The Annals of Statistics
1983, Vol. 11, No. 1, 95–103

ON THE CONVERGENCE PROPERTIES OF THE EM ALGORITHM[1]

BY C. F. JEFF WU

University of Wisconsin, Madison

Two convergence aspects of the EM algorithm are studied: (i) does the EM algorithm find a local maximum or a stationary value of the (incomplete-data) likelihood function? (ii) does the sequence of parameter estimates generated by EM converge? Several convergence results are obtained under conditions that are applicable to many practical situations. Two useful special cases are: (a) if the unobserved complete-data specification can be described by a curved exponential family with compact parameter space, all the limit points of any EM sequence are stationary points of the likelihood function; (b) if the likelihood function is unimodal and a certain differentiability condition is satisfied, then any EM sequence converges to the unique maximum likelihood estimate. A list of key properties of the algorithm is included.

1. Introduction. Dempster, Laird and Rubin (1977) (henceforth abbreviated DLR) introduced the EM algorithm for computing maximum likelihood estimates from incomplete data. The essential ideas underlying the EM algorithm have been presented in special cases by many authors; see DLR for a detailed account. Among them we mention Baum et al. (1970), Hartley and Hocking (1971), Orchard and Woodbury (1972), Sundberg (1974). The DLR paper has made three significant contributions: (i) it recognizes the expectation step (E-step) and the maximization step (M-step) in their general forms, (ii) it gives some theoretical properties of the algorithm, and (iii) it recognizes and gives a wide range of applications in statistics.

However, the proof of convergence of EM sequences in DLR contains an error. The implication from (3.13) to (3.14) in their Theorem 2 fails due to an incorrect use of the triangle inequality. Additional comments on this proof are given in Section 2.2. Therefore the convergence of EM sequence as proved in their Theorems 2 and 3 is cast in doubt. Other results on the monotonicity of likelihood sequence and the convergence rate of EM sequence (Theorems 1 and 4 of DLR) remain valid.

Despite its slow numerical convergence, the EM algorithm has become a very popular computational method in statistics. Contrary to the general experience in numerical optimization, the implementation of the E-step and M-step is easy for many statistical problems, thanks to the nice form of the complete-data likelihood function. Solutions of the M-step often exist in closed form. In many cases the M-step can be performed with a standard statistical package, thus saving programming time. Another reason for statisticians to prefer EM is that it does not require large storage space. These two features are especially attractive to those with free access to small computers.

In this paper, instead of patching up the original proof of DLR, we study more broadly two convergence aspects of the EM algorithm. Our approach is to view EM as a special optimization algorithm and to utilize existing results in the optimization literature.

Formally we have two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. Instead of observing the "complete data" $\mathbf{x}$ in $\mathcal{X}$, we observe the "incomplete data" $\mathbf{y} = \mathbf{y}(\mathbf{x})$ in $\mathcal{Y}$. Let the density function of $\mathbf{x}$ be $f(\mathbf{x} \mid \phi)$ with parameters $\phi \in \Omega$ and let the density function of $\mathbf{y}$ be given by

Received May 1981; revised August 1982.

[1] Research supported by the National Science Foundation Grant No. MCS-7901846.

AMS 1980 *subject classification.* Primary 62F10, 90C30.

Key words and phrases. EM algorithm, GEM algorithm, incomplete data, curved exponential family, maximum likelihood estimate.

$$g(\mathbf{y}\,|\,\phi) = \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x}\,|\,\phi)\,d\mathbf{x},$$

where $\mathcal{X}(\mathbf{y}) = \{\mathbf{x}\colon \mathbf{y}(\mathbf{x}) = \mathbf{y}\}$. The parameters ϕ are to be estimated by the method of maximum likelihood, i.e., by maximizing $g(\mathbf{y}\,|\,\phi)$ over $\phi \in \Omega$. In many statistical problems, maximization of the complete-data specification $f(\mathbf{x}\,|\,\phi)$ is simpler than that of the incomplete-data specification $g(\mathbf{y}\,|\,\phi)$. A main feature of the EM algorithm is maximization of $f(\mathbf{x}\,|\,\phi)$ over $\phi \in \Omega$ (M-step). Since $\mathbf{x}$ is unobservable, we replace $\log f(\mathbf{x}\,|\,\phi)$ by its conditional expectation given $\mathbf{y}$ and the current fit ϕ_p(E-step). To this end, let $k(\mathbf{x}\,|\,\mathbf{y}, \phi) = f(\mathbf{x}\,|\,\phi)/g(\mathbf{y}\,|\,\phi)$ be the conditional density of $\mathbf{x}$ given $\mathbf{y}$ and ϕ. Then the log-likelihood

$$L(\phi') = \log g(\mathbf{y}\,|\,\phi') = Q(\phi'\,|\,\phi) - H(\phi'\,|\,\phi), \tag{1}$$

where $Q(\phi'\,|\,\phi) = E\{\log f(\mathbf{x}\,|\,\phi')\,|\,\mathbf{y}, \phi\}$ and $H(\phi'\,|\,\phi) = E\{\log k(\mathbf{x}\,|\,\mathbf{y}, \phi')\,|\,\mathbf{y}, \phi\}$ are assumed to exist for all pairs (ϕ', ϕ). We now define the EM iteration $\phi_p \to \phi_{p+1} \in M(\phi_p)$ as follows:

E-STEP. Determine $Q(\phi\,|\,\phi_p)$.

M-STEP. Choose ϕ_{p+1} to be any value of $\phi \in \Omega$ which maximizes $Q(\phi\,|\,\phi_p)$.

Note that M is a point-to-set map, i.e., $M(\phi_p)$ is the set of ϕ values which maximizes $Q(\phi\,|\,\phi_p)$ over $\phi \in \Omega$. Many applications of EM are for the curved exponential family, for which the E-step and M-step take special forms.

Sometimes it may not be numerically feasible to perform the M-step. DLR defined a generalized EM algorithm (a GEM algorithm) to be an iterative scheme $\phi_p \to \phi_{p+1} \in M(\phi_p)$, where $\phi \to M(\phi)$ is a point-to-set map, such that

$$Q(\phi'\,|\,\phi) \geq Q(\phi\,|\,\phi) \quad \text{for all} \quad \phi' \in M(\phi). \tag{2}$$

EM is a special case of GEM. For any instance $\{\phi_p\}$ of a GEM algorithm,

$$L(\phi_{p+1}) \geq L(\phi_p) \tag{3}$$

follows from the definition of GEM and the inequality

$$H(\phi\,|\,\phi) \geq H(\phi'\,|\,\phi) \quad \text{for any} \quad \phi' \in \Omega. \tag{4}$$

For proofs of (3) and (4), see Lemma 1 and Theorem 1 of DLR.

In Sections 2.1 and 2.2 we inspect two convergence aspects of the EM and GEM algorithms and discuss the relationship of our results to previous ones in the literature. In Section 3 we summarize the key properties of EM. Potential EM users that are not patient with mathematical details may read the summary before consulting the results in Section 2.

2. Does EM do the job? The original purpose of the EM algorithm was to provide iterative computation of the maximum-likelihood estimates. For a bounded sequence $L(\phi_p)$, (3) implies that $L(\phi_p)$ converges monotonically to some L^*. We want to know whether L^* is the global maximum of $L(\phi)$ over Ω. If not, is it a local maximum or a stationary value? This problem is studied in Section 2.1. A related problem of the convergence of an EM or GEM sequence $\{\phi_p\}$ is studied in Section 2.2. We make the following assumptions for the rest of the paper:

$$\Omega \text{ is a subset in the } r\text{-dimensional Euclidean space } R^r, \tag{5}$$

$$\Omega_{\phi_0} = \{\phi \in \Omega\colon L(\phi) \geq L(\phi_0)\} \text{ is compact for any } L(\phi_0) > -\infty, \tag{6}$$

$$L \text{ is continuous in } \Omega \text{ and differentiable in the interior of } \Omega. \tag{7}$$

As a consequence of (5), (6), (7), we have

$$\{L(\phi_p)\}_{p\geq 0} \text{ is bounded above for any } \phi_0 \in \Omega. \tag{8}$$

The compactness assumption in (6) can be restrictive when no realistic compactification of the original parameter space is available. Such may happen in, say, variance components models and factor analysis. To avoid trivialities, we assume that the starting point ϕ_0 of an EM or GEM algorithm satisfies $L(\phi_0) > -\infty$. When we compute the derivatives of L, Q, H at ϕ_p, *we assume that* ϕ_p *is in the interior* of Ω. Such an assumption is implied, for example, by

(9) Ω_{ϕ_0} is in the interior of Ω for $\phi_0 \in \Omega$.

2.1 *Convergence to global maximum, local maximum or stationary value?* From (3) and (8), $L(\phi_p)$ converges monotonically to some L^*. There is no guarantee that L^* is the global maximum of $L(\phi)$ over Ω for the EM algorithm. Although a global maximization of Q is involved in the M-step, the other term H in $L = Q\text{-}H$ may not cooperate. Even the question of convergence to a local maximum cannot be satisfactorily answered. For simplicity we assume that ϕ_p converges to some ϕ^* in the interior of Ω, that the Hessian matrices $D^{20}\, Q(\phi^* \mid \phi^*)$ and $D^{20}H(\phi^* \mid \phi^*)$ with respect to the first ϕ^* variable exist, and that $D^{20}Q(\phi' \mid \phi)$ is continuous in (ϕ', ϕ). Then $-D^{20}Q(\phi^* \mid \phi^*)$ is non-negative definite (n.n.d.) according to the definition of the M-step, and $-D^{20}H(\phi^* \mid \phi^*)$ is n.n.d. according to Lemma 2 of DLR. Since $D^2L(\phi^*) = D^{20}Q(\phi^* \mid \phi^*) - D^{20}H(\phi^* \mid \phi^*)$, ϕ^* may not be a local maximum.

We give an example (Murray, 1977) to illustrate the possibility of converging to a stationary value but not to a local maximum. The twelve observations in the display below come from a bivariate normal distribution with zero means, correlation coefficient ρ and variances σ_1^2, σ_2^2, where asterisks represent missing values. For these data the likelihood has two

Variable 1:	1	1	−1	−1	2	2	−2	−2	*	*	*	*
Variable 2:	1	−1	1	−1	*	*	*	*	2	2	−2	−2

global maxima at $\rho = \pm$ ½, $\sigma_1^2 = \sigma_2^2 = \frac{8}{3}$ and a saddle point at $\rho = 0$, $\sigma_1^2 = \sigma_2^2 = \frac{5}{2}$. If the starting point of an EM sequence has $\rho = 0$, then $\rho = 0$ remains true for all the subsequent iterations and the sequence converges to the saddle point. If ρ is bounded away from zero in the EM iterations, then the sequence converges to either maximum. We will come back to this example after Corollary 1.

In general, if the log-likelihood L has several (local or global) maxima and stationary points, convergence of the EM sequence to either type of point depends on the choice of starting point. This phenomenon has also been reported in Hasselblad (1969), Wolfe (1970), Haberman (1974), Laird (1978), Rubin and Thayer (1982).

The above discussion on the convergence to local or global maximum may seem redundant, since it is known that no general optimization algorithms are guaranteed to converge to local maxima. Over the last few years some misconception of the power of EM has developed, partly because of the global maximization nature of its M-step. We hope that our discussion will help to remove the misconception.

We now consider the issue of convergence to stationary values. The main theorems of this section rely on the following result. A map A from points of X to subsets of X is called a point-to-set map on X. It is said to be *closed* at x if $x_k \to x$, $x_k \in X$ and $y_k \to y$, $y_k \in A(x_k)$, imply $y \in A(x)$. For point-to-point map, continuity implies closedness.

GLOBAL CONVERGENCE THEOREM. *Let the sequence* $\{x_k\}_{k=0}^{\infty}$ *be generated by* $x_{k+1} \in M(x_k)$, *where* M *is a point-to-set map on* X. *Let a solution set* $\Gamma \subset X$ *be given, and suppose that:* (i) *all points* x_k *are contained in a compact set* $S \subset X$; (ii) M *is closed over the complement of* Γ; (iii) *there is a continuous function* α *on* X *such that* (a) *if* $x \notin \Gamma$, $\alpha(y) > \alpha(x)$ *for all* $y \in M(x)$, *and* (b) *if* $x \in \Gamma$, $\alpha(y) \geq \alpha(x)$ *for all* $y \in M(x)$.

Then all the limit points of $\{x_k\}$ *are in the solution set* Γ *and* $\alpha(x_k)$ *converges monotonically to* $\alpha(x)$ *for some* $x \in \Gamma$.

PROOF. See Zangwill (1969, page 91).

Let M be the point-to-set map in a GEM iteration and let $\alpha(x)$ be the log-likelihood function L. Take the solution set Γ to be

$$\mathscr{M} = \text{set of local maxima in the interior of } \Omega,$$

or

$$\mathscr{S} = \text{set of stationary points in the interior of } \Omega.$$

Condition (iii)(b) follows from (3) and condition (i) follows from (3), (8). Then Theorem 1 follows as a special case of the above theorem.

THEOREM 1. *Let* $\{\phi_p\}$ *be a* GEM *sequence generated by* $\phi_{p+1} \in M(\phi_p)$, *and suppose that* (i) M *is a closed point-to-set map over the complement of* $\mathscr{S}$ (*resp.* $\mathscr{M}$), (ii) $L(\phi_{p+1}) > L(\phi_p)$ *for all* $\phi_p \notin \mathscr{S}$ (*resp.* $\mathscr{M}$).

Then all the limit points of $\{\phi_p\}$ *are stationary points* (*local maxima*) *of* L, *and* $L(\phi_p)$ *converges monotonically to* $L^* = L(\phi^*)$ *for some* $\phi^* \in \mathscr{S}$ (*resp.* $\mathscr{M}$).

For the EM algorithm, it is easy to show that a simple sufficient condition for the closedness of M is that

$$Q(\psi \mid \phi) \text{ is continuous in both } \psi \text{ and } \phi. \tag{10}$$

This condition is very weak and should be satisfied in most practical situations. For convergence to stationary values it turns out to be the only required regularity condition (in addition to those given before). The following theorem is most useful in that it covers a broad range of statistical applications.

THEOREM 2. *Suppose* Q *satisfies the continuity condition* (10). *Then all the limit points of any instance* $\{\phi_p\}$ *of an* EM *algorithm are stationary points of* L *and* $L(\phi_p)$ *converges monotonically to* $L^* = L(\phi^*)$ *for some stationary point* ϕ^*.

PROOF. Since (10) is sufficient for (i) of Theorem 1, it remains to prove (ii) of Theorem 1 for all $\phi_p \notin \mathscr{S}$. Consider a ϕ_p, which is in the interior of Ω by (9). Since ϕ_p maximizes $H(\phi \mid \phi_p)$ over $\phi \in \Omega$ according to (4), $D^{10}H(\phi_p \mid \phi_p) = 0$. Therefore $DL(\phi_p) = D^{10}Q(\phi_p \mid \phi_p) \neq 0$ for any $\phi_p \notin \mathscr{S}$ from the definition of $\mathscr{S}$, implying that ϕ_p is not a local maximum of $Q(\phi \mid \phi_p)$ over $\phi \in \Omega$. From the definition of the M-step, $Q(\phi_{p+1} \mid \phi_p) > Q(\phi_p \mid \phi_p)$. Together with (4), this proves $L(\phi_{p+1}) > L(\phi_p)$ for all $\phi_p \notin \mathscr{S}$. The desired result follows. □

The same argument does not apply when $\mathscr{S}$ is replaced by $\mathscr{M}$. This is easily demonstrated by considering a ϕ_p in $\mathscr{S}$ but not in $\mathscr{M}$. Here $DL(\phi_p) = D^{10}Q(\phi_p \mid \phi_p) = 0$ and ϕ_p may indeed maximize $Q(\phi \mid \phi_p)$ over $\phi \in \Omega$. The EM iteration terminates at ϕ_p, a stationary point but not a local maximum, and $L(\phi_{p+1}) > L(\phi_p)$ does not hold for such ϕ_p. Thus to guarantee convergence to a local maximum, we need a further condition such as (11) below. Since (11) holds for any $\phi \notin \mathscr{S}$, (4) and (11) imply (ii) of Theorem 1 for all $\phi_p \notin \mathscr{M}$. The following theorem is a special case of Theorem 1.

THEOREM 3. *Suppose* Q *satisfies the continuity condition* (10) *and*

$$\sup_{\phi' \in \Omega} Q(\phi' \mid \phi) > Q(\phi \mid \phi) \quad \textit{for any} \quad \phi \in \mathscr{S} \setminus \mathscr{M}. \tag{11}$$

Then all the limit points of any instance $\{\phi_p\}$ *of an* EM *algorithm are local maxima of* L *and* $L(\phi_p)$ *converges monotonically to* $L^* = L(\phi^*)$ *for some local maximum* ϕ^*.

Since (11) is typically hard to verify, the utility of Theorem 3 is somewhat limited. An important class of densities that satisfy (10) is the curved exponential family

(12) $$f(x|\phi) = b(x)\exp\{\phi^T t(x)\}/a(\phi),$$

where the parameters ϕ lie in a *compact submanifold* Ω_0 of the r-dimensional convex region $\Omega = \{\phi: a(\phi) = \int_{\mathcal{X}} b(x)\exp[\phi^T t(x)]\, dx < \infty\}$. Here $Q(\phi'|\phi) = -\log a(\phi') + E\{\log b(x)|y, \phi\} + \phi'^T E\{t(x)|y, \phi\}$. From the compactness of Ω_0 and properties of exponential family, one can easily verify condition (10). We emphasize that the continuity of Q also hold true for many densities outside the exponential family.

Theorem 1 is the most general result for EM and GEM algorithms. The result in Theorem 2 was obtained by Baum et al. (1970) and Haberman (1977) for two special models. Boyles (1980) obtained a similar result for general models but under stronger regularity conditions. One key condition in Baum et al. (1970) and Boyles (1980) is that M is a continuous point-to-point map over Ω, which is stronger than a closed point-to-set map over $\Omega \setminus \mathcal{S}$ (assumed in Theorem 1). Boyles showed that his Lemma 4.2 covers the regular exponential family, a case not frequently encountered in practice, while our Theorem 2 covers a much broader class of applications.

Murray's example illustrates Theorems 2 and 3 well. Convergence of any EM sequence to a stationary value follows from Theorem 2 or Corollary 1. But why does an EM sequence converge to a saddle point in this example? From the assumptions of Theorem 3, it becomes clear that condition (11) is being violated at the saddle point $\rho = 0$, $\sigma_1^2 = \sigma_2^2 = 5/2$. In fact the saddle point maximizes the likelihood over all the parametric specifications with $\rho = 0$. On the other hand, if a particular EM sequence is not attracted toward the hyperplane $\rho = 0$, then violation of (11) is avoided and convergence of this EM sequence to local maxima is guaranteed.

2.2 *Convergence of an* EM *or* GEM *sequence* $\{\phi_p\}$. The convergence of $L(\phi_p)$ to L^* studied in Section 2.1 does not automatically imply the convergence of ϕ_p to a point ϕ^*. (The implication in the reverse direction is true if $D^{10}H(\phi'|\phi)$ is continuous.) Convergence of EM in the latter sense usually requires more stringent regularity conditions and will be addressed in this subsection.

Theorem 2 of DLR incorrectly claimed that, under their conditions (1) and (2), a GEM sequence $\{\phi_p\}$ converges to a point ϕ^* in the closure of Ω. A GEM (but not EM) sequence was given in Boyles (1983) as a counterexample. The DLR argument would be valid only if condition (2) of their theorem could be replaced by the following:

(13) $$Q(\phi_{p+1}|\phi_p) - Q(\phi_p|\phi_p) \geq \lambda\|\phi_{p+1} - \phi_p\| \quad \text{for some} \quad \lambda > 0 \quad \text{and all } p.$$

However, if $\phi_p \to \phi^*$, then (13) implies $\|D^{10}Q(\phi^*|\phi^*)\| = \|DL(\phi^*)\| \geq \lambda > 0$, and ϕ^* can not be a stationary point. There appears to be no simple way to fix their proof. From the numerical viewpoint, the convergence of ϕ_p is not as important as the convergence of $L(\phi_p)$ to stationary values or local maxima. Since the original DLR claim was made in terms of the convergence of $\{\phi_p\}$ and since subsequent EM users often quoted this result, a rigorous study of this problem is worthwhile.

Define $\mathcal{S}(a) = \{\phi \in \mathcal{S}: L(\phi) \equiv a\}$ and $\mathcal{M}(a) = \{\phi \in \mathcal{M}: L(\phi) = a\}$. Under the assumptions of Theorem 1, $L(\phi_p) \to L^*$ and all the limit points of $\{\phi_p\}$ are in $\mathcal{S}(L^*)$(resp. $\mathcal{M}(L^*)$). If $\mathcal{S}(L^*)$(resp. $\mathcal{M}(L^*)$) consists of a single point ϕ^*, i.e., there can not be two different stationary points (resp. local maxima) with the same L^*, then $\phi_p \to \phi^*$.

THEOREM 4. *Let* $\{\phi_p\}$ *be an instance of a* GEM *algorithm satisfying conditions* i) *and* ii) *of Theorem* 1. *If* $\mathcal{S}(L^*)$(resp. $\mathcal{M}(L^*)$) $= \{\phi^*\}$, *where* L^* *is the limit of* $L(\phi_p)$ *in Theorem* 1, *then* $\phi_p \to \phi^*$.

The above assumption $\mathcal{S}(L^*) = \{\phi^*\}$ can be greatly relaxed if we assume $\|\phi_{p+1} - \phi_p\| \to 0$ as $p \to \infty$, a condition *necessary* for the desired result $\phi_p \to \phi^*$.

THEOREM 5. *Let* $\{\phi_p\}$ *be an instance of a* GEM *algorithm satisfying conditions* i) *and* ii) *of Theorem* 1. *If* $\|\phi_{p+1} - \phi_p\| \to 0$ *as* $p \to \infty$, *then all the limit points of* $\{\phi_p\}$ *are in*

a connected and compact subset of $\mathcal{S}(L^*)$ (*respectively* $\mathcal{M}(L^*)$), *where* L^* *is the limit of* $L(\phi_p)$ *in Theorem* 1. *In particular, if* $\mathcal{S}(L^*)$(resp. $\mathcal{M}(L^*)$) *is discrete, i.e. its only connected components are singletons, then* ϕ_p *converges to some* ϕ^* *in* $\mathcal{S}(L^*)$(*resp.* $\mathcal{M}(L^*)$).

PROOF. From assumption (6), $\{\phi_p\}$ is a bounded sequence. According to Theorem 28.1 of Ostrowski (1966), the set of limit points of a bounded sequence $\{\phi_p\}$ with $\|\phi_{p+1} - \phi_p\| \to 0$ as $p \to \infty$ is connected and compact. From Theorem 1, all the limit points of $\{\phi_p\}$ are already in $\mathcal{S}(L^*)$(resp. $\mathcal{M}(L^*)$). The desired result now follows. □

Note that condition (ii) of Theorem 1 for $\phi_p \notin \mathcal{S}$ is automatically satisfied for any EM sequence as demonstrated in the proof of Theorem 2. Theorem 5 was proved in Boyles (1980, 1982, 1983), under different regularity conditions. Theorem 4 was obtained by Hartley and Hocking (1971) for a special model.

Convergence of ϕ_p to a stationary point can be proved without recourse to Theorem 1. Let $\mathcal{L}(L) = \{\phi \in \Omega: L(\phi) = L\}$.

THEOREM 6. *Let* $\{\phi_p\}$ *be an instance of a* GEM *algorithm with the additional property* $D^{10}Q(\phi_{p+1}|\phi_p) = 0$. *Suppose* $D^{10}Q(\phi'|\phi)$ *is continuous in* ϕ' *and* ϕ. *Then* ϕ_p *converges to a stationary point* ϕ^* *with* $L(\phi^*) = L^*$, *the limit of* $L(\phi_p)$, *if either* (a) $\mathcal{L}(L^*) = \{\phi^*\}$, *or* (b) $\|\phi_{p+1} - \phi_p\| \to 0$ *as* $p \to \infty$ *and* $\mathcal{L}(L^*)$ *is discrete.*

PROOF. As in the proofs of Theorems 4 and 5, we can show $\phi_p \to \phi^*$ for some $\phi^* \in \mathcal{L}(L^*)$. $DL(\phi^*) = D^{10}Q(\phi^*|\phi^*) = 0$ follows from (4), $D^{10}Q(\phi_{p+1}|\phi_p) = 0$ and the continuity of $D^{10}Q(\phi'|\phi)$ in ϕ' and ϕ.

Note that EM satisfies $D^{10}Q(\phi_{p+1}|\phi_p) = 0$. Since $\mathcal{M}(L)$ and $\mathcal{S}(L)$ are subsets of $\mathcal{L}(L)$, conditions (a) and (b) in Theorem 6 are stronger than the corresponding ones in Theorems 4 and 5 respectively. The advantage of Theorem 6 is that it does not require conditions (i) and (ii) of Theorem 1. An important special case is the following.

COROLLARY 1. *Suppose that* $L(\phi)$ *is unimodal in* Ω *with* ϕ^* *being the only stationary point and that* $D^{10}Q(\phi'|\phi)$ *is continuous in* ϕ *and* ϕ'. *Then for any* EM *sequence* $\{\phi_p\}$, ϕ_p *converges to the unique maximizer* ϕ^* *of* $L(\phi)$.

From the viewpoint of the user, Theorem 2 and Corollary 1 are the most useful results since they require conditions that are very easy to verify. Their applications to specific statistical problems are numerous, e.g., Haberman (1977), Redner and Walker (1982), Turnbull (1976), Turnbull and Mitchell (1981), Vardi (1982).

There are two key assumptions (in addition to those discussed in Section 2.1) in Theorem 5 for the convergence of $\{\phi_p\}$: discreteness of $\mathcal{S}(L^*)$ and $\|\phi_{p+1} - \phi_p\| \to 0$ as $p \to \infty$. The latter condition may be hard to verify in practice. For the special cases stated below, the condition holds true.

Sufficient conditions for $\|\phi_{p+1} - \phi_p\| \to 0$ *as* $p \to \infty$:

CONDITION 1. There exists a forcing function σ such that

$$Q(\phi_{p+1}|\phi_p) - Q(\phi_p|\phi_p) \geq \sigma(\|\phi_{p+1} - \phi_p\|) \quad \text{for all } p. \tag{14}$$

(A mapping σ: $[0, \infty) \to [0, \infty)$ is said to be a *forcing function* if, for any sequence $\{t_k\}$ in $[0, \infty)$, $\lim_{k\to\infty}\sigma(t_k) = 0$ implies $\lim_{k\to\infty} t_k = 0$.) From (4),

$$L(\phi_{p+1}) - L(\phi_p) \geq Q(\phi_{p+1}|\phi_p) - Q(\phi_p|\phi_p). \tag{15}$$

Since $L(\phi_p) \to L^*$ and σ is a forcing function, (14) and (15) imply $\|\phi_{p+1} - \phi_p\| \to 0$. By

taking $\sigma(t) = \lambda t^2$, $\lambda > 0$, (14) becomes

$$Q(\phi_{p+1}\,|\,\phi_p) - Q(\phi_p\,|\,\phi_p) \geq \lambda \,\|\phi_{p+1} - \phi_p\|^2 \quad \text{for all } p. \tag{16}$$

Except for the regular exponential family, it is difficult to verify (16).

We single out a subclass of GEM algorithms for which (14) is satisfied. Let $B(\phi_p)$ be a positive definite $r \times r$ matrix continuous in ϕ_p. Then a sufficient condition is obtained by letting the GEM iteration $\phi_p \to \phi_{p+1}$ be defined by (17) and (18)

$$\phi_{p+1} = \phi_p + \xi_p B(\phi_p) D^{10} Q(\phi_p\,|\,\phi_p) / \|D^{10} Q(\phi_p\,|\,\phi_p)\|^s, \tag{17}$$

where $s < 1$ and $0 < \xi_p \leq 1$ is chosen such that ϕ_{p+1} is in the interior of Ω,

$$Q(\phi_{p+1}\,|\,\phi_p) - Q(\phi_p\,|\,\phi_p) \geq \alpha D^{10} Q(\phi_p\,|\,\phi_p)^T (\phi_{p+1} - \phi_p), \tag{18}$$

where $0 < \alpha < 1$ is independent of p. For example, for steepest ascent, we would take $B(\phi_p) = I$. The Armijo line search method (Polak, 1971, page 36; Ortega and Rheinboldt, 1970, page 491) is a finite-terminating algorithm designed for finding the steplength ξ_p specified in (17) and (18). Such a ξ_p always exists because ϕ_p is in the interior by assumption, $Q(\phi\,|\,\phi_p)$ is locally increasing in the direction $B(\phi_p) D^{10} Q(\phi_p\,|\,\phi_p)$, and $0 < \alpha < 1$. By the nature of the GEM iteration (17) and (18), the algorithm terminates at a ϕ_p with $D^{10} Q(\phi_p\,|\,\phi_p) = 0$. It is based solely on the directional derivatives of Q and line searches along the directions. We now proceed to prove that the GEM iteration (17) and (18) satisfies (14) with $\sigma(t) = ct^{(2-s)/(1-s)}$. Let

$$\lambda^* = \sup_{\phi \in \Omega_{\phi_0}} \{\text{max eigenvalue of } B(\phi)\}.$$

From (17) and (18),

$$\begin{aligned} Q(\phi_{p+1}\,|\,\phi_p) - Q(\phi_p\,|\,\phi_p) &\geq \alpha \xi_p^{-1} (\phi_{p+1} - \phi_p)^T B^{-1}(\phi_p)(\phi_{p+1} - \phi_p) \,|\, D^{10} Q(\phi_p\,|\,\phi_p)|^s \\ &\geq \alpha \xi_p^{-1} \lambda^{*-1} \|\phi_{p+1} - \phi_p\|^2 (\xi_p^{-1} \lambda^{*-1} \|\phi_{p+1} - \phi_p\|)^{(s)/(1-s)} \\ &\geq \alpha (\lambda^*)^{-(1-s)^{-1}} \|\phi_{p+1} - \phi_p\|^{(2-s)/(1-s)}. \end{aligned}$$

Note that $\lambda^* < \infty$ since Ω_{ϕ_0} is compact. For other iteration schemes, sufficient conditions for $\|\phi_{p+1} - \phi_p\| \to 0$ can be found in Ortega and Rheinboldt (1970, Chapter 14).

We conclude this subsection with a discussion of the discreteness assumption on $\mathscr{S}(L^*)$, which was not mentioned in DLR. If $L(\phi)$ has a ridge of stationary points in which $L(\phi) = L^*$, i.e. $\mathscr{S}(L^*)$ is not discrete, does an EM sequence with $\|\phi_{p+1} - \phi_p\| \to 0$ converge to a ϕ^* in $\mathscr{S}(L^*)$? Or may it sometimes move indefinitely on the ridge? DLR (1977, page 10) made an unwarranted claim that the former must be true. Although (16), an assumption made in DLR, implies $\sum_{p=0}^{\infty} \|\phi_{p+1} - \phi_p\|^2 < \infty$ and $\|\phi_{p+1} - \phi_p\| \to 0$, it is still possible to have $\sum_{p=0}^{\infty} \|\phi_{p+1} - \phi_p\| = \infty$. A sequence with this property may move indefinitely on a ridge. Although we do not have a counter-example to this claim by DLR, our general impression is that, if $\mathscr{S}(L^*)$ is not discrete, convergence of $\{\phi_p\}$ can only be guaranteed under conditions stronger than those assumed in their paper. This belief is further supported by the GEM counterexample of Boyles (1983).

3. Summary of properties of EM algorithm.

(i) Any EM sequence $\{\phi_p\}$ increases the likelihood and $L(\phi_p)$, if bounded above, converges to some L^*.

(ii) If $Q(\psi\,|\,\phi)$ is continuous in both ψ and ϕ, L^* is a stationary value of L. The continuity of Q holds for the important case of a curved exponential family. If ϕ_p converges to some point ϕ^*, ϕ^* is a stationary point under the continuity condition of $D^{10} Q(\phi'\,|\,\phi)$ in ϕ' and ϕ.

(iii) If, in addition to (ii), Q is not trapped at any point ϕ_0 that is a stationary point but not a local maximum of L, i.e. $\sup_{\phi \in \Omega} Q(\phi \mid \phi_0) > Q(\phi_0 \mid \phi_0)$, then L^* is also a local maximum of L. But this condition may be difficult to verify. Since the convergence to stationary value or local maximum or global maximum depends on the choice of starting points, we recommend that several EM iterations be tried with different starting points that are representative of the parameter space.

(iv) If, in addition to (ii) or (iii), $\| \phi_{p+1} - \phi_p \| \to 0$ as $p \to \infty$ and the set of stationary points (local maxima) with a given L value is discrete, then ϕ_p converges to a stationary point (local maximum).

(v) If, in addition to (ii) or (iii), there cannot exist two different stationary points (local maxima) with the same L value, then ϕ_p converges to a stationary point (local maximum).

(vi) A sufficient condition for $\| \phi_{p+1} - \phi_p \| \to 0$ is condition (14), which is satisfied by a class of optimization algorithms (17) and (18), whose iteration scheme is based on the local directional derivatives of Q and Armijo line searches along the chosen direction. For a regular exponential family, a special case (16) of (14) is satisfied by the EM algorithm.

(vii) If $L(\phi)$ is unimodal in Ω and has only stationary point and $D^{10}Q(\phi' \mid \phi)$ is continuous in ϕ and ϕ', then ϕ_p converges to the unique maximizer ϕ^* of $L(\phi)$.

(viii) If the set of stationary points (local maxima) with a given L value, denoted $\mathscr{S}(L)$ (respectively $\mathscr{M}(L)$), is not discrete, and $\| \phi_{p+1} - \phi_p \| \to 0$, then ϕ_p converges to a compact, connected component of $\mathscr{S}(L)$(resp. $\mathscr{M}(L)$) but not necessarily to a point. The original claim by DLR that ϕ_p converges to a point does not seem to hold under the conditions they stated. But we emphasize that the convergence of the EM sequence ϕ_p is not as important as the convergence of $L(\phi_p)$ to desired locations on the log-likelihood surface, an issue largely resolved in the present article.

Acknowledgments. The author wishes to thank Wing-Hung Wong and a referee for helpful comments.

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.

BOYLES, R. A. (1983). On the convergence of the EM algorithm. *J. Roy. Statist. Soc. B* **44**. To appear.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser.* **39** 1–38.

HABERMAN, S. J. (1974). Loglinear models for frequency tables derived by indirect observation: maximum likelihood equations. *Ann. Statist.* **2** 911–924.

HABERMAN, S. J. (1977). Product models for frequency tables involving indirect observation. *Ann. Statist.* **5** 1124–1147.

HARTLEY, H. O. and HOCKING, R. R. (1971). The analysis of incomplete data. *Biometrics* **27** 783–808.

HASSELBLAD, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *J. Amer. Statist. Assoc.* **64** 1459–1471.

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

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

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: theory and applications. *Proc. 6th Berkeley Symposium on Math. Statist. and Probab.* **1** 697–715.

ORTEGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic, New York.

OSTROWSKI, A. M. (1966). *Solution of Equations and Systems of Equations.* 2nd Edition. Academic, New York.

POLAK, E. (1971). *Computational Methods in Optimization.* Academic, New York.

REDNER, R. A. and WALKER, H. F. (1982). Mixture densities, maximum likelihood, and the EM algorithm. Preprint.

RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76.

SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1** 49–58.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. B* **38** 290–295.

TURNBULL, B. W. and MITCHELL, T. J. (1981). Nonparametric estimation of the time to onset for specific diseases in survival/sacrifice experiments. *Biometrics. To appear.*

VARDI, Y. (1982). Nonparametric estimation in renewal processes. *Ann. Statistics* **10** 772–785.

WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* **5** 329–350.

ZANGWILL, W. I. (1969). *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 WEST DAYTON ST.
MADISON, WISCONSIN 53706