*J. R. Statist. Soc.* B (1983),
**45**, *No.* 1, *pp.* 47–50

# On the Convergence of the *EM* Algorithm

By RUSSELL A. BOYLES

*Lawrence Livermore National Laboratory, USA*

[Received August 1980. Final revision April 1982]

SUMMARY

An example is given showing that a sequence generated by a *GEM* algorthm need not converge under the conditions stated in Dempster *et al.*, (1977). Two general convergence results are presented which suggest that in practice a *GEM* sequence will converge to a compact connected set of local maxima of the likelihood function; this limit set may or may not consist of a single point.

*Keywords*: MAXIMUM LIKELIHOOD; EM ALGORITHM; GEM ALGORITHM; INCOMPLETE DATA

## 1. INTRODUCTION AND EXAMPLE

Dempster *et al.*, (1977) (hereinafter abbreviated DLR) discuss the generalized *EM* (*GEM*) algorithm for iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. Unfortunately, Theorem 2 in DLR regarding convergence of parameter sequences generated by *GEM* is incorrect. We first present a counterexample, recalling notation from DLR. We then present two theorems on convergence properties of *GEM* sequences under general conditions. The first is viewed as a correct version of Theorem 2 in DLR, while the second employs sharper assumptions and deals with convergence to stationary points.

For the counterexample, let the complete-data model $f(x \mid \phi)$ be a bivariate normal density with mean $\phi = (\phi_1, \phi_2)$ and unit covariance matrix. Assume there is no missing data, so that the observed data is $y = x$. Thus the incomplete-data log-likelihood $L(\phi)$ is identical to the estimated complete-data log-likelihood, i.e.

$$Q(\phi \mid \phi') \equiv E\{\log f(x \mid \phi) \mid y, \phi'\} = \log f(x \mid \phi) = L(\phi)$$

for all $\phi, \phi' \in \Omega$, where the parameter space $\Omega$ in this example is two-dimensional Euclidean space. A mapping $M$: $\Omega \to \Omega$ corresponds to an *EM* algorithm if $M\phi$ maximizes $Q(\cdot \mid \phi)$. For a *GEM* algorithm we require only $Q(M\phi \mid \phi) \geqslant Q(\phi \mid \phi)$ for all $\phi$, which in the present example is equivalent to

$$L(M\phi) \geqslant L(\phi), \quad \phi \in \Omega. \tag{1}$$

A *GEM* sequence $\{\phi^{(p)}\}$ has the form $\phi^{(p)} = M(\phi^{(p-1)})$ where $M$ satisfies (1). We exhibit below such a sequence satisfying the hypotheses but not the conclusion of Theorem 2 in DLR.

Let $(r, \theta)$ denote polar co-ordinates centred at the observation $y = (y_1, y_2)$. Define $r_0 = 2$, $\theta_0 = 0$ and

$$r_p = 1 + (p+1)^{-1}, \quad \theta_p = \sum_{k=1}^{p} (k+1)^{-1} \tag{2}$$

for $p \geqslant 1$. The required sequence $\{\phi^{(p)}\}$ is given by

*Present address*: Mathematics and Statistics Division, Lawrence Livermore National Laboratory, L-316, Livermore, CA 54550, USA.

 0035–9246/83/45047 $2.00

$$\phi_1^{(p)} = y_1 + r_p \cos\theta_p$$
$$\phi_2^{(p)} = y_2 + r_p \sin\theta_p. \tag{3}$$

It is straightforward to verify that the sequence given by (2) and (3) is generated by iteration of the mapping $M$ defined (in terms of polar co-ordinates) by

$$M(r,\theta) = \begin{cases} (2-r^{-1}, \theta + 1 - r^{-1}), & r > 1, \\ (r, \theta + 1 - r^{-1}), & r \leqslant 1 \end{cases}$$

using the initial values $r_0 = 2, \theta_0 = 0$. To verify (1), note that $L(\phi) = L(r,\theta) = -\ln(2\pi) - r^2/2$, and that $0 < 2 - r^{-1} < r$ for $r > 1$. Thus, $L(M\phi) - L(\phi) = \max\{0, r^2/2 - (2-r^{-1})^2/2\} \geqslant 0$, showing that $\{\phi^{(p)}\}$ is an instance of a *GEM* algorithm.

The remaining hypotheses of Theorem 2 in DLR are

$$\{L(\phi^{(p)})\} \text{ is a bounded sequence} \tag{4}$$

and

$$Q(\phi^{(p+1)} \mid \phi^{(p)}) - Q(\phi^{(p)} \mid \phi^{(p)}) \geqslant \lambda \| \phi^{(p+1)} - \phi^{(p)} \|^2 \tag{5}$$

for some scalar $\lambda > 0$ and all $p$. For the present example, (4) is obvious. To verify (5), note that $Q(\phi^{(p+1)} \mid \phi^{(p)}) - Q(\phi^{(p)} \mid \phi^{(p)}) = L(\phi^{(p+1)}) - L(\phi^{(p)}) = \frac{1}{2}(r_p^2 - r_{p+1}^2)$, while a little algebra yields $\| \phi^{(p+1)} - \phi^{(p)} \|^2 = r_p^2 + r_{p+1}^2 - 2r_p r_{p+1} \cos((p+2)^{-1})$. Taking $\lambda = \frac{1}{2}$, (5) is thus equivalent to $r_{p+1}[r_p \cos((p+2)^{-1}) - r_{p+1}] \geqslant 0$, which in turn is equivalent to $\cos((p+2)^{-1}) \geqslant 1 - (p+2)^{-2}$. The latter inequality is valid for all $p \geqslant 0$, hence the sequence $\{\phi^{(p)}\}$ satisfies all the hypotheses of Theorem 2 in DLR. However, it is obvious from (2) and (3) that $\{\phi^{(p)}\}$ converges to the circle of unit radius and centre $y$, not to a single point $\phi^*$ as claimed by DLR.

## 2. CONVERGENCE RESULTS

Although assumptions (4) and (5) do not imply that $\{\phi^{(p)}\}$ is a convergent sequence, together with the *GEM* property (1) they do imply

$$\| \phi^{(p+1)} - \phi^{(p)} \| \to 0, \quad \text{as } p \to \infty. \tag{6}$$

In general, property (6) is the only benefit derived from assumption (5), and (5) is difficult to verify in practice. Therefore, in the following discussion we adopt (6) as one of our basic assumptions; this is less restrictive than assuming (5) as in DLR. The important problem of finding useful sufficient conditions for the key property (6) is left for future research. (See Wu, 1981 for new work on this and other aspects of *GEM* convergence.)

In the following discussion $\Omega$ is a subset of some finite-dimensional Euclidean space, the norm $\| \cdot \|$ in (6) being the usual one. Let $\Omega^0$ denote the interior of $\Omega$, and for real numbers $\lambda$ let $\{L \geqslant \lambda\}$ denote the set $\{\phi \in \Omega: L(\phi) \geqslant \lambda\}$, with a similar interpretation for the notation $\{L = \lambda\}$. To avoid trivialities, we assume $\lambda_0 \equiv L(\phi^{(0)}) > -\infty$. In addition to this and (6), we assume:

$$\{L \geqslant \lambda_0\} \text{ is compact} \tag{7}$$

and

$$L \text{ is continuous on } \{L \geqslant \lambda_0\}. \tag{8}$$

Assumption (7) reflects our interest in convergence to *finite* parameter values. In some cases, unboundedness of the set $\{L \geqslant \lambda_0\}$ may be avoided by suitable choice of parametrization. In general, if prior information is available in the form of a proper density on $\Omega$, (7) will usually be satisfied by the log-posterior even if not by $L$.

We view Theorem 1 below as a correct version of Theorem 2 in DLR. Assumptions (7) and (8) have replaced their consequence (4), but on the other hand (6) has replaced the stronger

assumption (5). Assumptions (7) and (8) are mild regularity conditions general enough to cover a wide range of practical *GEM* applications.

*Theorem* 1. Assume (6), (7) and (8). Then there exists $\lambda^* \in [\lambda_0, +\infty)$ such that $L(\phi^{(p)}) \uparrow \lambda^*$ as $p \to \infty$. Moreover, $\{\phi^{(p)}\}$ converges to a compact, connected component of $\{L = \lambda^*\}$.

*Proof.* The first conclusion follows from (4), as noted in DLR, and (4) is a consequence of (7) and (8). Since (1) and (7) imply $\{\phi^{(p)}\}$ is bounded, Theorem 28.1 of Ostrowski (1966) asserts that (6) implies a compact, connected set of limit points for $\{\phi^{(p)}\}$. For any such limit point $\phi^*$, (8) implies $L(\phi^*) = \lambda^*$.

In the example of Section 1, $\{\phi^{(p)}\}$ converges to the circle of unit radius surrounding the maximum likelihood estimate. In general, if the connected components of $\{L \geqslant \lambda\}$ are convex for $\lambda$ sufficiently close to $\lambda^*$, Theorem 1 implies that eventually $\{\phi^{(p)}\}$ will not leap over valleys in the log-likelihood surface. Of course, if $\{L = \lambda^*\}$ is a discrete set, Theorem 1 implies convergence of $\{\phi^{(p)}\}$ to some $\phi^* \in \{L = \lambda^*\}$.

A result stronger than Theorem 1 will now be obtained under sharper assumptions. Here we are concerned with convergence to points in $S = \{\phi \in \Omega^0 : DL(\phi) = 0\}$, the set of stationary points of $L$. Appropriate assumptions are:

$$\{L \geqslant \lambda_0\} \text{ is a compact subset of } \Omega^0, \tag{7}'$$

$$L \text{ and } Q(\cdot \mid \phi),\ \phi \in \{L \geqslant \lambda_0\}, \text{ are differentiable in an open set containing } \{L \geqslant \lambda_0\}, \tag{8}'$$

$$M \text{ is continuous on } \{L \geqslant \lambda_0\}, \tag{9}$$

$$D^{10} Q(M\phi \mid \phi) = 0 \quad \text{for all } \phi \in \{L \geqslant \lambda_0\}. \tag{10}$$

*Theorem* 2. Assume (6), (7)', (8)', (9) and (10). Then $\{\phi^{(p)}\}$ converges to a compact, connected component of $S \cap \{L = \lambda^*\}$, where $\lambda^*$ is given in Theorem 1.

*Proof.* By (6) and (9), any limit point $\phi^*$ of $\{\phi^{(p)}\}$ satisfies $M\phi^* = \phi^*$, and (1) implies $\phi^* \in \{L \geqslant \lambda_0\}$. As discussed by DLR and Baum *et al.*, (1970), assumption (10) now implies $DL(\phi^*) = 0$. The rest follows as in Theorem 1.

Theorem 2 does not in itself guarantee convergence of $\{\phi^{(p)}\}$ to a set of local maxima. However, as pointed out by DLR (p. 10) and Murray (1977), convergence of *GEM* to saddle points of $L$ will not be a problem in practice. Since *GEM* cannot converge to a (strict) local minimum, the practical implication of Theorem 2 is that under the stated assumptions $\{\phi^{(p)}\}$ will seek an isolated plateau of local maxima, and for $p$ sufficiently large will not leap over valleys to neighbouring plateaux. In particular, if $S$ is a discrete set then we have $\phi^{(p)} \to \phi^* \in S \cap \{L = \lambda^*\}$ as $p \to \infty$.

## 3. ACKNOWLEDGEMENTS

I am indebted to a referee for suggesting the example in Section 1. Also, the possibility of using Theorem 28.1 of Ostrowski (1966) to give a streamlined proof of Theorem 1 was brought to my attention by Wu (1981). In an earlier version (Boyles, 1980) Ostrowski's result is essentially proven "from scratch" in the presence of an additional (unnecessary) technical assumption on $L$.

## REFERENCES

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

Boyles, R. A. (1980) Convergence results for the *EM* algorithm. Technical Report No. 13, Division of Statistics, University of California, Davis.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Statist. Soc.* B, **39**, 1–38.

Murray, G. D. (1977) Contribution to discussion of paper by A. P. Dempster, N. M. Laird and D. B. Rubin. *J. R. Statist. Soc.* B, **39**, 27–28.

Ostrowski, A. M. (1966) *Solution of Equations and Systems of Equations*, 2nd edition. New York: Academic Press.

Wu, C. F. (1981) On the convergence of *EM* algorithm. Technical Report No. 642, Department of Statistics, University of Wisconsin, Madison.