



A Graphical Illustration of the EM Algorithm

William Navidi

The American Statistician, Vol. 51, No. 1 (Feb., 1997), 29-31.

Stable URL:

<http://links.jstor.org/sici?sici=0003-1305%28199702%2951%3A1%3C29%3AAGIOTE%3E2.0.CO%3B2-2>

The American Statistician is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

A Graphical Illustration of the EM Algorithm

William NAVIDI

The EM algorithm is a method for producing a sequence of parameter estimates that, under mild regularity conditions, converges to the MLE. The EM algorithm is well regarded, in part because of two monotonicity properties: convergence to the MLE is monotone, and the value of the likelihood function increases with each iteration. A graphical illustration of the EM algorithm makes these properties intuitively apparent in the one-parameter case. In addition, a well-known result regarding the rate of convergence of the algorithm can be inferred.

KEY WORDS: Convergence rate; EM algorithm.

1. DESCRIPTION OF THE EM ALGORITHM

The EM algorithm (Dempster, Laird, and Rubin 1977) is a very widely used procedure for maximizing a likelihood function. In a typical application a random vector \mathbf{X} ("the complete data") is postulated to have a density in some family $f(\mathbf{x}|\theta)$, $\theta \in \Theta$. The random vector \mathbf{X} is not observed. Instead, we observe a random vector \mathbf{Y} (the "incomplete data") that is the image of \mathbf{X} under some many-to-one transformation. Let $g(\mathbf{y}|\theta)$ denote the density of \mathbf{Y} . A natural estimator of θ is $\hat{\theta}$, the maximizer of $g(\mathbf{y}|\theta)$. Often it is numerically infeasible to maximize $g(\mathbf{y}|\theta)$ directly. If the complete data likelihood $f(\mathbf{x}|\theta)$ can be maximized, and if certain conditional expectations given \mathbf{Y} can be calculated, then the EM algorithm can be used to produce a sequence of values θ_i that converges to $\hat{\theta}$.

The EM algorithm is valid in many commonly occurring situations. For example, \mathbf{X} and \mathbf{Y} may be multinomial, with \mathbf{Y} representing the collapsing of some of the cells of \mathbf{X} . More generally, if data are missing in \mathbf{Y} , then \mathbf{X} would be taken to be the vector with no missing values. Recently, problems in genetic epidemiology have been modeled where \mathbf{Y} is a vector with each component representing the disease status of a subject in a study, and \mathbf{X} is a vector of ordered pairs representing both the disease status and the unknown genotype of a subject. Many other examples are given in Dempster et al. (1977).

In most cases where the EM algorithm is attractive, the density of \mathbf{X} , $f(\mathbf{x}|\theta)$, can be written in exponential family form:

$$\log f(\mathbf{x}|\theta) = \mathbf{S}(\mathbf{x})^T \theta - a(\theta) + b(\mathbf{x}), \quad (1)$$

where $\mathbf{S}(\mathbf{x})$ is a vector of sufficient statistics for the complete data. Note that we have assumed without loss of generality that θ is the natural parameter of the exponential

family. In this case the sequence of values generated by the EM algorithm is easy to describe. Given an element θ_k in the sequence, θ_{k+1} is generated by the following two-step process:

- The E-step: Compute $E(\mathbf{S}|y, \theta_k)$.
- The M-step: Substitute $E(\mathbf{S}|y, \theta_k)$ for \mathbf{S} in (1). Define θ_{k+1} to be the maximizer of (1).

2. PROPERTIES OF THE EM ALGORITHM

Under mild regularity conditions on the likelihood function (see Wu 1983 for details), and assuming that the initial value in the sequence is not too far from the maximum, the EM algorithm has two important properties that account for its usefulness:

Property I: If θ is a real parameter, the sequence of EM approximations converges monotonically to the MLE, that is, $\theta_n \uparrow \hat{\theta}$ (or $\theta_n \downarrow \hat{\theta}$).

Property II: The EM sequence increases the likelihood at each step, that is, for all k , $g(\mathbf{y}|\theta_{k+1}) \geq g(\mathbf{y}|\theta_k)$.

In the case that $f(\mathbf{x}|\theta)$ is a regular exponential family (the parameter space Θ contains an m -dimensional rectangle where m is the number of sufficient statistics), it turns out that the MLE $\hat{\theta}$ can be characterized as the parameter value under which \mathbf{S} and \mathbf{Y} are uncorrelated, that is, the conditional expectation of \mathbf{S} given \mathbf{Y} is the same as the unconditional expectation (Dempster et al. 1977; Sundberg 1974). We have

$$E(\mathbf{S}|\mathbf{Y}, \hat{\theta}) = E(\mathbf{S}|\hat{\theta}). \quad (2)$$

This result provides considerable insight into Properties I and II. We present this insight in a simple graphical form for the case where $f(\mathbf{x}|\theta)$ is a one-parameter regular exponential family. A result regarding the rate of convergence of the EM algorithm then becomes apparent as well. We begin by deriving the one-parameter version of (2). Our derivation parallels that of Dempster et al. (1977).

Let X be a real random variable distributed according to a one-parameter exponential family. That is,

$$\log f(x|\theta) = S\theta - a(\theta) + b(x). \quad (3)$$

Let Y be the image of X under some many-to-one function. Let $L(\theta) = \log g(y|\theta)$, the log likelihood we wish to maximize. Let $k(x|y, \theta) = f(x|\theta)/g(y|\theta)$, the conditional density of X given Y . Because $\log k(x|y, \theta) = \log f(x|\theta) - \log g(y|\theta)$,

$$\log k(x|y, \theta) = S\theta - a(\theta) - L(\theta) + b(x). \quad (4)$$

Because $E[\partial/\partial\theta \log f(x|\theta)] = 0$ and $E[\partial/\partial\theta \log k(x|y, \theta)|y] = 0$, it follows by differentiating and taking expectations of the right-hand sides of (3) and (4) that

$$E(S|\theta) = a'(\theta) \quad (5)$$

William Navidi is Associate Professor, Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401.

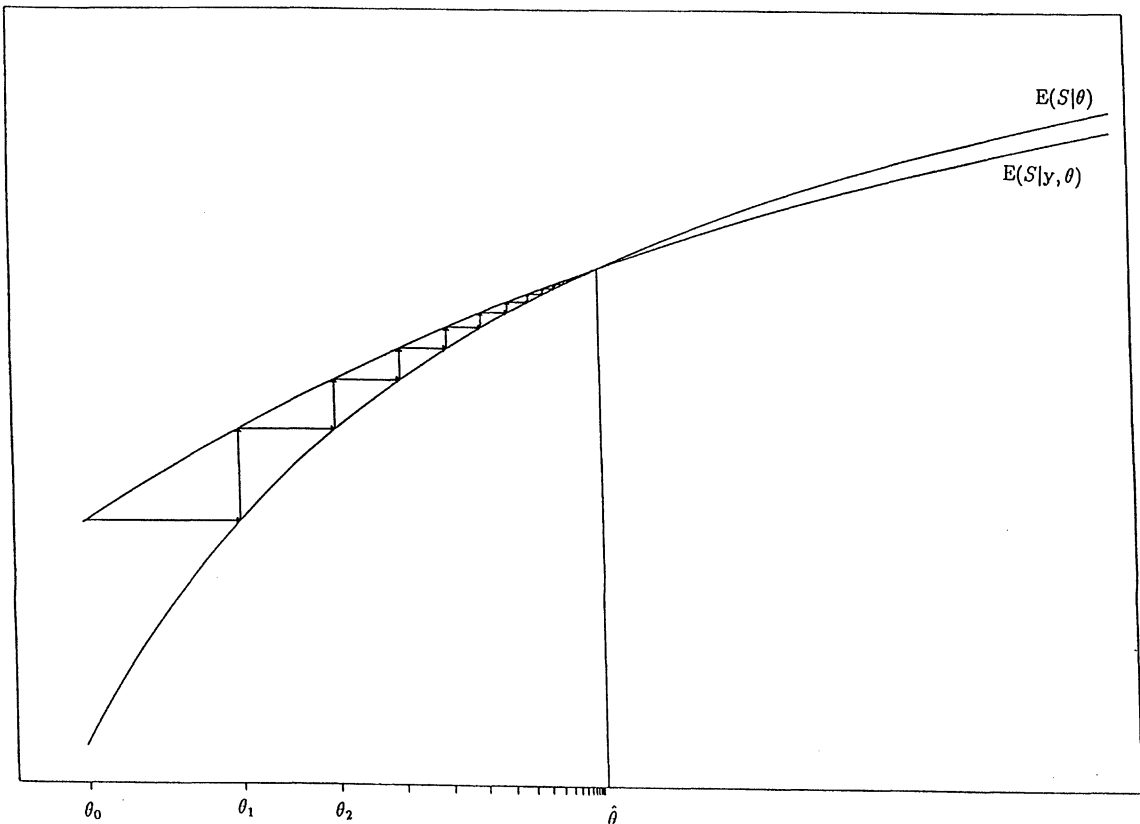


Figure 1. The sequence of EM approximations is obtained by moving rightward and upward between the curves. The sequence converges to the MLE $\hat{\theta}$.

and

$$E(S|y, \theta) = a'(\theta) + L'(\hat{\theta}). \quad (6)$$

Equations (5) and (6) can be combined to yield an expression for $L'(\theta)$:

$$L'(\theta) = E(S|y, \theta) - E(S|\theta). \quad (7)$$

A general form of equation (7) was given by Sundberg (1974). Because $L'(\hat{\theta}) = 0$, the one-parameter version of (2) is now apparent:

$$E(S|y, \hat{\theta}) = E(S|\hat{\theta}). \quad (8)$$

Now let θ_k be the k th term in the sequence of EM approximations to $\hat{\theta}$. The $k + 1$ st term is obtained by computing $E(S|y, \theta_k)$ (the E-step), substituting this quantity for S on the right-hand side of (3), and maximizing over θ (the M-step). One thus obtains θ_{k+1} as the solution to $a'(\theta_{k+1}) = E(S|y, \theta_k)$, which by (5) is equivalent to

$$E(S|\theta_{k+1}) = E(S|y, \theta_k). \quad (9)$$

For all θ , $\text{var}(S|\theta) = a''(\theta)$, and $\text{var}(S|y, \theta) = a''(\theta) + L''(\theta)$, so $a''(\theta) > 0$ and $a''(\theta) + L''(\theta) > 0$. Furthermore, $L''(\theta) < 0$ in some suitably small neighborhood of the MLE $\hat{\theta}$. It follows that both $E(S|\theta)$ and $E(S|y, \theta)$ are increasing functions of θ in a neighborhood of $\hat{\theta}$, with $E(S|\theta)$ increasing more rapidly.

Figure 1 shows a schematic plot of $E(S|\theta)$ and $E(S|y, \theta)$ versus θ . Equation (9) indicates that the sequence $\{\theta_n\}$ of

EM approximations is obtained by moving between the curves, following the arrows as indicated in the figure. It is thus clear that $\theta_n \uparrow \hat{\theta}$, which is Property I.

Equation (7) indicates that the derivative of the log likelihood, $L'(\theta)$, is the distance between the two curves for $\theta < \hat{\theta}$, and the negative of the distance for $\theta > \hat{\theta}$. Now it is clear that for all k , $L(\theta_{k+1}) > L(\theta_k)$, which is Property II.

Figure 1 was generated by taking X to have the exponential distribution with mean $1/\theta$, and Y equal to the greatest integer in the quantity $(X + .99)$. The sufficient statistic S is equal to X . In this simple example it is possible to compute the density of Y as $P(Y = y) = (e^{.99\theta} - e^{-.01\theta})e^{-\theta y}$. Of course, the intuition provided by Figure 1 applies to any exponential family.

The monotonicity of convergence shown in Figure 1 suggests that if one can find starting values θ_0 and θ'_0 whose EM sequences are monotone in opposite directions, then one can conclude that the sequences bracket $\hat{\theta}$. It may then be possible to speed convergence by using averages of estimates from these sequences.

3. RATE OF CONVERGENCE

A diagram similar to Figure 1 illustrates the rate of convergence of the EM algorithm. Let $I_x(\theta)$ and $I_{x|y}(\theta)$ represent the Fisher information associated with $g(y|\theta)$ and $k(x|y, \theta)$, respectively. Differentiating (3) and (4) twice and taking expectations yields

$$I_x(\theta) = a''(\theta) \quad (10)$$

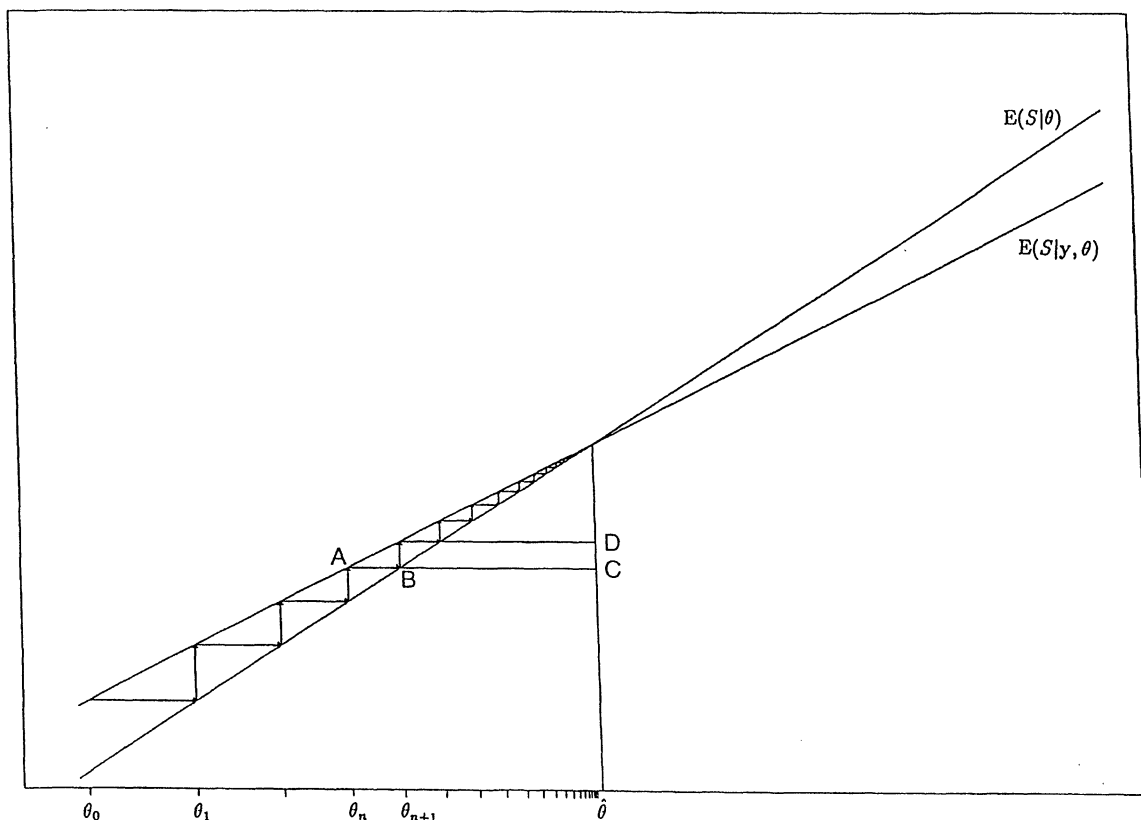


Figure 2. The sequence of EM approximations converges to the MLE $\hat{\theta}$. The limiting rate of convergence is $\overline{BC}/\overline{AC} = I_{x|y}(\hat{\theta})/I_x(\hat{\theta})$.

and

$$I_{x|y}(\theta) = a''(\theta) + L''(\theta). \quad (11)$$

It follows from (5) and (6) that $I_x(\theta)$ and $I_{x|y}(\theta)$ are the derivatives of $E(S|\theta)$ and $E(S|y, \theta)$, respectively. As an aside we note that

$$L''(\theta) = I_{x|y}(\theta) - I_x(\theta). \quad (12)$$

Equation (12) can be found in Louis (1982) and in a more general form in Sundberg (1974) and in Dempster et al. (1977).

As $n \rightarrow \infty$, $\theta_n \rightarrow \hat{\theta}$, and it is appropriate to use the linear approximations

$$E(S|\theta) = E(S|\hat{\theta}) + I_x(\hat{\theta})(\theta - \hat{\theta}) \quad (13)$$

and

$$E(S|y, \theta) = E(S|y, \hat{\theta}) + I_{x|y}(\hat{\theta})(\theta - \hat{\theta}). \quad (14)$$

Figure 2 is the same as Figure 1, except that the curves $E(S|\theta)$ and $E(S|y, \theta)$ are replaced by their linear approximations, with slopes $I_x(\hat{\theta})$ and $I_{x|y}(\hat{\theta})$, respectively. Now it is clear from the figure that

$$\begin{aligned} \overline{AC} &= \hat{\theta} - \theta_n, & \overline{BC} &= \hat{\theta} - \theta_{n+1}, \\ I_x(\hat{\theta}) &= \overline{CD}/\overline{BC}, & I_{x|y}(\hat{\theta}) &= \overline{CD}/\overline{AC} \end{aligned} \quad (15)$$

from which it follows that

$$\lim_{n \rightarrow \infty} \frac{\theta_{n+1} - \hat{\theta}}{\theta_n - \hat{\theta}} = \frac{I_{x|y}(\hat{\theta})}{I_x(\hat{\theta})}. \quad (16)$$

Because $I_x(\hat{\theta})$ is the information in the complete data X , and $I_{x|y}(\hat{\theta})$ can be interpreted as the information in X that is not in the observed data Y , equation (16) shows that the rate of convergence of the EM algorithm is determined by the proportion of information lost due to incompleteness. A more detailed discussion of convergence rates in a general multivariate setting is given in Dempster et al. (1977).

[Received October 1994. Revised May 1996.]

REFERENCES

- Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 39, 1-38.
- Louis, T. A. (1982), "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 44, 226-233.
- Sundberg, R. (1974), "Maximum Likelihood Theory for Incomplete Data from an Exponential Family," *Scandinavian Journal of Statistics*, 1, 49-58.
- Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, 11, 95-103.