CS 281 Machine Learning
Spring 1998 Stuart Russell

The EM Algorithm

The EM algorithm (Dempster et al., 1977) is one of the most widely used algorithms in statistics. Every year, 200–300 research papers are published in which EM is the topic or the main tool. Applications range from finding new types of stars to separating out different types of tissue in X-ray images to identifying categories of consumers from their buying behaviour. Neural networks and belief networks can be trained using EM as well as more "traditional" gradient-based methods. McLachlan and Krishnan (1997) devote an entire book to EM.

Given some data $\mathbf{X}$ and a model family parameterized by θ, the goal of EM in its basic form is to find θ such that the likelihood $P(\mathbf{X}|\theta)$ is maximized. In general, EM can find only a local maximum. Each cycle revises the value of θ so as to increase the likelihood until a maximum is reached. The purpose of this document is to derive the algorithm in its most general form from first principles and to give a short proof of its convergence. The derivation extends the mixture-model derivation from Bishop (1995, pp. 65–66) and leads to the algorithm given in Mitchell (1997, p.195).

Suppose we define the log likelihood function $L(\theta) = \ln P(\mathbf{X}|\theta)$ and suppose that our current estimate for the optimal parameters is θ_i. We will examine what happens to L when a new value θ is computed by the algorithm:

$$L(\theta) - L(\theta_i) = \ln P(\mathbf{X}|\theta) - \ln P(\mathbf{X}|\theta_i) = \ln \frac{P(\mathbf{X}|\theta)}{P(\mathbf{X}|\theta_i)}$$

Depending on what we choose for θ, the value of L could go up or down. We would like to choose θ to maximize the right-hand side of the equation above. In general, this cannot be done; the core idea of EM is to introduce some *unobserved* variables $\mathbf{Z}$, appropriate for the model family under consideration, such that *if* $\mathbf{Z}$ *were known* the optimal value θ could be computed easily. Mathematically, $\mathbf{Z}$ is brought into the equations by conditioning:

$$L(\theta) - L(\theta_i) = \ln \frac{\sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{X}|\theta_i)}$$

Unfortunately, this expression is the logarithm of a sum, which is hard to deal with. Jensen's inequality allows one to replace this with a sum of logarithms:

$$\ln \sum_j \lambda_j y_j \geq \sum_j \lambda_j \ln y_j \qquad \text{if } \sum_j \lambda_j = 1$$

To use this in our problem, we need some coefficients $\lambda_{\mathbf{z}}$ such that $\sum_{\mathbf{z}} \lambda_{\mathbf{z}} = 1$. Given that the E-step of EM computes the expected value of the hidden variables given the current data and parameters, it seems reasonable to introduce coefficients $P(\mathbf{z}|\mathbf{X},\theta_i)$ as follows:

$$L(\theta) - L(\theta_i) = \ln \frac{\sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{X}|\theta_i)} \frac{P(\mathbf{z}|\mathbf{X},\theta_i)}{P(\mathbf{z}|\mathbf{X},\theta_i)}$$

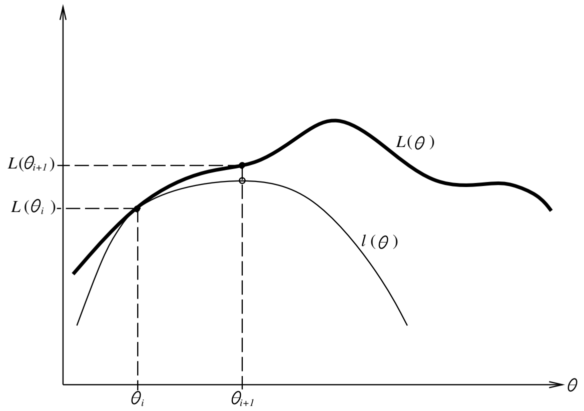
Now we can apply Jensen's inequality to obtain

$$L(\theta) - L(\theta_i) \geq \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_i) \ln \frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{X}|\theta_i)P(\mathbf{z}|\mathbf{X},\theta_i)}$$

which can be rewritten as $L(\theta) \geq L(\theta_i) + \Delta(\theta|\theta_i) = l(\theta|\theta_i)$ where

$$\Delta(\theta|\theta_i) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_i) \ln \frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{X}|\theta_i)P(\mathbf{z}|\mathbf{X},\theta_i)}$$

Now $L(\theta)$ and $l(\theta|\theta_i)$ are both functions of θ, and $L(\theta)$ is everywhere greater than or equal to $l(\theta|\theta_i)$. Furthermore, it is straightforward to show that $\Delta(\theta_i|\theta_i) = 0$, so the two functions must be exactly equal at $\theta = \theta_i$. Thus, we have the following picture:



Diagrammatic representation of how EM works. The X–axis ranges over the possible values of the parameters θ; the Y-axis gives the likelihood of the data. EM computes the function $l(\theta)$ using the current estimate θ_i and computes the new estimate θ_{i+1} as the maximum point of $l(\theta)$.

Now the idea is to find the value of θ that maximizes $L(\theta_i) + \Delta(\theta|\theta_i)$. (Because of the inclusion of the hidden variables, we hope this will be easier than the original maximization problem for L.)

$$\begin{aligned}
\theta_{i+1} &= \arg\max_\theta L(\theta_i) + \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_i) \ln \frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{X}|\theta_i)P(\mathbf{z}|\mathbf{X},\theta_i)} \\
&= \arg\max_\theta \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_i) \ln[P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)] = \arg\max_\theta E_{\mathbf{z}|\mathbf{X},\theta_i}[\ln P(\mathbf{X},\mathbf{z}|\theta)]
\end{aligned}$$

This equation is the EM algorithm in a nutshell. The E-step is the computation of the expectation and the M-step is the maximization. Notice that quite a bit of work may be required to apply this to any particular model family, as illustrated by the rederivation of the mixture model algorithm given by Mitchell.

Convergence is shown as follows:

- θ_{i+1} maximizes Δ, therefore $\Delta(\theta_{i+1}|\theta_i) \geq \Delta(\theta_i|\theta_i) = 0$. Therefore, at each iteration, $L(\theta)$ cannot decrease.
- When EM reaches a fixed point at some θ_i, we know that θ_i is a maximum of $l(\theta)$; furthermore, L and l are equal at θ_i; hence, provided L and l are differentiable, θ_i must also be a *stationary* point of L—not necessarily a local maximum. In fact, McLachlan and Krishnan (1997) give actual examples showing convergence to saddle points and local minima. Such behaviour is rare in practice provided differentiability holds; if not, as in the case of fitting a Gaussian with zero variance to a single data point, there are no guarantees at all.

Notice that the convergence result simply relies on the assertion that $\Delta(\theta_{i+1}|\theta_i) \geq \Delta(\theta_i|\theta_i)$. Although this is clearly satisfied by choosing θ_{i+1} to maximize Δ, it is also satisfied by any process that can improve the value of Δ from its current value—for example, a gradient method using $\partial\Delta/\partial\theta$. Thus, even in cases where the M-step is intractable and the graident of L cannot be computed easily, it is still possible to guarantee progress using this generalized verison of EM—the so-called GEM algorithm.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford.

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

McLachlan, G. J., & Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley, New York.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill, New York.