$. By Jensen's inequality, we have

$$B(\mathbf{\Theta};\mathbf{\Theta}^t) \triangleq \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J}) \log \frac{P(\mathbf{U},\mathbf{J},\mathbf{\Theta})}{f^t(\mathbf{J})} <= \log \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J})\frac{P(\mathbf{U},\mathbf{J},\mathbf{\Theta})}{f^t(\mathbf{J})}$$

Note that we have transformed a log of sums into a sum of logs, which was the prime motivation.

2.1 Finding an Optimal Bound

EM goes one step further and tries to find the *best* bound, defined as the bound $B(\mathbf{\Theta};\mathbf{\Theta}^t)$ that touches the objective function $\log P(\mathbf{U},\mathbf{\Theta})$ at the current guess $\mathbf{\Theta}^t$. Intuitively, finding the best bound at each iteration will guarantee that we obtain an improved estimate $\mathbf{\Theta}^{t+1}$ when we locally maximize the bound with respect to $\mathbf{\Theta}$. Since we know $B(\mathbf{\Theta};\mathbf{\Theta}^t)$ to be a lower bound, the optimal bound at $\mathbf{\Theta}^t$ can be found by maximizing

$$B(\mathbf{\Theta}^t;\mathbf{\Theta}^t) = \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J}) \log \frac{P(\mathbf{U},\mathbf{J},\mathbf{\Theta}^t)}{f^t(\mathbf{J})} \tag{3}$$

with respect to the distribution $f^t(\mathbf{J})$. Introducing a Lagrange multiplier λ to enforce the constraint $\sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J}) = 1$, the objective becomes

$$G(f^t) = \lambda\left[1 - \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J})\right] + \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J}) \log P(\mathbf{U},\mathbf{J},\mathbf{\Theta}^t) - \sum_{\mathbf{J}\in\mathcal{J}^n} f^t(\mathbf{J}) \log f^t(\mathbf{J})$$

Taking the derivative

$$\frac{\partial G}{\partial f^t(\mathbf{J})} = -\lambda + \log P(\mathbf{U},\mathbf{J},\mathbf{\Theta}^t) - \log f^t(\mathbf{J}) - 1$$

and solving for $f^t(\mathbf{J})$ we obtain

$$f^t(\mathbf{J}) = \frac{P(\mathbf{U},\mathbf{J},\mathbf{\Theta}^t)}{\sum_{\mathbf{J}\in\mathcal{J}^n} P(\mathbf{U},\mathbf{J},\mathbf{\Theta}^t)} = P(\mathbf{J}|\mathbf{U},\mathbf{\Theta}^t)$$

By examining the value of the resulting optimal bound at $\mathbf{\Theta}^t$ we see that it indeed touches the objective function:

$$B(\mathbf{\Theta}^t;\mathbf{\Theta}^t) = \sum_{\mathbf{J}\in\mathcal{J}^n} P(\mathbf{J}|\mathbf{U},\mathbf{\Theta}^t) \log \frac{P(\mathbf{U},\mathbf{J},\mathbf{\Theta}^t)}{P(\mathbf{J}|\mathbf{U},\mathbf{\Theta}^t)} = \log P(\mathbf{U},\mathbf{\Theta}^t)$$

2.2 Maximizing The Bound

In order to maximize $B(\mathbf{\Theta};\mathbf{\Theta}^t)$ with respect to $\mathbf{\Theta}$, note that we can write it as

$$\begin{aligned} B(\mathbf{\Theta};\mathbf{\Theta}^t) &\triangleq \langle \log P(\mathbf{U},\mathbf{J},\mathbf{\Theta})\rangle + \mathcal{H} \\ &= \langle \log P(\mathbf{U},\mathbf{J}|\mathbf{\Theta})\rangle + \log P(\mathbf{\Theta}) + \mathcal{H} \\ &= Q^t(\mathbf{\Theta}) + \log P(\mathbf{\Theta}) + \mathcal{H} \end{aligned}$$

where $\langle . \rangle$ denotes the expectation with respect to $f^t(\mathbf{J}) \triangleq P(\mathbf{J}|\mathbf{U},\mathbf{\Theta}^t)$, and

- $Q^t(\mathbf{\Theta})$ is the expected complete log-likelihood, defined as:

$$Q^t(\mathbf{\Theta}) \triangleq \langle \log P(\mathbf{U},\mathbf{J}|\mathbf{\Theta})\rangle$$

- $P(\mathbf{\Theta})$ is the prior on the parameters $\mathbf{\Theta}$
- $\mathcal{H} \triangleq -\langle \log f^t(\mathbf{J})\rangle$ is the entropy of the distribution $f^t(\mathbf{J})$

Since $\mathcal{H}$ does not depend on $\mathbf{\Theta}$, we can maximize the bound with respect to $\mathbf{\Theta}$ using the first two terms only:

$$\mathbf{\Theta}^{t+1} = \underset{\mathbf{\Theta}}{\text{argmax}}\;\; B(\mathbf{\Theta};\mathbf{\Theta}^t) = \underset{\mathbf{\Theta}}{\text{argmax}}\;\; \left[Q^t(\mathbf{\Theta}) + \log P(\mathbf{\Theta})\right] \tag{4}$$

2.3 The EM Algorithm

At each iteration, the EM algorithm first finds an optimal lower bound $B(\mathbf{\Theta};\mathbf{\Theta}^t)$ at the current guess $\mathbf{\Theta}^t$ (equation 3), and then maximizes this bound to obtain an improved estimate $\mathbf{\Theta}^{t+1}$ (equation 4). Because the bound is expressed as an expectation, the first step is called the "expectation-step" or E-step, whereas the second step is called the "maximization-step" or M-step. The EM algorithm can thus be conveniently summarized as:

- E-step: calculate $f^t(\mathbf{J}) \triangleq P(\mathbf{J}|\mathbf{U},\mathbf{\Theta}^t)$
- M-step: $\mathbf{\Theta}^{t+1} = \text{argmax}_{\;\mathbf{\Theta}} \left[Q^t(\mathbf{\Theta}) + \log P(\mathbf{\Theta})\right]$

It is important to remember that $Q^t(\mathbf{\Theta})$ is calculated in the E-step by evaluating $f^t(\mathbf{J})$ using the *current guess* $\mathbf{\Theta}^t$ (hence the superscript t), whereas in the M-step we are optimizing $Q^t(\mathbf{\Theta})$ with respect to the *free variable* $\mathbf{\Theta}$ to obtain the new estimate $\mathbf{\Theta}^{t+1}$. It can be proved that the EM algorithm converges to a local maximum of $\log P(\mathbf{U},\mathbf{\Theta})$, and thus equivalently maximizes the log-posterior $\log P(\mathbf{\Theta}|\mathbf{U})$ (Dempster et al., 1977; McLachlan and Krishnan, 1997).

A Relation to the Expected Log-Posterior

Note that we have chosen to define $Q^t(\mathbf{\Theta})$ as the expected log-likelihood as in (Dempster et al., 1977; McLachlan and Krishnan, 1997), i.e.,

$$Q^t(\mathbf{\Theta}) \triangleq \langle \log P(\mathbf{U}, \mathbf{J}|\mathbf{\Theta})\rangle$$

An alternative route is to compute the expected log-posterior (Tanner, 1996):

$$\langle \log P(\mathbf{\Theta}|\mathbf{U}, \mathbf{J})\rangle = \langle \log P(\mathbf{U}, \mathbf{J}|\mathbf{\Theta}) + \log P(\mathbf{\Theta}) - \log P(\mathbf{U}, \mathbf{J})\rangle \tag{5}$$

Here the second term does not depend on $\mathbf{J}$ and can be taken out of the expectation, and the last term does not depend on $\mathbf{\Theta}$. Hence, maximizing (5) with respect to $\mathbf{\Theta}$ is equivalent to (4):

$$\begin{aligned}\underset{\mathbf{\Theta}}{\text{argmax}} \ \langle \log P(\mathbf{\Theta}|\mathbf{U}, \mathbf{J})\rangle &= \underset{\mathbf{\Theta}}{\text{argmax}} \ [\langle \log P(\mathbf{U}, \mathbf{J}|\mathbf{\Theta})\rangle + \log P(\mathbf{\Theta})] \\ &= \underset{\mathbf{\Theta}}{\text{argmax}} \ \left[Q^t(\mathbf{\Theta}) + \log P(\mathbf{\Theta})\right]\end{aligned}$$

This is of course identical to (4).

References

[1] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[2] Hartley, H. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194.

[3] McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley series in probability and statistics. John Wiley & Sons.

[4] Minka, T. (1998). Expectation-Maximization as lower bound maximization. Tutorial published on the web at http://www-white.media.mit.edu/ tpminka/papers/em.html.

[5] Neal, R. and Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M., editor, *Learning in Graphical Models*. Kluwer Academic Press.

[6] Tanner, M. (1996). *Tools for Statistical Inference*. Springer Verlag, New York. Third Edition.