# EM

Ananth Ranganathan

13th October

**Abstract**

The EM derivation in one place. This place is not unique as a google search will easily reveal!

## 1 EM

The EM algorithm is simple a way to calculate the maximum likelihood parameters of the data in some model. We denote the parameters by $\theta$ and the data by $y$. EM deals with the special case where the model has some hidden variables $x$ on which the likelihood depends, i.e the likelihood is given as

$$\begin{aligned} f(y;\theta) &= p(y|\theta) \\ &= \int_x p(y,x|\theta) \\ &= \int_x p(y|x,\theta)p(x|\theta) \end{aligned}$$

For the sake of convenience, we deal with the log-likelihood instead of the likelihood function itself

$$L(\theta) = \log \int_x p(y,x|\theta)$$

We now trivially introduce a distribution on the hidden variables $x$ and use Jensen's inequality to convert this into a lower bound on the log-likelihood

$$\begin{aligned} L(\theta,\hat{\theta}) &= \log \int_x \frac{p(y,x|\hat{\theta})q(x|\theta)}{q(x|\theta)} \\ &\geq \int_x q(x|\theta) \log \frac{p(y,x|\hat{\theta})}{q(x|\theta)} \qquad (1) \end{aligned}$$

In the E-step, we find the distribution $q(x|\theta)$ that maximizes the bound on $L(\theta,\hat{\theta})$ for a fixed $\hat{\theta}$, while in the M-step we find the $\hat{\theta}$ that maximizes $L(\theta,\hat{\theta})$ for a fixed $q(.|.)$ distribution. The E-step can be derived in two ways as demonstrated below.

## 1.1 The E-step

### 1.1.1 Method 1 - Minimizing the KL divergence

Let us denote the lower bound by $Q(\theta, \hat{\theta})$. For the $i$th iteration of EM, we get

$$\begin{aligned}
Q(\theta, \theta^i) &= \int_x q^i(x|\theta) \log \frac{p(y, x|\theta^i)}{q(x|\theta)} \\
&= \int_x q^i(x|\theta) \log \frac{p(x|y, \theta^i) p(y|\theta^i)}{q^i(x|\theta)} \\
&= \int_x q^i(x|\theta) \log p(y|\theta^i) - \int_x q^i(x|\theta) \log \frac{q^i(x|\theta)}{p(x|y, \theta^i)} \\
&= \log p(y|\theta^i) - \mathrm{KL}(q||p)
\end{aligned}$$

where $\mathrm{KL}(q||p)$ is the KL-divergence. To maximize $Q(\theta, \hat{\theta})$ , we need to minimize the KL-divergence. But since the KL-divergence attains its minimum value of zero when $q^i(x|\theta) = p(x|y, \theta^i)$.

Hence in the E-step, the lower bound is maximized when the distribution over $x$ is taken to be $p(x|y, \theta^i)$.

### 1.1.2 Method 2 - Variational calculus

As in Method 1, we define $Q(\theta, \hat{\theta})$ as the lower bound

$$\begin{aligned}
Q(\theta, \theta^i) &= \int_x q^i(x|\theta) \log \frac{p(y, x|\theta^i)}{q^i(x|\theta)} \\
&= \int_x q^i(x|\theta) \log \frac{p(y, x|\theta^i)}{q^i(x|\theta)}
\end{aligned}$$

To maximize $Q$ wrt $q^i(x|\theta)$, we define the objective function by introducing a Lagrangian multiplier for the condition that the $q^i(x|\theta)$ function is a probability distribution

$$F = \int_x q^i(x|\theta) \log p(y, x|\theta^i) - \int_x q^i(x|\theta) \log q^i(x|\theta) + \lambda \left(1 - \int_x q^i(x|\theta)\right)$$

Functionally differentiating this wrt $q^i(x|\theta)$ and setting the derivative to zero, we get

$$\frac{\partial Q}{\partial q^i(x|\theta)} = 0 = \log p(y, x|\theta^i) - \log q^i(x|\theta) - 1 - \lambda$$

whence

$$\begin{aligned}
q^i(x|\theta) &= p(y, x|\theta^i) e^{-1-\lambda} \\
&\propto p(y, x|\theta^i)
\end{aligned}$$

and since the distribution has to be normalized, we get

$$\begin{aligned} q^i(x|\theta) &= \frac{p(y,x|\theta^i)}{\int_x p(y,x|\theta^i)} \\ &= p(x|y,\theta^i) \end{aligned} \tag{2}$$

which is the same result we obtained in Method 1.

## 1.2 The M-step

In the M-step, we maximize $\theta$ for a fixed distribution on the hidden variables $x$. Substituting the value of $q$ obtained from the E-step into (1), we get

$$\begin{aligned} L(\theta,\theta^i) &= \int_x p(x|y,\theta^i) \log \frac{p(x,y|\theta)}{p(x|y,\theta^i)} \\ &= \int_x p(x|y,\theta^i) \log p(x,y|\theta) - \int_x p(x|y,\theta^i) \log p(x|y,\theta^i) \end{aligned}$$

Note that the inequality gets transformed into an equality because the bound calculated in the E-step is tight. Also, second term in the above equation does not depend on $\theta$ and can be ignored for the purpose of maximizing $L$ wrt $\theta$. Hence, the M-step is

$$\theta^{i+1} = \underset{\theta}{\operatorname{argmax}} \int_x p(x|y,\theta^i) \log p(x,y|\theta) \tag{3}$$