

THE EM ALGORITHM

Michael Turmon

29 July 1992

1. Motivation and EM View of Data
2. The Algorithm and Its Properties
3. Examples
4. (Relations to Other Methods)
5. Convergence Issues

References:

A. Dempster, N. Laird, and D. Rubin. “Maximum likelihood from incomplete data via the EM algorithm.” *Jour. Royal Stat. Soc. Ser. B*, pp. 1–39, 1977.

J. A. Fessler and A. O. Hero, “Space-alternating generalized expectation-maximization algorithm,” *IEEE Trans. Sig. Proc.*, pp. 2664–2677.

I. Meilijson, “A fast improvement to the EM algorithm on its own terms,” *Jour. Royal. Stat. Soc. Ser. B*, pp. 127–138, 1989.

I. Csiszár and G. Tusnády. “Information geometry and alternating minimization procedures.” *Statistics and Decisions*, supplement issue, pp. 205–237, 1984.

PROBLEM STATEMENT

Observe random variable Y modeled by one of $\{g(y|\theta)\}_{\theta \in \Omega}$.

Estimate the state of nature θ with maximum likelihood.

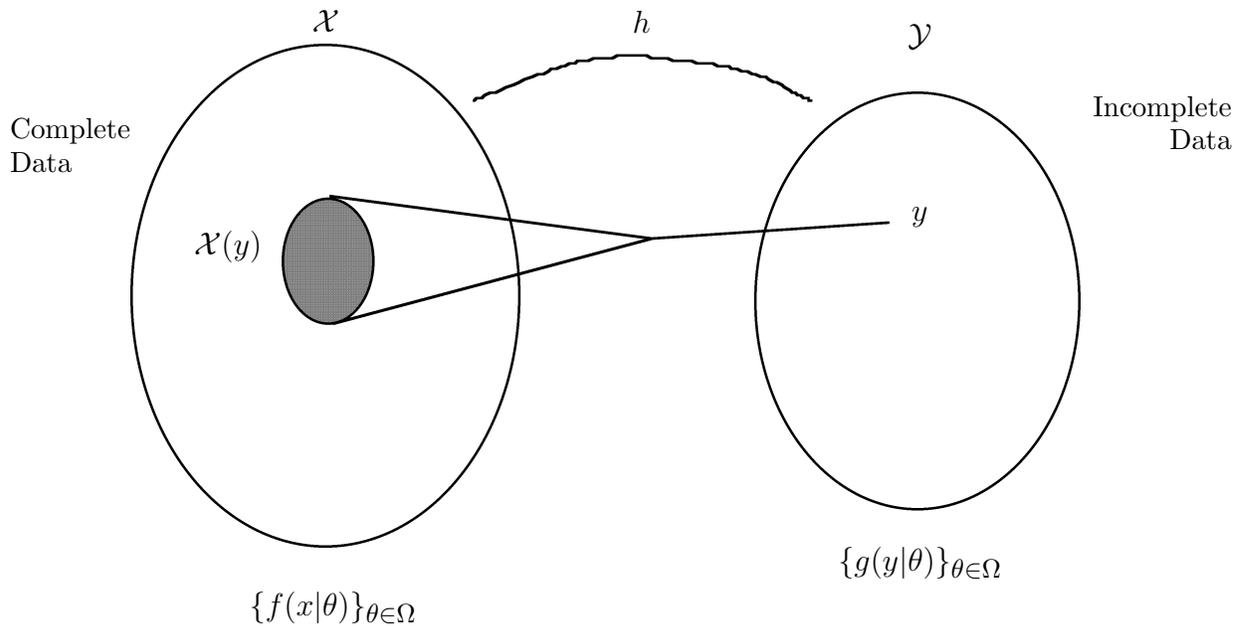
$$\hat{\theta} = \arg \max_{\theta \in \Omega} \log g(y|\theta)$$

$$\left. \frac{\partial}{\partial \theta} \log g(y|\theta) \right|_{\theta=\hat{\theta}} = 0 \quad (1)$$

- Difficult to solve transcendental equation.
Often resort to approximations or numerical solutions.
- The constraint $\theta \in \Omega$ may complicate the derivative condition (1) at boundaries of Ω .

INCOMPLETE-DATA MODEL

Imagine two sample spaces, \mathcal{X} and \mathcal{Y} , and a map $h : \mathcal{X} \rightarrow \mathcal{Y}$.



- Occurrence of $x \in \mathcal{X}$ implies occurrence of $y = h(x) \in \mathcal{Y}$.
- However, only $y = h(x)$ can actually be observed.
Such an observation reveals only that $x \in \mathcal{X}(y)$.

- The sampling densities g and f are related by

$$g(y|\theta) = \int_{\mathcal{X}(y)} f(x|\theta) dx$$

- In a given problem \mathcal{Y} is fixed but \mathcal{X} can be chosen.

THE EM ALGORITHM

The complete-data space is chosen somehow.

The EM algorithm then consists of repeating two steps.

Define the expectation of the complete-data log likelihood

$$Q(\theta|\theta') = E[\log f(X|\theta) | y, \theta'] \quad .$$

Let $\theta^{(0)} \in \Omega$ be any first approximation to θ^* . Then repeat

E-step Compute $Q(\theta|\theta^{(p)})$

M-step Let $\theta^{(p+1)} = \arg \max_{\theta \in \Omega} Q(\theta|\theta^{(p)})$

We want $\theta^{(p+1)}$ to maximize $\log f(x|\theta)$, which is unknown.

Do the next best thing by maximizing its expectation given the data y and the current fit $\theta^{(p)}$.

Why the EM algorithm?

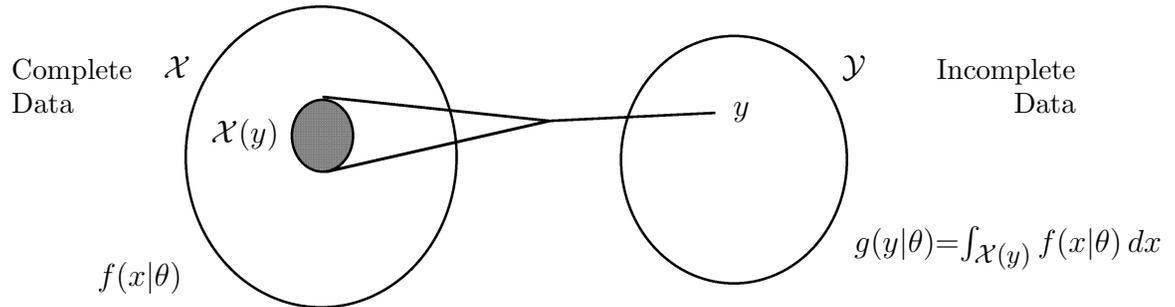
- The constraint $\theta \in \Omega$ can be incorporated into the M-step.
- The likelihood $g(y|\theta^{(p)})$ of the estimates is nondecreasing.
- It is simple and stable.

What about \mathcal{X} ?

\mathcal{X} is chosen to simplify the E and M steps. Often the incomplete-data y is augmented with data that makes estimating θ easy.

For example, if $g(y|\theta)$ is a family of mixtures of M densities, choosing $\mathcal{X} = \mathcal{Y} \times \{1 \dots M\}$ allows each observation have a mark indicating the density it came from.

NON-DECREASING LIKELIHOOD



The conditional density of X given $Y = y$ is

$$k(x|y, \theta) = \begin{cases} \frac{f(x|\theta)}{\int_{\mathcal{X}(y)} f(x|\theta) dx} = \frac{f(x|\theta)}{g(y|\theta)} & x \in \mathcal{X}(y) \\ 0 & x \notin \mathcal{X}(y) \end{cases}$$

So on $\mathcal{X}(y)$, $\log g(y|\theta) = \log f(x|\theta) - \log k(x|y, \theta)$.

Taking the conditional expectation converts this into

$$\log g(y|\theta) = Q(\theta|\theta^{(p)}) - E^{k(x|y, \theta^{(p)})} \log k(X|y, \theta)$$

so that the increase in likelihood between iterations is

$$\begin{aligned} & \left[Q(\theta^{(p+1)}|\theta^{(p)}) - Q(\theta^{(p)}|\theta^{(p)}) \right] \\ & + E^{k(x|y, \theta^{(p)})} \left[\log k(X|y, \theta^{(p)}) - \log k(X|y, \theta^{(p+1)}) \right] \end{aligned}$$

The first term is ≥ 0 because $\theta^{(p+1)}$ maximizes $Q(\cdot|\theta^{(p)})$.

The second is

$$\begin{aligned} -E^{k(x|y, \theta^{(p)})} \log \left[\frac{k(X|y, \theta^{(p+1)})}{k(X|y, \theta^{(p)})} \right] & \geq -\log E^{k(x|y, \theta^{(p)})} \left[\frac{k(X|y, \theta^{(p+1)})}{k(X|y, \theta^{(p)})} \right] \\ & = -\log \int_{\mathcal{X}} k(x|y, \theta^{(p+1)}) dx \\ & = 0 \end{aligned}$$

PROPERTIES OF THE EM ALGORITHM

What the EM algorithm gives you:

1. The EM algorithm is simple and fairly robust.
2. EM can incorporate parameter constraints.
3. EM guarantees monotonically nondecreasing likelihood.
4. A strict maximizer of the likelihood is therefore a stable point of an EM algorithm...
5. ...and if the likelihood is bounded above, the sequence of likelihoods has a finite limit L^* .
 L^* may be the global maximum, a local maximum, a stationary value, or just a stable value of the EM iteration.
6. If $Q(\theta|\theta')$ is continuous in both arguments, then L^* is at least a stationary value.
7. Under conditions, L^* is a local maximum of the likelihood.
8. Under more conditions, if $\theta^{(p)}$ converges then its limit θ^* is a local maximizer.

The chief disadvantage of EM is slow convergence.

Gradient algorithms, such as Newton's method for finding roots of $\frac{\partial}{\partial \theta} g(y|\theta)$, give quadratic convergence in a neighborhood of the maximizer.

A SIMPLE EXAMPLE

Observe $Y = S + N$ where $X \perp N$ and

$$N \sim N(0, \sigma), \sigma \text{ known,}$$

$$S \sim N(0, \theta), \theta \text{ unknown.}$$

Estimate θ given the observation $Y = y$. The MLE is

$$\hat{\theta} = \arg \max_{\theta \geq 0} \log g(y|\theta)$$

where of course

$$\begin{aligned} g(y|\theta) &= N(0, \sigma + \theta) \\ &= (2\pi(\sigma + \theta))^{-1/2} \exp -y^2/2(\sigma + \theta) \end{aligned}$$

Find maximizing θ :

$$\begin{aligned} \frac{\partial}{\partial \theta} \log g(y|\theta) &= -\frac{1}{2} \frac{1}{\sigma + \theta} + \frac{1}{2} \frac{y^2}{(\sigma + \theta)^2} = 0 \\ \theta^{\max} &= y^2 - \sigma \end{aligned}$$

If $y^2 - \sigma \geq 0$, set $\theta^* = \theta^{\max}$.

Otherwise, note $g(y|\theta)$ is decreasing on $[y^2 - \sigma, \infty)$, so put θ^* as close to θ^{\max} as possible:

$$\theta^* = \max(0, y^2 - \sigma)$$

THE SAME EXAMPLE WITH EM

Apply EM to this problem with complete-data $X = (S, N)$.
Its density is $f(x|\theta) = f_S(s|\theta)f_N(n)$.

(1) E-step:

$$\begin{aligned} E[\log f(X|\theta) | y, \theta^{(p)}] &= E[\log f_S(S|\theta) | y, \theta^{(p)}] \\ &\quad + E[\log f_N(N) | y, \theta^{(p)}] \\ &= E\left[-\frac{1}{2} \log \theta - \frac{S^2}{2\theta} \mid y, \theta^{(p)}\right] \\ &\quad + E\left[-\frac{1}{2} \log \sigma - \frac{N^2}{2\sigma} \mid y, \theta^{(p)}\right] + \text{o. t.} \\ &= -\frac{1}{2} \log \theta - \frac{1}{2\theta} E[S^2 | y, \theta^{(p)}] + \text{o. t.} \end{aligned}$$

(2) M-step:

$$\theta^{(p+1)} = E[S^2 | y, \theta^{(p)}] \quad (\text{as expected})$$

Now let $\mu_S(y) = E[S | y, \theta^{(p)}]$, and note

$$E[S^2 | y, \theta^{(p)}] = E[(S - \mu_S(y))^2 | y, \theta^{(p)}] + \mu_S(y)^2$$

Given $Y = y$, S is $N(\theta y / (\sigma + \theta), \sigma\theta / (\sigma + \theta))$, so

$$\theta^{(p+1)} = \frac{\sigma\theta^{(p)}}{\sigma + \theta^{(p)}} + \left(\frac{\theta^{(p)}y}{\sigma + \theta^{(p)}} \right)^2$$

THE EM ITERATION

In this simple case, we can find the fixed point analytically.

At the fixed point, $\theta^{(p)} = \theta^{(p+1)} = \theta^*$:

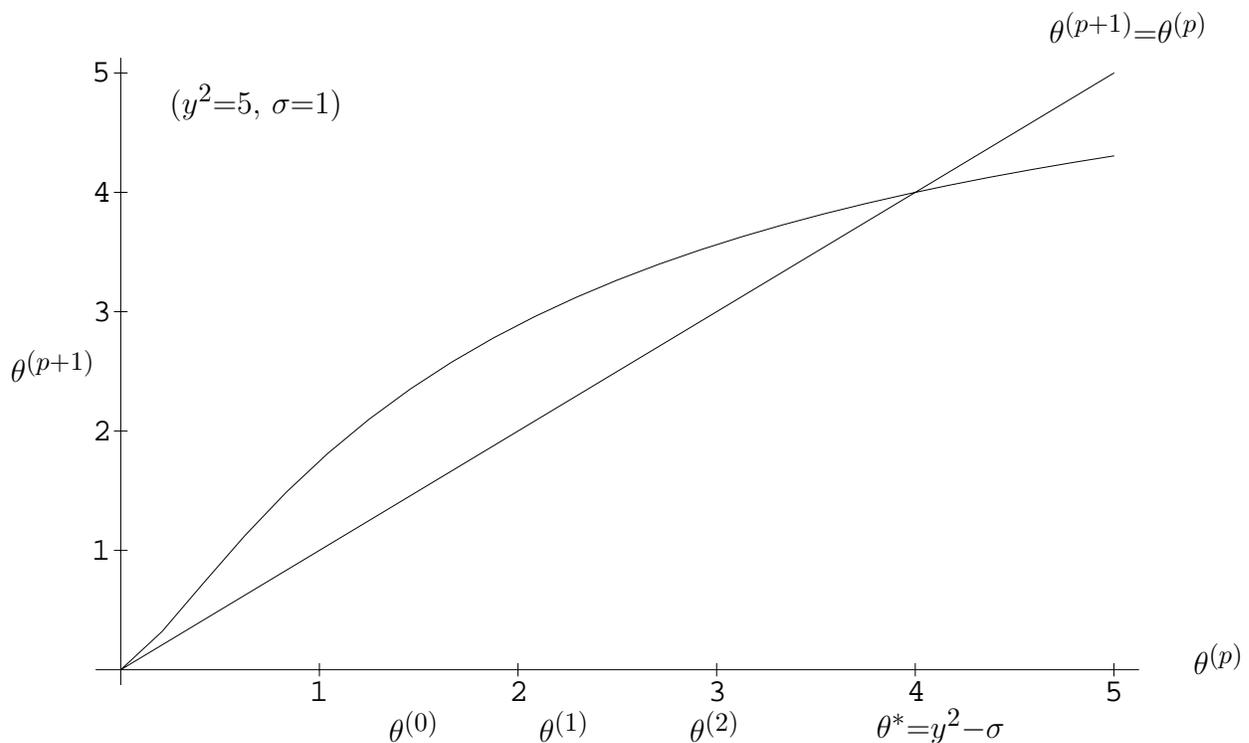
$$\theta^* = y^2 \left(\frac{\theta^*}{\sigma + \theta^*} \right)^2 + \sigma \frac{\theta^*}{\sigma + \theta^*}$$

whose only solutions are 0 and $y^2 - \sigma$, whence

$$\theta^* = \max(0, y^2 - \sigma)$$

just as before.

A sample EM iteration is shown below.



A BOUND ON RELATIVE ENTROPY

Recall the relative entropy $D(p\|q) = E^p \log \frac{p(X)}{q(X)}$.

Let $\mathfrak{D}(S)$ denote the set of densities having S as a support.

For densities $p \in \mathfrak{D}(\mathcal{X}(y))$,

$$\begin{aligned}
 -D(p(x)\|f(x|\theta)) &= -E^p \log \frac{p(X)}{f(X|\theta)} \\
 &= E^p \log \frac{f(X|\theta)}{p(X)} \\
 &\leq \log E^p \frac{f(X|\theta)}{p(X)} \\
 &= \log \int_{\mathcal{X}(y)} p(x) \frac{f(x|\theta)}{p(x)} dx \\
 &= \log \int_{\mathcal{X}(y)} f(x|\theta) dx \\
 &= \log g(y|\theta) \quad ,
 \end{aligned}$$

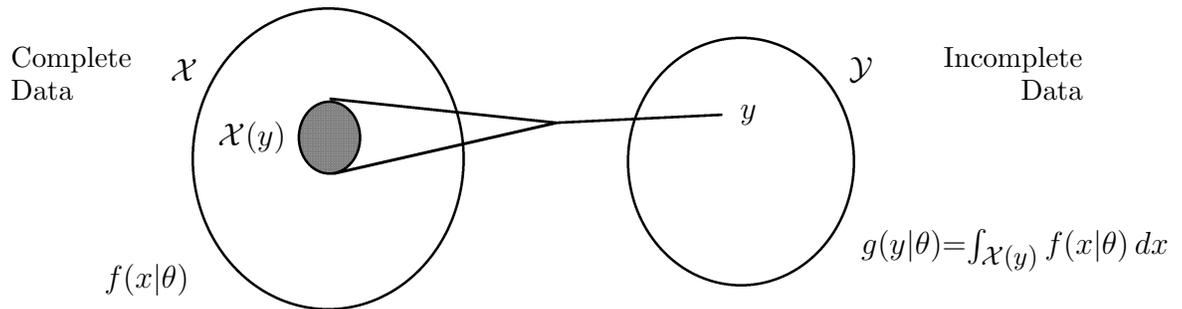
allowing a bound on the distance of any density on $\mathcal{X}(y)$ from $f(x|\theta)$:

$$D(p(x)\|f(x|\theta)) \geq -\log g(y|\theta)$$

Since $g(y|\theta) = P\{X \in \mathcal{X}(y)\} \leq 1$, the bound is ≥ 0 , and positive unless $\mathcal{X}(y)$ is the entire space \mathcal{X} .

The domain constraint thus allows a positive bound.

ENLARGING THE ORIGINAL PROBLEM



Recall the conditional density of X given $Y = y$ is

$$k(x|y, \theta) = \begin{cases} \frac{f(x|\theta)}{g(y|\theta)} & x \in \mathcal{X}(y) \\ 0 & x \notin \mathcal{X}(y) \end{cases}$$

On $\mathcal{X}(y)$, $g(y|\theta) = \int_{\mathcal{X}(y)} f(x|\theta) dx$, so

$$-\log g(y|\theta) = E^{k(x|y, \theta)} \log \frac{k(X|y, \theta)}{f(X|\theta)} = D(k(x|y, \theta) \| f(x|\theta))$$

Max'ing likelihood is thus equivalent to min'ing relative entropy:

$$\hat{\theta} = \arg \max_{\theta \in \Omega} \log g(y|\theta) = \arg \min_{\theta \in \Omega} D(k(x|y, \theta) \| f(x|\theta))$$

More interesting, the conditional $k(x|y, \theta)$ has $\mathcal{X}(y)$ as a support, and it attains the lower bound on $D(p \| f(x|\theta))$. It is thus the member of $\mathfrak{D}(\mathcal{X}(y))$ which is closest to $f(x|\theta)$:

$$k(x|y, \theta) = \arg \min_{p \in \mathfrak{D}(\mathcal{X}(y))} D(p(x) \| f(x|\theta))$$

We might as well solve this more complicated extremal problem:

$$\hat{\theta} = \arg \min_{\theta \in \Omega} \min_{p \in \mathfrak{D}(\mathcal{X}(y))} D(p(x) \| f(x|\theta))$$

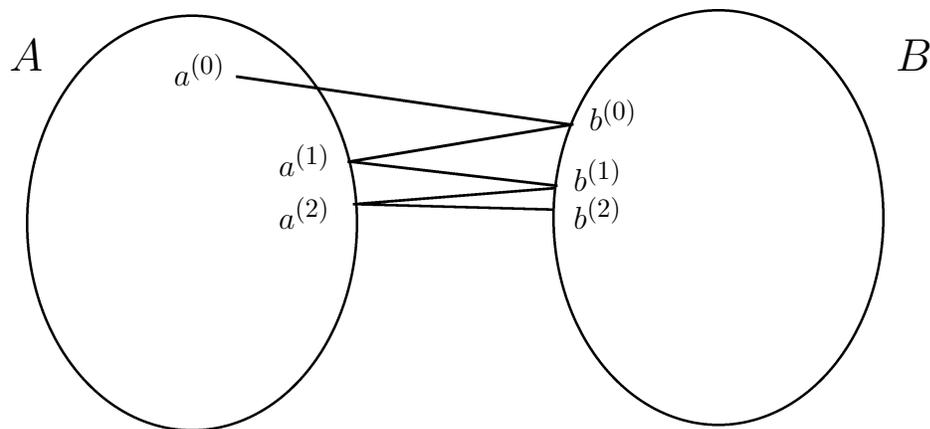
ALTERNATING MINIMIZATION

This is an instance of the alternating minimization of Csiszár and Tusnády for finding the distance between convex sets A and B :

$$\hat{d} = \min_{a \in A} \min_{b \in B} d(a, b)$$

The initial estimate is some $a^{(0)} \in A$, and then

$$\begin{aligned} b^{(0)} &= \arg \min_{b \in B} d(a^{(0)}, b); & a^{(1)} &= \arg \min_{a \in A} d(a, b^{(0)}); \\ b^{(1)} &= \arg \min_{b \in B} d(a^{(1)}, b); & a^{(2)} &= \arg \min_{a \in A} d(a, b^{(1)}), \text{ etc.} \end{aligned}$$



The distance at each stage clearly decreases.

Csiszár and Tusnády have shown that if the distance measure d satisfies certain conditions, the alternating minimization converges to the minimum distance between A and B .

EM AS ALTERNATING MINIMIZATION

For the EM algorithm, the convex sets are Ω and $\mathfrak{D}(\mathcal{X}(y))$, and the distance measure is relative entropy.

The minimization starts with $\theta^{(0)}$ and then

$$k^{(0)} = \arg \min_{p \in \mathfrak{D}(\mathcal{X}(y))} D(p(x) \| f(x|\theta^{(0)}))$$

$$\theta^{(1)} = \arg \min_{\theta \in \Omega} D(k^{(0)}(x) \| f(x|\theta))$$

$$k^{(1)} = \arg \min_{p \in \mathfrak{D}(\mathcal{X}(y))} D(p(x) \| f(x|\theta^{(1)})) \text{ , etc.}$$

The odd-numbered minimizations are easily done, because we know already that the conditional density $k(x|y, \theta^{(p)})$ minimizes entropy relative to $f(x|\theta^{(p)})$. Thus $k^{(p)}(x) = k(x|y, \theta^{(p)})$.

As for the even steps, note

$$D(k^{(p)} \| f(x|\theta)) = E^{k(x|y, \theta^{(p)})} \log \frac{k(X|y, \theta^{(p)})}{f(X|\theta)} = -Q(\theta|\theta^{(p)}) + \text{o. t.},$$

so minimization of the entropy is equivalent to maximization of $Q(\theta|\theta^{(p)})$, which is the E and M steps of the EM algorithm rolled in to one.

Other instances of the Csiszár and Tusnády alternating minimization are the Blahut-Arimoto algorithm for finding the capacity of a communication channel, and Cover's algorithm for finding the log-optimal portfolio for the stock market.

SUMMARY

- The EM algorithm can be used in classical and Bayesian settings when the goal is to maximize likelihood.
- The method works by defining a set of complete-data which, if it were available, would make the problem simpler. The original maximization in the incomplete-data space is transformed into a series of simpler maximizations in the complete-data space.
- The EM algorithm guarantees nondecreasing likelihood, and that maximizers of the likelihood are stable points of the iteration.
- The EM algorithm is one of a family of methods that transform a single difficult minimization into a series of simpler ones. The general paradigm is the alternating minimization of the distance between convex sets.