Proof that EM increases data likelihood

November 26, 2018

1 Introduction

Suppose a data item is specified by values of variables \( z, x \), and you have a model of their probability, determined by some parameters \( \Theta - P(z, x; \Theta) \). EM is a (re-)estimation procedure for finding a setting of \( \Theta \) when you have just incomplete data \( x^1 \ldots x^D \) to work with, the values on \( z \) being hidden. EM takes some starting estimate \( \theta^1 \) and iteratively derive new estimates with the vital property that the likelihood of the data never decreases. As this is the same as increasing the log-likelihood of the data, abbreviating this as \( l(x^1 \ldots x^D; \Theta) \), this amounts to the property that

\[
l(x^1 \ldots x^D; \theta^m) \leq l(x^1 \ldots x^D, \theta^{m+1})
\]

There is an elegant proof of (1), a version of which is given in the following pages.

2 The Q function formulation of EM

Let \( \theta^m \) be some setting of the parameters of the model. Under these parameters, there is the distribution \( P(z|x^d; \theta^m) \) of the values of \( z \) given what is known. For any quantity \( \Phi^d(z) \), involving \( z \), you could formulate what is technically the expectation of \( \Phi^d(z) \), under the distribution \( P(z|x^d; \theta^m) \).

\[
E_{z|x^d;\theta^m}[\Phi^d(z)] := \sum_z \left[ p(z|x^d;\theta^m)\Phi^d(z) \right]
\]

Setting \( \Phi^d(z) \) to be \( \log(p(z, x^d; \Theta)) \), where \( \Theta \) is any arbitrary setting, you can form the following 'expectation', which is conventionally referred to as \( Q^d(\Theta, \theta^m) \)

\[
Q^d(\Theta, \theta^m) = \mathbb{E}_{z|x^d;\theta^m}[\log(p(z, x^d; \Theta))]
\]

If you put in some particular value for \( \Theta \), the \( \log(p(z, x^d; \Theta)) \) gives the log probability of a complete data point, completed with a particular value for \( z \). The expectation is creating the weighted sum of these, weighting each by \( P(z|x^d; \theta^m) \), the conditional probability of \( z \) under parameter setting \( \theta^m \). Additionally, let \( Q(\Theta, \theta^m) \) be the sum of this over all \( d \), i.e. \( Q(\Theta, \theta^m) := \sum_d Q^d(\Theta, \theta^m) \). So for a given
$\theta^m$, $Q(\Theta, \theta^m)$ is a function, giving for each $\Theta$ a particular expectation of a log probability: see red line in Fig 1.

What’s this got to do with the virtual counts considered before? The ‘count’ due to item $d$ of outcome $z$, $\gamma_d(z)$, is $P(z|x^d; \theta^m)$. $\sum_z \gamma_d(z) \log(p(z, x^d; \Theta))$ is the log-likelihood of the virtual corpus of completions of item $x^d$, and this is exactly what $Q^d(\Theta, \theta^m)$ defines. Furthermore it is not hard to show that the log-likelihood of the virtual corpus of completions of all items $x^d$, is

$$\sum_d \left[ \sum_z \gamma_d(z) \log(p(z, x^d; \Theta)) \right] = \sum_d Q^d(\Theta, \theta^m)$$

Thus maximising $Q(\Theta, \theta^m)$ (wrt. $\Theta$) is the same as maximising the log-likelihood of the virtual corpus of completions of all items $x^d$ – which is how the M step is thought of in the virtual counts presentation.

The most general formulation of EM is in terms of this $Q$-function and it is:

**Q-function formulation of E-M step** Suppose $\theta^m$ is current parameter estimate. Derive new estimate $\theta^{m+1}$ by following E then M step

**E step** : given $\theta^m$, ‘determine’ the $Q$ function $Q(\Theta, \theta^m)$, which is $\sum_d \left[ \mathbb{E}_{z|x^d, \theta^m} \left\{ \log(p(z, x^d; \Theta)) \right\} \right]$. Practically this means calculating $P(z|x^d; \theta^m)$ for each $z$

**M step** : set $\theta^{m+1}$ to maximiser of $Q$:

$$\theta^{m+1} = \arg \max_{\Theta} Q(\Theta, \theta^m) = \arg \max_{\Theta} \sum_d Q^d(\Theta, \theta^m) = \arg \max_{\Theta} \sum_d \left[ \mathbb{E}_{z|x^d, \theta^m} \left\{ \log(p(z, x^d; \Theta)) \right\} \right]$$

In the cases that were looked at in the course the solution to this M step is always accomplished by taking various ratios of expected counts, which might take a bit of believing, but can be rigorously proved.

Thus far not much seems to have been accomplished: a familiar formulation of EM in terms of repeated parameter estimation from a virtual corpus of completions has been re-described as maximising a summed expectation of a particular log, in particular the expectation of the log of $p(z, x^d; \Theta)$ taken wrt to $P(z|x^d; \theta^m)$, the conditional probability of the completions given what is visible and the current parameter setting $\theta^m$. However, expressing things this way turns out to be crucial to carrying out the proof of increasing data likelihood, which will invoke Jensen’s inequality involving expectations of logs$^1$:

$$\log\left( \mathbb{E}_{p(X)} [\Phi(X)] \right) \geq \mathbb{E}_{p(X)} [\log(\Phi(X))]$$

$^1$log(x), where $x$ is a probability, is a function whose second derivative is always negative, and Jensen’s inequality applies to any such function $f$, $f(\mathbb{E}_{p(X)} [x]) \geq \mathbb{E}_{p(X)} [f(x)]$
or spelling out the expectation:

\[
\log \left( \sum_{x \in \text{val}(X)} [p(x)\Phi(x)] \right) \geq \sum_{x \in \text{val}(X)} [p(x)\log(\Phi(x))] \]

The quantity that you really want to maximise is the log likelihood of the data, \( \log(p(x^1 \ldots x^D; \Theta)) \). Abbreviating this to \( l(x^1 \ldots x^D; \Theta) \), this is

\[
l(x^1 \ldots x^D; \Theta) = \sum_d \left[ \log(\sum_z p(z, x^d; \Theta)) \right] \tag{2}\]

The quantity that is getting maximised in each M step is

\[
\sum_d Q_d(\Theta, \theta^m) = \sum_d \left[ \mathbb{E}_{z|x^d, \theta^m} [\log(p(z, x^d; \Theta))] \right] \tag{3}\]

Something is going to have to happen to relate (2) to (3), and its going to have to do with switching the relative order of log and sum in the two expressions.

## 3 Relating Q to the Log Likelihood

It turns out you can find a lower bound for \( l(x^1 \ldots x^D; \Theta) \) in terms of \( Q(\Theta, \theta^m) \)

\[
l(x^1 \ldots x^D; \Theta) \geq Q(\Theta, \theta^m) + h(\theta^m) \tag{4}\]

It also turns out that this lower bound is actually equal to the likelihood when \( \Theta = \theta^m \):

\[
l(x^1 \ldots x^D; \theta^m) = Q(\theta^m, \theta^m) + h(\theta^m) \tag{5}\]

It follows that if \( \Theta \) is such that \( Q(\Theta, \theta^m) \geq Q(\theta^m, \theta^m) \) then \( l(x^1 \ldots x^D; \Theta) \geq l(x^1 \ldots x^D; \theta^m) \). This is sufficient for showing that EM increases probability – that is (1) – because at the M step \( \theta^{m+1} \) is a maximiser of \( Q(\Theta, \theta^m) \)

(4) and (5) are both proven below. But before that its probably useful to look in slow motion at how they fit together to indeed show monotone increase of data likelihood. Consider the picture in Fig 1. This shows \( \Theta \) on the horizontal axis, pretending that it is just 1-dimensional. The bold line is the log likelihood of the data under various settings of \( \Theta \) ie. \( l(x^1 \ldots x^D; \Theta) \). Suppose some EM iteration starts with the parameters at \( \theta^m \). The red line (dashes) shows \( Q(\Theta, \theta^m) \), which can be of pretty arbitrary shape, but (4) and (5) force two things. The first is that when \( h(\theta^m) \) is added to the \( Q \) line, you get something which is always below \( L \) – shown as the blue (dash-dot) line, and (ii) that the value of this sum at \( \theta^m \) is actually equal to \( L \). It should be clear then from the picture that by moving from \( \theta^m \) to any \( \theta' \) which

\[\text{recall this is } \sum_d Q_d(\Theta, \theta^m)\]
Figure 1: $\Theta$ on the horizontal axis, pretending it is just 1-dimensional. The bold line is the log likelihood of the data under various settings of $\Theta$ i.e. $l(x^1 \ldots x^D; \Theta)$. The red line (dashes) shows $Q(\Theta, \theta^m)$. The blue (dash-dot) line shows when $h(\theta^m)$ is added to the $Q$ line. (i) The blue line is always below $L$ (ii) the blue line meets $L$ at $\theta^m$. (iii) so if $\theta^{m+1}$ is the max of $Q$, which is also the max of $Q + h$, $L$ is above that, so $L(\theta^{m+1})$ is above $L(\theta^m)$

represents climbing on the red $Q$ line, and therefore on the blue $Q + h$ line, means moving to a $\theta'$ with a higher $L$ value too. The EM algorithm actually chooses $\theta^{m+1}$ to give the maximum possible value of the $Q$ line, $\theta^{m+1} = \arg \max_{\theta} Q(\Theta, \theta^m)$. This is still not necessarily going to give a maximum value of the $L$ line, but there was an increase. A next iteration of the EM algorithm would be represented by a similar configuration of lines drawn further to the right, involving new $Q(\Theta, \theta^{m+1})$ and $h(\theta^{m+1})$ lines.

4 The actual proofs

Proof of (4) **Lower bound for** $l(x^1 \ldots x^D; \Theta)$
\[ l(x^1 \ldots x^D; \Theta) = \sum_d \left[ \log(\sum_z p(z, x^d; \Theta)) \right] \]  

(6)

\[ = \sum_d \left[ \log(\sum_z p(z|x^d; \theta_m) \frac{p(z, x^d; \Theta)}{p(z|x^d; \theta_m)}) \right] \text{ seemingly pointless mult and divide (7)} \]

\[ = \sum_d \left[ \log(\mathbb{E}_{z|x^d, \theta_m} \left[ \frac{p(z, x^d; \Theta)}{p(z|x^d; \theta_m)} \right]) \right] \text{ re-expressing as an expectation (8)} \]

\[ \geq \sum_d \left[ \mathbb{E}_{z|x^d, \theta_m} \left[ \log(p(z, x^d; \Theta)) \right] \right] \text{ Jensen’s inequality (9)} \]

\[ = \sum_d \left[ \mathbb{E}_{z|x^d, \theta_m} \left[ log(p(z, x^d; \Theta)) \right] \right] + \mathbb{E}_{z|x^d, \theta_m} \left[ -\log(p(z|x^d; \theta_m)) \right] \]  

(10)

\[ = \sum_d [Q^d(\Theta, \theta_m)] + h(\theta_m) \]  

(11)

Note we are just defining \( h(\theta_m) \) to be what you get by gathering up everything not mentioning \( \Theta \) ie. \( \sum_d \mathbb{E}_{z|x^d, \theta_m} \left[ -\log(p(z|x^d; \theta_m)) \right] \)

Proof of (5): this lower bound is actually equal to the likelihood when \( \Theta = \theta_m \)

this means

\[ l(x^1 \ldots x^D; \theta_m) = \sum_d [Q^d(\theta_m, \theta_m)] + h(\theta_m) \]

Backing up to the line where Jensen’s inequality is used

\[ \sum_d \left[ \mathbb{E}_{z|x^d, \theta_m} \left[ \log(p(z, x^d; \Theta)) \frac{p(z|x^d; \theta_m)}{p(z|x^d; \theta_m)} \right] \right] \]

If \( \Theta \) is set to \( \theta_m \), looking just at the term whose expectation is being taken:

\[ \log \frac{p(z, x^d; \theta_m)}{p(z|x^d; \theta_m)} = \log \left( \frac{p(z, x^d; \theta_m)}{p(z|x^d; \theta_m)/p(x^d; \theta_m)} \right) = \log(p(x^d; \theta_m)) \]

As this term does not contain \( z \) it can be taken out the summation that is implied by the \( \mathbb{E}_{z|x^d, \theta_m} \) term:
\[
\mathbb{E}_{z|x^d,\theta^m} \left[ \log( \frac{p(z, x^d; \Theta)}{p(z|x^d; \theta^m)} ) \right] = \sum_z p(z|x^d; \theta^m) \log(p(x^d; \theta^m)) \\
= \log(p(x^d; \theta^m)) \sum_z p(z|x^d; \theta^m) \\
= \log(p(x^d; \theta^m))
\]

Summed over all \(d\) this just is \(l(x^1 \ldots x^D; \Theta)\) and this means that the expression giving the lower-bound for \(l(x^1 \ldots x^D; \Theta)\) is actually tight for \(\Theta = \theta^m\).