C EM

C.1 Introduction

Everyone seems to have a different understanding of the EM algorithm. This is mine.

EM is a numerical algorithm for maximising probabilities. Say you want to maximise the probability of some parameters, θ. Normally you differentiate and equate to zero to solve for θ. When this is in some sense difficult, you can begin with some estimate of θ and aim to find a better estimate, θ'. EM is concerned with the difficulty being a logarithm of a sum; it basically swaps them around into a sum of logarithms.

EM stands for Expectation Maximisation; the term comes from Dempster et al. (1977). It's misleading though.

C.2 Baum's derivation

The first paper that is really important is by Baum et al. (1970). To get both θ and θ' in the same equation, they begin with a ratio, then proceed to apply Jensen's inequality:

$$\log \frac{p(\theta')}{p(\theta)} = \log \int_{\phi} d\phi \, \frac{p(\phi, \theta')}{p(\theta)} \tag{1}$$

$$= \log \int_{\phi} d\phi \, \frac{p(\phi, \theta)}{p(\theta)} \frac{p(\phi, \theta')}{p(\phi, \theta)} \tag{2}$$

$$\geqslant \int_{\phi} d\phi \, \frac{p(\phi, \theta)}{p(\theta)} \log \frac{p(\phi, \theta')}{p(\phi, \theta)} \tag{3}$$

$$= \frac{1}{p(\theta)} \left[Q(\theta, \theta') - Q(\theta, \theta) \right] \geqslant 0, \tag{4}$$

which implies that

$$Q(\theta, \theta') = \int_{\phi} d\phi \, p(\phi, \theta) \log p(\phi, \theta'). \tag{5}$$

The point here is that if you maximise $Q(\theta, \theta')$ such that it's greater than $Q(\theta, \theta)$, that implies that $p(\theta') > p(\theta)$, so iterating will move in the right direction. $Q(\theta, \theta')$ is known as Baum's auxiliary function.

More often than not, that integral is actually a summation over either mixture components or state sequences. The thing you need to differentiate is inside the logarithm; it can be quite complicated, but the logarithm tends to separate out terms.

C.3 Lower bounds

There is a nicer derivation that goes like this: Introduce some data, x, which we know were generated by a parametric model with parameters θ. First formulate it as a maximum a-posteriori (MAP) problem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \, p(\theta \mid x) = \underset{\theta}{\operatorname{argmax}} \, \frac{p(x, \theta)}{p(x)}, \tag{6}$$

and as the denominator is independent of the maximisation we proceed using the joint distribution.

Now, say the data is dependent upon something else; something that we don't know and don't necessarily care about. Call it ϕ. The correct approach is to integrate it out, or marginalise:

$$p(x, \theta) = \int d\phi \, p(x, \phi, \theta). \tag{7}$$

where that integral can be a summation if ϕ is discrete.

We need both θ and θ' in the same equation. Using Baum's derivation as a clue, introduce the other one like this:

$$\begin{aligned} p(x, \theta') &= \int d\phi\, p(x, \phi, \theta') && (8)\\ &= \int d\phi\, p(x, \phi, \theta') \cdot \frac{p(\phi \mid x, \theta)}{p(\phi \mid x, \theta)} && (9)\\ &= \int d\phi\, p(\phi \mid x, \theta)\, \frac{p(x, \phi, \theta')}{p(\phi \mid x, \theta)}. && (10) \end{aligned}$$

The logarithm (or any concave function) of equation 10 has the form of Jensen's inequality. Applying it,

$$\begin{aligned} \log p(x, \theta') &= \log \left(\int d\phi\, p(\phi \mid x, \theta)\, \frac{p(x, \phi, \theta')}{p(\phi \mid x, \theta)} \right) && (11)\\ &\geqslant \int d\phi\, p(\phi \mid x, \theta) \log \frac{p(x, \phi, \theta')}{p(\phi \mid x, \theta)}. && (12) \end{aligned}$$

That's basically it, but there are a few spiffy things to point out about equation 12:

1. If we write

$$\begin{aligned} \log p(x, \theta') &\geqslant \int d\phi\, p(\phi \mid x, \theta) \log \frac{p(x, \phi, \theta')}{p(\phi \mid x, \theta)} && (13)\\ &= \int d\phi\, p(\phi \mid x, \theta) \log \frac{p(\phi \mid x, \theta')\, p(x, \theta')}{p(\phi \mid x, \theta)} && (14)\\ &= \log p(x, \theta') - \int d\phi\, p(\phi \mid x, \theta) \log \frac{p(\phi \mid x, \theta)}{p(\phi \mid x, \theta')}. && (15) \end{aligned}$$

   i.e., there is a relationship with Kullback-Leibler "distance".

2. If we set $\theta' = \theta$, equation 15 becomes

$$\log p(x, \theta') - \int d\phi\, p(\phi \mid x, \theta) \log \frac{p(\phi \mid x, \theta)}{p(\phi \mid x, \theta')} = \log p(x, \theta)\,. \tag{16}$$

   So, we have a function that is strictly smaller than another that we want to maximise, except at one point, where they are the same. This is pretty much sufficient to prove that iteratively maximising the smaller function will converge to the local maximum of the larger. That's the lower bound thing.

3. Notice that

$$\operatorname*{argmax}_{\theta'} \int d\phi\, p(\phi \mid x, \theta) \log \left(\frac{p(x, \phi, \theta')}{p(\phi \mid x, \theta)} \right) = \operatorname*{argmax}_{\theta'} \int d\phi\, p(\phi \mid x, \theta) \log \left(p(x, \phi, \theta') \right), \tag{17}$$

   so we get back to something like Baum's auxiliary function, albeit with the data included too.

4. You can write

$$\int d\phi\, p(\phi \mid x, \theta) \log \left(p(x, \phi, \theta') \right) = \mathbb{E} \left(\log p(x, \phi, \theta') \mid x, \theta \right), \tag{18}$$

   where the notation as a conditional expectation leads to the EM (Expectation Maximisation) terminology. Thinking of it as an expectation just clouds the situation though.

5. For the ML solution, just drop the prior on θ'

$$\int d\phi\, p(\phi \mid x, \theta) \log \left(p(x, \phi \mid \theta') \right) = \mathbb{E} \left(\log p(x, \phi \mid \theta') \mid x, \theta \right), \tag{19}$$

The thing about lower bounds is that none of this is actually tied to either logarithms or even to Jensen's inequality. Jensen is true for any concave function; any inequality that simplifies the situation will do.