EM algorithm review*

Konstantinos G. Derpanis
kosta@cs.yorku.ca
Version 1.0

March 8, 2005

1 Background

- Expectation Maximization (EM) is a method to design algorithms and not an algorithm in itself
- Dempster, Laird, and Rubin first proposed it in 1977 [1]
- early predecessors include: Iteratively Reweighted Least-Squares and Newcomb in 1886 proposed a two class problem
- EM-like methods were used long before EM was formalized

2 EM algorithm

- start at some random θ^0
- $$Q(\theta, \theta^t) = E_z\{\ln p(x, z|\theta)|x, \theta^t\} \tag{1}$$

  x represents observed data and z is the missing data (e.g., cluster label in EM clustering). We seek a value θ that maximizes (1),

  $$\theta^{t+1} = \arg\max_{\theta} Q(\theta, \theta^t) \tag{2}$$

  $$\theta^{t+1} = \arg\max_{\theta} E_z\{\ln p(x, z|\theta)|x, \theta^t\} \tag{3}$$

  observation x, unobserved z, parameters θ, guess θ_t, complete $y = x \cup z$
- 3 classes of problems for EM: lost data (e.g., in transmission), inaccessible data (e.g., the life span of a light bulb but only have a short period of time to gather data) and uninteresting data (e.g., in clustering we might not be interested in the membership data)

*The review is adapted from a lecture given by Minas Spetsakis (York University)

- We show the monotonicity of EM

$$L(\theta|x) = p(x|\theta) = \int p(x,z|\theta)dz \tag{4}$$

Taking the ln of both sides, yields,

$$\ln L(\theta|x) = \ln p(x|\theta) = \ln \int p(x,z|\theta)dz \tag{5}$$

- could formulate non-linear optimization problem (e.g., Newton-Raphson) but generally HARD
- add a self-cancelling term to (5), notice the introduction of θ^t,

$$\ln \int p(x,z|\theta)dz = \ln \int \left(\frac{p(x,z|\theta)}{p(z|x,\theta^t)}\right) p(z|x,\theta^t)dz \tag{6}$$

- Apply Jensen's inequality to (6),

$$\begin{aligned}
\ln \int \left(\frac{p(x,z|\theta)}{p(z|x,\theta^t)}\right) p(z|x,\theta^t)dz &\geq \int \ln\left(\frac{p(x,z|\theta)}{p(z|x,\theta^t)}\right) p(z|x,\theta^t)dz && (7)\\
&= \int \ln\Big(p(x,z|\theta)\Big) p(z|x,\theta^t)dz - \int \ln\Big(p(z|x,\theta^t)\Big) p(z|x,\theta^t)dz && (8)\\
&= Q(\theta,\theta^t) - Q'(\theta^t) && (9)
\end{aligned}$$

- From (7), if $\theta = \theta^t$,

$$\begin{aligned}
\frac{p(x,z|\theta^t)}{p(z,|x,\theta^t)} &= \frac{p(x|\theta^t)p(z|x,\theta^t)}{p(z,|x,\theta^t)} && (10)\\
&= p(x|\theta^t) && (11)
\end{aligned}$$

since (11) is independent of z, Jensen's strict equality in (7) holds,

$$\ln L(\theta^t|x) = Q(\theta^t,\theta^t) - Q'(\theta^t) \tag{12}$$

- So far we know:

$$\ln L(\theta|x) \geq Q(\theta,\theta^t) - Q'(\theta^t) \tag{13}$$

and

$$\ln L(\theta^t|x) = Q(\theta^t,\theta^t) - Q'(\theta^t) \tag{14}$$

- Let $\theta^{t+1} = \arg\max_\theta Q(\theta,\theta^t)$ and $Q(\theta^{t+1}) > Q(\theta^t,\theta^t)$, e.g., θ^t is not already maximizing $Q(\theta,\theta^t)$,

$$\ln L(\theta^t|x) = Q(\theta^t,\theta^t) - Q'(\theta^t) < Q(\theta^{t+1},\theta^t) - Q'(\theta^t) \leq \ln L(\theta^{t+1}|x) \tag{15}$$

References

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximal likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39:185–197, 1977.

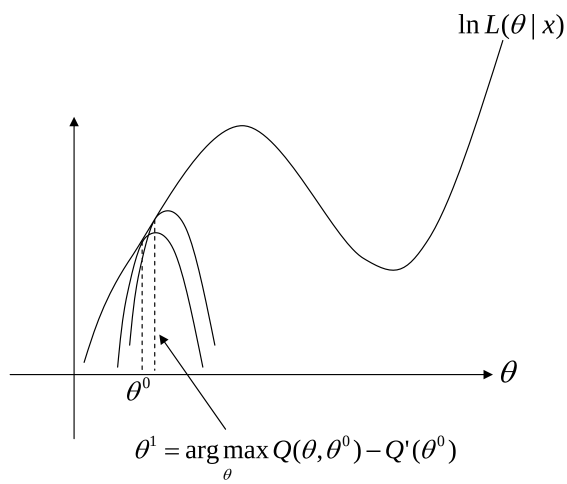

Figure 1. Illustrative conceptualization of EM algorithm.