# EM algorithm

Saki Kuzushima

June 27, 2019

## 1 Intro

This memo sketches the proof of the monotonicity of the EM algorithm.

## 2 Set up

We assume a simple mixture model. Let $\phi, \gamma$ be the parameters and $\gamma$ is the latent parameter. Let $y$ be data. We want to maximize the marginal probability of $\phi$ given $y$

$$\max p(\phi|y) \tag{1}$$

This often corresponds to the MAP (maximum a posteriori) estimate in the Bayesian framework. However, the functional form of (1) is often intractable. EM is useful when (1) is intractable but the joint distribution of $\phi$ and $\gamma$, $p(\phi, \gamma|y)$ is tractable. One of the typical cases is mixture models.

## 3 EM algorithm

(1) Start with a rough estimate $\phi = \phi^0$
(2) While the increase in the maximand $p(\phi|y)$ becomes less than the arbitrary threshold, Do;
(2-1): E step. Obtain the Q function. i.e. the expectation of the complete log likelidhood.

$$Q = \mathbb{E}_\gamma[\log p(\phi, \gamma|y)] = \int \log p(\phi, \gamma|y) p(\gamma|\phi^t, y) d\gamma \tag{2}$$

(2-2): M step. Maximize the Q function with respect to $\phi$.

$$\phi^{t+1} = \text{argmax} \mathbb{E}_\gamma[\log p(\phi, \gamma|y)] = \text{argmax} \int \log p(\phi, \gamma|y) p(\gamma|\phi^t, y) d\gamma \tag{3}$$

## 4 Monotonicity proof

We want to show

$$p(\phi^{t+1}|y) > p(\phi^t|y) \tag{4}$$

where $t$ is the current iteration and $t+1$ is the next iteration. First, we can we write (1)

$$p(\phi|y) = \frac{p(\phi, \gamma|y)}{p(\gamma|\phi, y)} \tag{5}$$

$$\iff \log p(\phi|y) = \log p(\phi, \gamma|y) - \log p(\gamma|\phi, y) \tag{6}$$

Taking the expectation with respect to $\gamma$ in the current iteration, $p(\gamma|\phi^t, y)$,

$$\log p(\phi|y) = \mathbb{E}_\gamma[\log p(\phi, \gamma|y)] - \mathbb{E}_\gamma[\log p(\gamma|\phi, y)] \tag{7}$$

Showing (4) is equivalent to showing

$$\log p(\phi^{t+1}|y) > \log p(\phi^t|y) \tag{8}$$

Using (7),

$$\log p(\phi^{t+1}|y) - \log p(\phi^t|y) \tag{9}$$

$$= \mathbb{E}_\gamma[\log p(\phi^{t+1}, \gamma|y)] - \mathbb{E}_\gamma[\log p(\phi^t, \gamma|y)] \tag{10}$$

$$- \big\{\mathbb{E}_\gamma[\log p(\gamma|\phi^{t+1}, y)] - \mathbb{E}_\gamma[\log p(\gamma|\phi^t, y)]\big\} \tag{11}$$

Observe that the last line is a KL divergence.

$$- \big\{\mathbb{E}_\gamma[\log p(\gamma|\phi^{t+1}, y)] - \mathbb{E}_\gamma[\log p(\gamma|\phi^t, y)]\big\} \tag{12}$$

$$= \mathbb{E}_\gamma[\log p(\gamma|\phi^t, y)] - \mathbb{E}_\gamma[\log p(\gamma|\phi^{t+1}, y)] \tag{13}$$

$$= \int \log p(\gamma|\phi^t, y) p(\gamma|\phi^t, y) d\gamma - \int \log p(\gamma|\phi^{t+1}, y) p(\gamma|\phi^t, y) d\gamma \tag{14}$$

$$= \int \log \frac{p(\gamma|\phi^t, y)}{p(\gamma|\phi^{t+1}, y)} p(\gamma|\phi^t, y) d\gamma \tag{15}$$

$$= KL\big(p(\gamma|\phi^t, y) \| p(\gamma|\phi^{t+1}, y)\big) \tag{16}$$

$$\geq KL\big(p(\gamma|\phi^t, y) \| p(\gamma|\phi^t, y)\big) = 0 \tag{17}$$

Therefore,

$$\log p(\phi^{t+1}|y) > \log p(\phi^t|y) \tag{18}$$

$$\iff \mathbb{E}_\gamma[\log p(\phi^{t+1}, \gamma|y)] > \mathbb{E}_\gamma[\log p(\phi^t, \gamma|y)] \tag{19}$$

As long as we set $\phi^{t+1}$ so that it satisfies (19), the EM algorithm is guaranteed to increase the maximand (1). Usually, we choose $\phi^{t+1} = \text{argmax} \mathbb{E}_\gamma[\log p(\phi^t, \gamma|y)]$.

## 5 Refernece

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A. and Rubin, D.B., 2013. Bayesian Data Analysis. Chapman and Hall/CRC., Chapter 13 p.320-321.