Statistica Sinica **15**(2005), 831-840

PARAMETER CONVERGENCE FOR EM AND MM ALGORITHMS

Florin Vaida

University of California at San Diego

Abstract: It is well known that the likelihood sequence of the EM algorithm is non-decreasing and convergent (Dempster, Laird and Rubin (1977)), and that the limit points of the EM algorithm are stationary points of the likelihood (Wu (1982)), but the issue of the convergence of the EM sequence itself has not been completely settled. In this paper we close this gap and show that under general, simple, verifiable conditions, any EM sequence is convergent. In pathological cases we show that the sequence is cycling in the limit among a finite number of stationary points with equal likelihood. The results apply equally to the optimization transfer class of algorithms (MM algorithm) of Lange, Hunter, and Yang (2000). Two different EM algorithms constructed on the same dataset illustrate the convergence and the cyclic behavior.

Key words and phrases: EM, MM algorithm.

1. Introduction

This paper contains new results concerning the convergence of the EM algorithm. The EM algorithm was brought into the limelight by Dempster, Laird and Rubin (1977) as a general iterative method of computing the maximum likelihood estimator by maximizing a simpler likelihood on an augmented data space. However, the problem of the convergence of the algorithm has not been satisfactory resolved. Wu (1983), the main theoretical contribution in this area, showed that the limit points of the EM algorithm are stationary points of the likelihood, and that when the likelihood is unimodal, any EM sequence is convergent. Boyles (1983) has a number of results along similar lines. These results still allow the possibility of a non-convergent EM sequence when the likelihood is not unimodal. More importantly, the EM algorithm is useful when the likelihood is hard to obtain directly; for these cases, the unimodality of the likelihood is very difficult to verify. Here we give simple, general, verifiable conditions for convergence: our main result (Theorem 3) is that any EM sequence is convergent, if the maximizer at the M-step is unique. This condition is almost always satisfied in practice (otherwise the particular EM data augmentation scheme would

have limited usefulness), and it is verifiable by direct examination of the E-step function.

In an interesting recent development, Lange, Hunter and Yang (2000) proposed a new class of algorithms, called optimization transfer algorithms, or the MM algorithm. This is a generalization of the EM algorithm which does not involve missing data. The MM algorithm has already proven useful in computations for various statistical applications (Lange, Hunter and Yang (2000) and Becker, Yang and Lange (1997)). The convergence results proven for EM extend immediately to the MM algorithm.

After setting the regularity conditions (notably, that the stationary points of the likelihood are isolated) in the next section, we establish two results regarding the convergence of the EM algorithm. The first is Theorem 2, which states that any EM sequence either converges to a stationary point, or to a finite cycle of stationary limit points. The second is the convergence theorem mentioned above. The two results are illustrated on an example with inference for multivariate normal data with missing observations: two different data augmentation schemes lead to EM algorithms with different convergence properties, one convergent and the other cyclical. Section 5 extends the convergence results to the MM algorithm. A brief discussion closes the paper.

2. Preliminaries

Let y denote the observed data, with distribution $p(y|\theta)$ and log-likelihood $l_y(\theta) = \log p(y|\theta)$, where θ is a point in the parameter space $\Omega \subset R^d$.

The EM algorithm (EM) is best applied in problems with a natural missing random information structure, such as partially missing observations, censored data or random effects, where the direct maximization of the log-likelihood is typically a difficult task. The EM iteratively maximizes a function based on the augmented data Y_{aug}. The augmented data can be written as $Y_{\text{aug}} = (y, Y_{\text{mis}})$, where Y_{mis} is the missing data; more generally, Y_{aug} is such that y is a random function of Y_{aug} and θ. The algorithm starts with $\theta^{(0)} \in \Omega$; after t steps, $\theta^{(t+1)}$ is defined as follows.

E-step (Expectation): Calculate $Q(\theta|\theta^{(t)}) = E[\log p(Y_{\text{aug}}|\theta)|y, \theta^{(t)}]$ for all $\theta \in \Omega$; $Q(\theta|\theta^{(t)})$ is the expected log-likelihood of the augmented data, conditional on the observed data and current parameter $\theta^{(t)}$.

M-step (Maximization): Put $\theta^{(t+1)} = M(\theta^{(t)})$ where, for any θ, $M(\theta)$ is the global maximizer of $Q(\cdot|\theta)$: $Q(M(\theta)|\theta) \geq Q(\theta'|\theta)$, for all $\theta' \in \Omega$.

We assume that the maximizer at the M-step, $M(\cdot)$, called the EM transition function, is well defined, i.e., either the global maximizer of $Q(\cdot|\theta)$ is unique, or there is some deterministic procedure by which $M(\theta)$ is chosen among several

global maxima. Let the score function be $s_y(\theta) = \partial l_y(\theta)/\partial\theta$. For $\theta, \theta' \in \Omega$, let $H(\theta'|\theta) = E[\log p(Y_{\text{aug}}|y, \theta') \mid y, \theta]$, and note that

$$l_y(\theta') = Q(\theta'|\theta) - H(\theta'|\theta). \tag{1}$$

The EM algorithm is completely defined by the likelihood l_y, together with the data augmentation scheme, Y_{aug}; alternatively, EM is characterized by the transition function $M(\cdot)$. An EM sequence $\{\theta^{(t)}\}$ is given by its starting value $\theta^{(0)}$ and by the iteration $\theta^{(t+1)} = M(\theta^{(t)})$. We denote the set of stationary points of the log-likelihood $l_y(\cdot)$ by $\mathcal{S}_l$, $\mathcal{S}_l \subset \Omega$, i.e., $\theta^* \in \mathcal{S}_l$ iff $s_y(\theta^*) = 0$.

For a function f of two variables, let $D^{10}f$ denote the first partial derivative with respect to the first argument. If $f(x) = x$, then x is a fixed point for the function f. The point θ is a limit point, or subconvergence point for a sequence $\{\theta_k\}$ if there is a subsequence $\{\theta_{k_j}\}$ that converges to θ. An element $\theta^* \in \Omega$ is called a subconvergence point for EM if there exists an EM sequence $\{\theta^{(t)}\}$ for which θ^* is a subconvergence point. If $\{\theta^{(t)}\}$ converges to θ^*, then the latter is called a convergence point for EM. The EM algorithm is called convergent if any EM sequence converges. Note that we do not require for a convergent EM algorithm that all sequences converge to the same point for all starting values, but merely that each EM sequence is convergent.

Regularity conditions

The following conditions will be assumed to hold throughout.

R1: Ω is an open set in R^d.

R2: $l_y(\cdot)$ is differentiable, with continuous derivative $s_y(\cdot)$.

R3: The level set $\Omega_\theta = \{\theta' \in \Omega : l_y(\theta') \geq l_y(\theta)\}$ is compact in R^d (i.e., closed and bounded.)

R4: The missing data distribution $p(Y_{\text{aug}}|y, \theta)$ has the same support for all $\theta \in \Omega$.

R5: $Q(\theta'|\theta)$ is continuous in both θ' and θ, and differentiable in θ'.

R6: All the stationary points in $\mathcal{S}_l$ are isolated.

The conditions R1−R3 refer to the likelihood, and are needed for the asymptotic maximum likelihood theory to hold; the compactness condition R3 ensures that the local maxima of $l_y(\cdot)$ do not occur on the boundary of Ω. R1−R3 are similar to equations (5)−(8) in Wu (1983). R4 and R5 are linked to the data augmentation procedure. In particular, R4 implies that $H(\theta|\theta) \geq H(\theta'|\theta)$, for all $\theta, \theta' \in \Omega$ (see e.g., Lehmann (1983, p.409)), and therefore $D^{10}H(\theta|\theta) = 0$ for all $\theta \in \Omega$ (the existence of $D^{10}H$ is ensured by R2, R5 and (1)). Under conditions R1−R5, for all $\theta \in \Omega$, the function $Q(\cdot|\theta)$ admits a point of maximum, i.e., $M(\theta)$ exists. The condition R6 effectively rules out the case where the likelihood is maximized on a "ridge". Boyles (1983) shows an example of a generalized EM which does not converge, with a maximum likelihood on a ridge. The condition

R6 is certainly satisfied when $\mathcal{S}_l$ is finite, i.e., when the score equation $s_y(\theta) = 0$ has a finite number of solutions.

The following is a simplifying assumption needed in the sequel.

C1: The function $M(\cdot)$ is continuous at the stationary points in $\mathcal{S}_l$.

This continuity assumption is hard to check directly, but it can be replaced by the following easily verifiable, stronger condition.

C2: For all $\theta \in \mathcal{S}_l$ there exists a unique global maximum of $Q(\cdot|\theta)$.

Their connection is shown by the following lemma:

Lemma 1. *For a given $\theta \in \Omega$, if $Q(\cdot|\theta)$ has a unique global maximum, then $M(\cdot)$ is continuous at θ. In particular if* C2 *holds, then* C1 *is satisfied.*

3. The Main Results

We start by summarizing the extant results regarding the limiting behavior of the EM algorithm. Dempster, Laird and Rubin (1977) and Wu (1983) established the following properties (see also the monograph of McLachlan and Krishnan (1996)).

(i) Monotonicity of EM: for all $\theta \in \Omega$, $l_y(M(\theta)) \geq l_y(\theta)$.
(ii) If $l_y(M(\theta)) = l_y(\theta)$, then θ is a stationary point of l_y and a maximizer of $Q(\cdot|\theta)$.
(iii) If θ is a fixed point of $M(\cdot)$, then θ is a convergence point of EM, and a stationary point of the likelihood.

The reverse of (iii) is not true in general, as it is possible to have stationary points that are not fixed points of $M(\cdot)$, that is, points from which the algorithm can "escape" — see the second EM in Section 4. Statement (iii) also asserts that a fixed point of $M(\cdot)$ is a point of convergence for EM. The reverse is true under the continuity condition C1: if $\theta^{(t)}$ converges to θ^*, then $\theta^* \in \mathcal{S}_l, M(\theta^{(t)})$ converges to $M(\theta^*)$, hence $M(\theta^*) = \theta^*$. The following is a key result.

Theorem 1.(Wu (1983)) *If θ^* is a limit point of the EM sequence $\{\theta^{(t)}\}$, then* (a) *θ^* is a stationary point of the likelihood, and* (b) *the sequence $\{l_y(\theta^{(t)})\}$ is nondecreasing and converges to $l_y(\theta^*)$.*

In other words, an EM sequence is either convergent, or it has a set of limit points with identical likelihood value. The restriction on the set of limit points is a strong one, and it will be exploited later in Theorem 2. Note that the limit point θ^* may be any kind of stationary point, i.e., either a local maximum, a saddle-point, or a local minimum of the likelihood. The theorem does not go so far as to show that $\theta^{(t)} \to \theta^*$, which is what we are after. Clearly, if no two

stationary points have the same likelihood, then the EM algorithm is convergent. However, this condition depends on the likelihood function, and is hard to verify in practice. A more gratifying result would replace it with some conditions that are easier to check, such as condition C2.

After the next lemma, we give a simple and explicit characterization of the limiting behavior of the EM algorithm in Theorem 2, followed by the main convergence result in Theorem 3.

Lemma 2. *Assuming* R6, then *for any real value l^* there is at most a finite number of stationary points θ^* such that $l(\theta^*) = l^*$.*

A finite set of distinct points $\theta_1^*, \ldots, \theta_m^*$ of Ω is called a cycle of length $m \geq 2$ for the transition function $M(\cdot)$ if $M(\theta_i^*) = \theta_{i+1}^*$, for $i = 1, \ldots, m-1$, and $M(\theta_m^*) = \theta_1^*$.

Theorem 2. *Let $\{\theta^{(t)}\}$ be an EM sequence, and assume that C1 holds. Then either $\theta^{(t)}$ converges to a stationary point θ^* with $M(\theta^*) = \theta^*$, or there exists a finite set $\mathcal{C} = \{\theta_1^*, \ldots, \theta_m^*\}$, such that*
(i) *$\theta_1^*, \ldots, \theta_m^*$ are stationary points ($\mathcal{C} \subset \mathcal{S}_l$) with the same likelihood value;*
(ii) *$M(\theta_1^*) = \theta_2^*$, $M(\theta_2^*) = \theta_3^*, \ldots,$ $M(\theta_m^*) = \theta_1^*$;*
(iii) *The parallel subsequences $\{\theta^{(mt+i)}; t \geq 1\}$ satisfy $\theta^{(mt+i)} \to \theta_i^*$ when $t \to \infty$, for $i = 1, \ldots, m$.*

It is clear from Theorem 2 that the limit points of an EM algorithm coincide with the fixed points and the cycles of the transition function $M(\cdot)$. The following known result (Wu (1983, Theorem 6)) can be alternatively be obtained as a direct consequence of Theorem 2.

Corollary 1. *Under* C1, *an EM sequence $\{\theta^{(t)}\}$ is convergent iff $||\theta^{(t+1)} - \theta^{(t)}|| \to 0$, when $t \to \infty$.*

In practice, the cyclical behavior asserted by Theorem 2 is rarely encountered, but not impossible; see Section 4 for an illustrative example. However, this theorem is instrumental in obtaining the main result regarding the convergence of the EM algorithm.

Theorem 3. *Assume that* C2 *holds. Then for any starting value $\theta^{(0)}$ of the EM sequence $\{\theta^{(t)}\}$, $\theta^{(t)} \to \theta^*$ when $t \to \infty$, for some stationary point $\theta^* \in \mathcal{S}_l$. Moreover, $M(\theta^*) = \theta^*$, and if $\theta^{(t)} \neq \theta^*$ for all t, the sequence of likelihood values $\{l_y(\theta^{(t)})\}$ is strictly increasing to $l_y(\theta^*)$.*

Theorem 3 offers a general, verifiable condition for convergence. This condition is almost always satisfied in practice (otherwise the particular EM data augmentation scheme would have limited usefulness), and can be established, for example, by checking the sign of the derivative of the function to be maximized at the M-step, as illustrated in the next example.

4. An Illustrative Example

Consider twelve independent observations $z_i = (x_i, y_i)$ from a bivariate normal distribution with mean zero and variance matrix $\sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, with σ^2, ρ unknown:

$$\begin{array}{c|cccccccccccc} x_i & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 3 & 3 & -3 & -3 \\ y_i & 1 & -1 & 1 & -1 & . & . & . & . & . & . & . & . \end{array},$$

where the dots correspond to missing values. The purpose is to compute the MLE for σ^2, ρ. This is a modification of the example of Murray (1977).

Direct calculation of the observed log-likelihood yields $l = -8\log\sigma^2 - 18/\sigma^2 - 2\log(1-\rho^2) - 4/[\sigma^2(1-\rho^2)]$. The log-likelihood is symmetric in ρ, and admits three stationary points: maxima at $\rho = \pm 1/\sqrt{3}, \sigma^2 = 3$, and a saddlepoint at $\rho = 0, \sigma^2 = 11/4$.

A convergent EM. The MLE can be computed using the EM algorithm, treating $y_5 \dots y_{12}$ as missing data. The E-step is

$$Q(\rho, \sigma^2 | \tilde{\rho}, \tilde{\sigma}^2) = -12\log\sigma^2 - 6\log(1-\rho^2) - \frac{4(A - 9\tilde{\rho}\rho)}{\sigma^2(1-\rho^2)},$$

where $A = 11/2 + 9\tilde{\rho}^2/2 + (1-\tilde{\rho}^2)\tilde{\sigma}^2$. Straightforward calculations at the M-step show that Q has a unique maximum at $\rho = 9\tilde{\rho}/A, \sigma^2 = A/3$. From Theorem 3 it follows that the EM algorithm will converge from any starting point $\rho_0 \in (-1, 1), \sigma_0^2 > 0$.

It is interesting to note, however, that different EM sequences may converge to different stationary points. As in the example of Murray (1977), for $\rho_0 < 0$ the sequence converges to $\rho = -1/\sqrt{3}, \sigma^2 = 3$, for $\rho_0 > 0$ to $\rho = 1/\sqrt{3}, \sigma^2 = 3$, and for $\rho_0 = 0$, to the saddlepoint $\rho = 0, \sigma^2 = 11/4$. Figure 1 shows the EM paths for different starting values.

The cyclic EM. Consider now an alternative EM algorithm, in which only $y_5 \dots y_8$ are taken as missing data. The E-step is now

$$Q_A(\rho, \sigma^2 | \tilde{\rho}, \tilde{\sigma}^2) = -10\log\sigma^2 - 4\log(1-\rho^2) - \frac{4 + 2(1-\tilde{\rho}^2)\tilde{\sigma}^2}{\sigma^2(1-\rho^2)} - \frac{18}{\sigma^2}.$$

This function has two symmetric maxima, for $\sigma^2 = 3, \rho = \pm\sqrt{2/3 - \tilde{\sigma}^2(1-\tilde{\rho}^2)/6}$; more specifically, σ^2 converges in one step to the MLE $\sigma^2 = 3$, after which ρ is updated by the equation $\rho = \pm\sqrt{1/6 + \tilde{\rho}^2/2}$. (In addition, Q_A has the saddlepoint $\rho = 0, \sigma^2 = 11/5 + (1-\tilde{\rho}^2)\tilde{\sigma}^2$.)

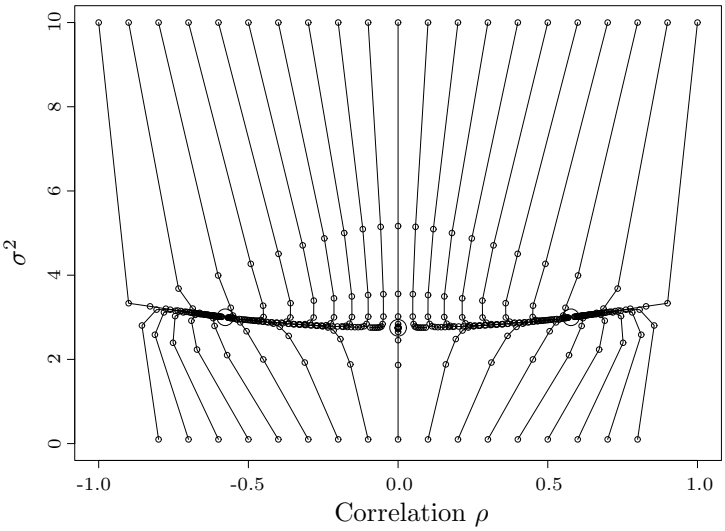

Figure 1. The standard EM algorithm: EM paths for several starting values (points, connected by lines; the starting points have $\sigma^2 = 0.1$ or $\sigma^2 = 10$). The two large side circles are the global maxima of the likelihood, and the middle large circle is the saddlepoint.

When defining the next value of ρ in the EM sequence, we have a choice between the two solutions. By convention, let us take ρ at the M-step as $\sqrt{1/6 + \tilde{\rho}^2/2}$ if $\tilde{\rho}$ is negative, and as $-\sqrt{1/6 + \tilde{\rho}^2/2}$ if $\tilde{\rho}$ is positive. The resulting EM sequence alternates between positive and negative values of ρ, and in the limit it cycles between the two likelihood maxima.

5. Convergence of the MM Algorithm

In a refreshing recent development, Lange, Hunter and Yang (2000) proposed the MM algorithm as a generalization of EM, which does not involve missing data. Briefly, if for the EM algorithm we put $U(\theta'|\theta) = Q(\theta'|\theta) - H(\theta'|\theta)$, at (1) we have that

$$l_y(M(\theta)) \geq U(\theta'|\theta) \quad \text{and } l_y(\theta) = U(\theta|\theta), \tag{2}$$

since $H(\theta'|\theta) \leq H(\theta|\theta)$ for all θ'. The MM algorithm is defined in general by assuming that we can find a function $U(\theta'|\theta)$ which satisfies (2). Usually l_y is the likelihood function, but the algorithm does not require it. Similarly to EM, the transition function $M(\cdot)$ is the global maximizer of $U(\cdot|\theta)$: $U(M(\theta)|\theta) \geq U(\theta'|\theta)$, for all $\theta \in \Omega$. The MM algorithm inherits the monotonicity and the convergence properties of the EM algorithm Lange et al. (2000). Since the transition function $M(\cdot)$ is defined by U, the surrogate function U in the MM algorithm plays the role

of Q in EM. Similarly, the convergence results in Theorems 2 and 3 apply to the MM algorithm, with minor modifications. Specifically, R4 is not necessary, and R5 and C2 refer to U instead of Q. With this obvious proviso, the proofs follow unchanged. In particular, if $U(\cdot|\theta)$ has a unique maximum at the stationary points $\theta \in \mathcal{S}_l$, then the MM algorithm is convergent.

6. Discussion

In this paper we established the convergence of the EM algorithm, and of its younger relative, the MM algorithm. Regarding the points of convergence of the EM algorithm, some misconceptions linger on. It is widely held that if EM converges, the limit is a local maximum or, in unfortunate cases, a saddle-point of the likelihood (see the previous section). However, Arslan, Constable and Kent (1993) show a simple example where the algorithm may converge to a local minimum. Their example also shatters another myth of the EM folklore, that the algorithm does not "overshoot", i.e., that it converges to the "closest" stationary point. The cycling EM presented here is one more instance of bizarre behavior.

How can cycling be avoided? A simple way is to impose additional conditions on the parameters: in the example in Section 4, if we enforce $\rho > 0$, the Q function has a unique maximum on the restricted parameter space, therefore the algorithm is convergent. However, in general, cycling is good and should be promoted, not avoided: a single cycling EM sequence identifies several likelihood maxima at once!

An alternative mathematical framework for dealing with multiple maxima at the M-step is to think of the EM transition function M as a set function taking the set Θ_t of all current parameter values to Θ_{t+1}, the set of all maxima of $Q(\cdot|\theta)$ over all $\theta \in \Theta_t$: $\Theta_{t+1} = M(\Theta_t)$. The starting set has a single element, $\Theta_0 = \{\theta_0\}$. Using the results in Section 3 we can show that Θ_t converges to a set Θ^* of stationary points of l_y of equal likelihood value. The cycling sequences appear through certain choices of points θ_t in Θ_t at step t of EM.

Pathological cases aside, in most cases of practical interest the EM and the MM algorithms are convergent. By checking on the unicity of the maxima at the M-step, the practitioners can establish this convergence, rather than merely hope for it.

Acknowledgements

This paper is dedicated to Professor Xiao-Li Meng, who introduced me to the EM algorithm, offered generous advice and support while writing this paper,

and suggested the extension to the MM algorithm. The helpful comments of an associate editor and one anonymous reviewer are also gratefully acknowledged.

Appendix

Proof of Lemma 1. To prove the continuity of $M(\cdot)$, we assume that $\theta_k \to \theta$ and show that $M(\theta_k) \to M(\theta)$ when $k \to \infty$. Since for any θ' and k, $Q(M(\theta_k)|\theta_k) \geq Q(\theta'|\theta_k)$, we have

$$\liminf_{k\to\infty} Q(M(\theta_k)|\theta_k) \geq Q(\theta'|\theta). \tag{3}$$

Because $\{M(\theta_k)\}$ is a bounded sequence (since $\{M(\theta_k)\} \subset \Omega_{\theta_0}$, and Ω_{θ_0} is bounded), it admits a limit point. Take $\{M(\theta_{k_n})\}_{n\geq 0}$ to be a convergent subsequence, with limit $\tilde{M}$. Then, from (3) and R5, $Q(\tilde{M}|\theta) \geq Q(\theta'|\theta)$ for all θ' in Ω, and from C2, we have $\tilde{M} = M(\theta)$. Therefore any limit point of $M(\theta_k)$ is $M(\theta)$, so $M(\theta_k) \to M(\theta)$.

Proof of Lemma 2. If $l_y(\theta^*) = l^*$ for some $\theta^* \in \mathcal{S}_l$, then $\mathcal{S}_l(l^*) \subset \Omega_{\theta^*}$, and therefore $\mathcal{S}_l(l^*)$ is bounded. Assume $\mathcal{S}_l(l^*)$ is infinite. Then there exists a sequence θ_k^* of distinct values of $\mathcal{S}_l(l^*)$ that converges to some value $\theta_0 \in \Omega_{\theta^*}$. Since the score function is continuous, $0 = s_y(\theta_k^*) \to s_y(\theta_0)$, hence $\theta_0 \in \mathcal{S}_l$; but now θ_0 is not an isolated stationary point, contradicting R6.

Proof of Theorem 2. If the sequence $\{\theta^{(t)}\}$ is convergent to θ^*, then the limit is a stationary point (from Wu's Theorem), and a fixed point of $M(\cdot)$, since $M(\theta^*) = M(\lim_t \theta^{(t)}) = \lim_t M(\theta^{(t)}) = \lim_t \theta^{(t+1)} = \theta^*$.

Assume now that the sequence $\{\theta^{(t)}\}$ is not convergent, and has the set of limit points $\mathcal{C}$. By Wu's Theorem, $\mathcal{C} \subset \mathcal{S}_l$ and, by the corollary to Lemma 2, this set is finite, hence (i).

To prove (ii), we note that since $M(\cdot)$ is continuous, the sequence $M(\theta^{(t)})$ has the limit set $\{M(\theta_i^*) : i = 1, \ldots, m\}$. But $M(\theta^{(t)}) = \theta^{(t+1)}$, so the limit set must coincide with $\mathcal{C}$. It follows that $M(\cdot)$ permutes the elements of $\mathcal{C}$; it remains to show that this permutation is a cycle, i.e., it does not have cycles of length smaller than m. Put $\sigma(i)$ for the corresponding permutation of the indexes $1, \ldots, m$: $M(\theta_i^*) = \theta_{\sigma(i)}^*$.

The positive integers may be partitioned into infinite sets $T_1, \ldots, T_m$ with $\theta^{(t)} \to \theta_i^*$ for $t \to \infty, t \in T_i$. Take $V_1, \ldots, V_m$ as disjoint neighborhoods of the points $\theta_1^*, \ldots, \theta_m^*$ respectively, and find t_0 large enough so that $t \in T_i$ whenever $\theta^{(t)} \in V_i$ and $t \geq t_0$. For each i in $\{1 \ldots m\}$ we have $M(\theta^{(t)}) \to M(\theta_i^*)$ when $t \to \infty$, $t \in T_i$, therefore $t + 1 \in T_{\sigma(i)}$. Then for some $t_i > 0, M(\theta^{(t)}) \in V_{\sigma(i)}$ whenever $t > t_i, t \in T_i$.

Take now $\tilde{t} = \max(t_0, \ldots, t_m)$. Then for $t \geq \tilde{t}$,

$$t \in T_i \qquad \text{implies} \qquad t + 1 \in T_{\sigma(i)} \qquad i = 1, \ldots, m. \tag{4}$$

For a cycle C of σ, put $T = \cup_{i \in C} T_i$. For $t \geq \tilde{t},\ t \in T$ implies $t+1 \in T$, therefore the algorithm 'stays' in T. It follows that the cycle has to contain all the set indexes $1, \ldots, m$, so $\sigma(\cdot)$ is a cyclic permutation, and $\mathcal{C}$ is a cycle of $M(\cdot)$. In particular, $\sigma^m(i) = i$ for $i = 1, \ldots, m$.

For (iii), notice now that from (4), when $t \geq \tilde{t}$, $t \in T_i$ implies that $t+m \in T_i$. Therefore the m parallel cyclic subsequences converge to each of the m points in $\mathcal{C}$.

Proof of Corollary 1. The direct implication is immediate. Assume now that the EM sequence is not convergent. Then the distances between consecutive EM iterations converge to the distances between the different limit points in the cycle $\mathcal{C}$ of Theorem 2, and they cannot go to 0.

Proof of Theorem 3. From Lemma 1 (i), $M(\cdot)$ is continuous at the points of $\mathcal{S}_l$, and the conditions of Theorem 2 are satisfied. Assume that θ_1^* and θ_2^* are distinct limit points of an EM sequence as in Theorem 2, with $M(\theta_1^*) = \theta_2^*$. Then $l_y(M(\theta_1^*)) = l_y(\theta_1^*)$, and from (ii) at the top of Section 3, θ_1^* is also a maximizer of $Q(\cdot|\theta_1^*)$, as is θ_2^*, contradicting the assumption of the theorem. It follows that the sequence converges to a stationary point θ^*, with $M(\theta^*) = \theta^*$. Assume now that for some t we have a non-increasing likelihood, i.e., $l_y(\theta^{(t)}) = l_y(\theta^{(t+1)})$. Then using the result (ii) at the top of Section 3 and C2 again, we get that $\theta^{(t+1)} = \theta^{(t)} = \theta^*$, which completes the proof of the theorem.

References

Arslan, O., Constable, P. D. L. and Kent, J. T. (1993). Domains of convergence for the EM algorithm: a cautionary tale in a location estimation problem, *Statist. Comput.* **3**, 103-108.

Becker, M. P., Yang, I. and Lange, K. (1997). EM algorithms without missing data. *Statist. Methods in Medical Res.* **6**, 38-54.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **45**, 47-50.

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. B* **39**, 1-38.

Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion), *J. Comput. Graph. Statist.* **9**, 1-20.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

McLachlan, G. J. and Krishnan, T. (1996). *The EM Algorithm and Extensions*. Wiley, New York.

Murray, G. D. (1977). Discussion of Dempster et al. maximum likelihood from incomplete data via the EM. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, UCSD School of Medicine, La Jolla, CA 92093-0645, U.S.A.

E-mail: vaida@ucsd.edu

(Received February 2003; accepted May 2004)