Neural Processing Letters **6:** 69–76, 1997.

Comparative Analysis on Convergence Rates of The EM Algorithm and Its Two Modifications for Gaussian Mixtures*

LEI XU
Dept of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, P.R. China
Email: lxu@cs.cuhk.edu.hk

Key words: convergence rate, EM algorithm and variants, Gaussian mixture, maximum likelihood, momentum

Abstract. For Gaussian mixture, a comparative analysis has been made on the convergence rate by the Expectation-Maximization (EM) algorithm and its two types of modifications. One is a variant of the EM algorithm (denoted by VEM) which uses the old value of mean vectors instead of the latest updated one in the current updating of the covariance matrices. The other is obtained by adding a momentum term in the EM updating equation, called the Momentum EM algorithm (MEM). Their up-bound convergence rates have been obtained, including an extension and a modification of those given in Xu & Jordan (1996). It has been shown that the EM algorithm and VEM are equivalent in their local convergence and rates, and that the MEM can speed up the convergence of the EM algorithm if a suitable amount of momentum is added. Moreover, a theoretical guide on how to add momentum is proposed, and a possible approach for further speeding up the convergence is suggested.

1. Introduction

The 'Expectation-Maximization' (EM) algorithm is a general methodology for maximum likelihood (ML) or maximum a posteriori (MAP) estimation [2]. It has been substantially investigated in the literature of both neural computing and statistics, e.g., see the reference list in [7]. The starting point for many of these studies is the fact that EM is generally a first order or linearly convergent algorithm, as can readily be seen by considering EM as a mapping $\Theta^{(k+1)} = M(\Theta^{(k)})$, with fixed point $\Theta^* = M(\Theta^*)$. Results on the convergence rate of EM are obtained by calculation of the information matrices for the missing data and the observed data [1,3].

In Xu and Jordan [7], the mathematical connection between the EM algorithm and gradient-based approach has been set up. That is, the EM step in parameter space is obtained from the gradient via a projection matrix P in an explicit expression. Based on it, a comparative study has been made on the advantages and

* Supported by HK RGC Earmarked Grant CUHK250/94E, CUHK484/95E and Ho Sin-Hang Education Endowment Fund HSH 95/02.

disadvantages of that EM and other algorithms for Gaussian mixtures. It has also been shown that the condition number associated with EM is smaller than the condition number associated with gradient ascent, providing a general guarantee of the dominance of the EM algorithm over the gradient algorithm. Moreover, in cases in which the mixture components are well separated, they showed that the condition number for EM approximately converges to one, corresponding to a local superlinear convergence rate. Furthermore, in a latest result by Ma, Xu and Jordan [2], it has been further shown that the asymptotic convergence rate of the EM algorithm for Gaussian mixtures around its true solution Θ^* is $o(e^{0.5-\varepsilon}(\Theta^*))$, where $\varepsilon > 0$ is an arbitrarily small number, $o(x)$ means that it is a higher order infinitesimal as $x \to 0$, and $e(\Theta^*)$ is a measure of the average overlap of Gaussians in the mixture. In other words, the large sample convergence rate for the EM algorithm tends to be asymptotically superlinear when $e(\Theta^*)$ tends to zero.

In fact, a slightly modified variant of the EM algorithm (denoted by VEM) has been studied in Xu and Jordan [7]. It differs from the standard EM algorithm (denoted by EM) in using the old value of mean vectors instead of the latest updated one in the current updating of the covariance matrices. This immediately rises a serious concern – whether EM shares the convergence properties of VEM. In this paper, it has been shown that the EM algorithm and VEM are actually equivalent in local convergence rate. Moreover, a quite number of publications [4, 5, 6] suggested to modifying the EM algorithm by adding a momentum term in the EM updating equation, called the Momentum EM algorithm (MEM). This paper has justified that the MEM can speed up the convergence of the EM algorithm if a suitable amount of momentum is added. Moreover, a theoretical guide on how to add momentum is proposed, and a possible approach for further speeding up the convergence is suggested.

2. Equivalence of EM and VEM on Convergence and Its Rate

For a Gaussian mixture given by :

$$P(x|\Theta) = \sum_{j=1}^{K} \alpha_j P(x|m_j, \Sigma_j), \tag{1}$$

and

$$P(x|m_j, \Sigma_j) = \frac{e^{-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)}}{\sqrt{(2\pi)^n}|\Sigma_j|^{\frac{1}{2}}},$$

where $\alpha_j \geq 0$, $\sum_{j=1}^{K} \alpha_j = 1$ and n is the dimension of x.

Given K and given n i.i.d. samples $\{x^{(t)}\}_1^N$, an estimate $\hat{\Theta}$ that maximizes the log likelihood

$$l(\Theta) = \log \prod_{t=1}^{N} P(x^{(t)}|\Theta) = \sum_{t=1}^{N} \log P(x^{(t)}|\Theta) \tag{2}$$

can be obtained via the standard EM algorithm (Dempster, Laird and Rubin, 1977) which is an iterative algorithm for maximum likelihood learning as follows:

$$\begin{aligned}
\alpha_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t)}{N}, \\
m_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t) x^{(t)}}{\sum_{t=1}^N h_j^{(k)}(t)} \\
\Sigma_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t)[x^{(t)} - \hat{m}][x^{(t)} - \hat{m}]^T}{\sum_{t=1}^N h_j^{(k)}(t)}, \quad where \;\; \hat{m} = m_j^{(k+1)}
\end{aligned} \tag{3}$$

where the posterior probabilities $h_j^{(k)}$ are defined as follows:

$$h_j^{(k)}(t) = \frac{\alpha_j^{(k)} P\left(x^{(t)} | m_j^{(k)}, \Sigma_j^{(k)}\right)}{\sum_{i=1}^K \alpha_i^{(k)} P\left(x^{(t)} | m_i^{(k)}, \Sigma_i^{(k)}\right)}.$$

The VEM algorithm studied in [7] is quite similar to the above standard EM algorithm, except that in Equation (3) we use the old value of mean vector

$$\hat{m} = m_j^{(k)}$$

to replace the latest updated one

$$\hat{m} = m_j^{(k+1)}.$$

To get some deep insight on the standard EM and VEM, we re-examine the origin of the M step in the standard EM algorithm. It follows from Dempster, Laird & Rubin [1] that Equation (3) comes from $\max_{m_j, \Sigma_j} Q(\{m_j, \Sigma_j\}_1^K)$ sequentially in the order of firstly $\max_{m_j} Q(\{m_j, \Sigma_j^{(k)}\}_1^K)$ and then $\max_{\Sigma_j} Q(\{m_j^{(k+1)}, \Sigma_j\}_1^K)$, where $Q(\{m_j, \Sigma_j\}_1^K)$ is a re-weighted criterion given by

$$\begin{aligned}
&Q(\{m_j, \Sigma_j\}_1^K) = \\
&-\tfrac{1}{2} \textstyle\sum_{t=1}^N \sum_{j=1}^K h_j^{(k)}(t)[\log|\Sigma_j| + (x^{(t)} - m_j)^T \Sigma_j^{-1}(x^{(t)} - m_j)]
\end{aligned}$$

We also have another choice for $\max_{m_j, \Sigma_j} Q(\{m_j, \Sigma_j\}_1^K)$. We can first get $\Sigma_j^{(k+1)}$ by $\max_{\Sigma_j} Q(\{m_j^{(k)}, \Sigma_j\}_1^K)$ and then get $m_j^{(k+1)}$ by $\max_{m_j} Q(\{m_j, \Sigma_j^{(k+1)}\}_1^K)$, which outcomes exactly the above VEM. It is obvious that this way of maximization of Q still increases Q. Thus, both the standard EM and VEM guarantee to converge to a maximum likelihood solution. Also, both share the properties of general convergence and automatic satisfaction of constraints discussed in [7].

From Xu and Jordan [7], each iteration step of VEM can be obtained by pre-multiplying the gradient of the log likelihood by a positive definite matrix $P_{\mathcal{A}}$ as follows:

$$\mathcal{A}^{(k+1)} - \mathcal{A}^{(k)} = P_{\mathcal{A}}^{(k)} \frac{\partial l}{\partial \mathcal{A}}|_{\Theta = \Theta^{(k)}},$$

$$\begin{aligned} m_j^{(k+1)} - m_j^{(k)} &= P_{m_j}^{(k)} \frac{\partial l}{\partial m_j}|_{\Theta=\Theta^{(k)}}, \\ \text{vec}[\Sigma_j^{(k+1)}] - \text{vec}[\Sigma_j^{(k)}] &= P_{\Sigma_j}^{(k)} \frac{\partial l}{\partial \text{vec}[\Sigma_j]}|_{\Theta=\Theta^{(k)}} \end{aligned} \tag{4}$$

with $\mathcal{A} = [\alpha_1, \cdots, \alpha_K]^T$, $\Theta = [m_1^T, \cdots, m_K^T, \text{vec}[\Sigma_1]^T, \cdots, \text{vec}[\Sigma_K]^T, \mathcal{A}^T]^T$ and

$$P_{\mathcal{A}}^{(k)} = \frac{1}{N}\{\text{diag}[\alpha_1^{(k)}, \cdots, \alpha_K^{(k)}] - \mathcal{A}^{(k)}(\mathcal{A}^{(k)})^T\},$$

$$P_{m_j}^{(k)} = \frac{\Sigma_j^{(k)}}{\sum_{t=1}^N h_j^{(k)}(t)}$$

$$P_{\Sigma_j}^{(k)} = \frac{2}{\sum_{t=1}^N h_j^{(k)}(t)} \Sigma_j^{(k)} \otimes \Sigma_j^{(k)}$$

are positive definite matrices with probability one under the constraints $\sum_{j=1}^K \alpha_j^{(k)} = 1$, where '$\otimes$' denotes the Kronecker product. Letting $P(\Theta) = \text{diag}[P_{m_1}, \cdots, P_{m_K}, P_{\Sigma_1}, \cdots, P_{\Sigma_K}, P_{\mathcal{A}}]$, we can have a single equation:

$$\Theta^{(k+1)} = \Theta^{(k)} + P(\Theta^{(k)}) \frac{\partial l}{\partial \Theta}|_{\Theta=\Theta^{(k)}}, \tag{5}$$

That is, *the direction $\Theta^{(k+1)} - \Theta^{(k)}$ of each VEM iteration has a positive projection on the gradient of the likelihood function l.*

We can link Equation (2) to VEM by letting

$$\begin{aligned} \Sigma_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t)[x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T}{\sum_{t=1}^N h_j^{(k)}(t)} + \\ &+ [m_j^{(k)} - m_j^{(k+1)}][m_j^{(k)} - m_j^{(k+1)}]^T \end{aligned}$$

Following a similar procedure in getting Equation (4) by Xu and Jordan [7], from the above equation we can have the following Equation (6) in the place of Equation (4) for the standard EM algorithm

$$\begin{aligned} &\text{vec}[\Sigma_j^{(k+1)}] - \text{vec}[\Sigma_j^{(k)}] = P_{\Sigma_j}^{(k)} \frac{\partial l}{\partial \text{vec}[\Sigma_j]}|_{\Theta=\Theta^{(k)}} + \\ &+ [m_j^{(k)} - m_j^{(k+1)}] \otimes [m_j^{(k)} - m_j^{(k+1)}] \end{aligned} \tag{6}$$

Strictly speaking, *the search direction $\Theta^{(k+1)} - \Theta^{(k)}$ of each iteration by the EM algorithm is no longer satisfies Equation (5) but is modified by an extra term resulted from* $[m_j^{(k)} - m_j^{(k+1)}] \otimes [m_j^{(k)} - m_j^{(k+1)}]$. From Equation (4) it is actually an 2nd order force by gradient $\frac{\partial l}{\partial m_j}|_{\Theta=\Theta^{(k)}}$.

Given a local area around a converged point Θ^*, for both the VEM and the EM we have $\|\Theta^{(k)} - \Theta^*\|^2 \to 0$ is an infinitesimal as $k \to \infty$. While the square

norm $\|[m_j^{(k)} - m_j^{(k+1)}] \otimes [m_j^{(k)} - m_j^{(k+1)}]\|^2 = \|m_j^{(k)} - m_j^{(k+1)}\|^4$ is a higher (≥ 2) order infinitesimal of $\|\Theta^{(k)} - \Theta^*\|^2$ because $\|m_j^{(k)} - m_j^{(k+1)}\|$ is a same order infinitesimal of $\|m_j^{(k)} - m_j^*\|$ and $m_j^{(k)} - m_j^*$ is a part of $\Theta^{(k)} - \Theta^*$ with $0 \leq \|m_j^{(k)} - m_j^*\|^4 / \|\Theta^{(k)} - \Theta^*\|^4 \leq 1$. Therefore, the role of the 2nd term in Equation (6) can be neglected, and Equation (6) and Equation (4) can be regarded the same in a local area around a converged point Θ^*. Moreover, an iterative algorithm is said to have a convergence rate of order $q \geq 1$ if $\|\Theta^{(k+1)} - \Theta^*\| / \|\Theta^{(k)} - \Theta^*\|^q \leq r + o(\|\Theta^{(k)} - \Theta^*\|)$ for k sufficiently large. The effect of the 2nd term in Equation (6) will be in $o(\|\Theta^{(k)} - \Theta^*\|)$ which is independent of q. Therefore, we have **Theorem 1** *For Gaussian mixture, the EM and the VEM share the same local convergence rate r, which is up-bounded by*

$$r \leq \| E^T(I + P(\Theta^*)H(\Theta^*)) \| \leq \sqrt{1 + \lambda_M^2 - 2\lambda_m} = r^u; \tag{7}$$

$$\lambda_M = \|E^T P(\Theta^*)H(\Theta^*)\| = \sqrt{\lambda_M[E^T P(\Theta^*)H^2(\Theta^*)P(\Theta^*)E]},$$

$$\lambda_m = \sqrt{\lambda_m[-E^T(P(\Theta^*)H(\Theta^*) + H(\Theta^*)P(\Theta^*))E]}, \tag{8}$$

where $H(\Theta^*)$ is the Hessian of the likelihood function l at Θ^*. $\|A\|$, $\lambda_M[A]$ and $\lambda_m[A]$ is the norm, the largest and smallest eigenvalues of the matrix A respectively, and $E = [e_1, \cdots, e_{m-1}]$ consists of unit orthogonal basis vectors that span the subspace $\mathcal{D} = \{\Theta : \sum_{j=1}^K \alpha_j = 0\}$ for the case that $\text{vec}[\Sigma_j]^T = [\sigma_{11}^{(j)}, \cdots, \sigma_{nn}^{(j)}]$ *with* $\Sigma_j = diag[\sigma_{11}^{(j)}, \cdots, \sigma_{nn}^{(j)}]$ *or* $\mathcal{D} = \{\Theta : \sum_{j=1}^K \alpha_j = 0,\ \sigma_{pq}^{(j)} = \sigma_{qp}^{(j)},\ for\ all\ q, p, j\}$ *for a general covariance matrix Σ_j.*

Equation (7) is the last inequality on p137 of Xu and Jordan [7], and Equation (8) is modification of Equation (5.5) there, which is approximately correct only but not strictly true. The detail proof of Theorem 1 will be given in Section 4.

3. The Momentum EM Algorithm (MEM) and Convergence Rate

For an iterative algorithm with a current incremental $\Delta\Theta = \Theta^{(k+1)} - \Theta^{(k)}$, we can always modify the obtained $\Theta^{(k+1)}$ into $(1 - \eta)\Theta^{(k)} + \eta\Theta^{(k+1)}$ or

$$\Theta^{(k+1)} = \Theta^{(k)} + \eta\Delta\Theta, \quad \eta > 0 \tag{9}$$

Usually, this is called the momentum approach. We can easily get the Momentum EM and the Momentum VEM by using Equation (9) to modify the incremental $\Delta\Theta = \Theta^{(k+1)} - \Theta^{(k)}$ given by Equation (3). The Momentum EM has been considered to be able to speed up the convergence [4,5,6] with an appropriate η that is usually chosen heuristically. Here, we analyze why the MEM can speed up the convergence and how to chose η.

From Equation (5), the Momentum VEM can be actually represented as

$$\Theta^{(k+1)} = \Theta^{(k)} + \eta P(\Theta^{(k)}) \frac{\partial l}{\partial \Theta}|_{\Theta=\Theta^{(k)}} \tag{10}$$

Moreover, similar to the previous section we can show that the Momentum EM and the Momentum VEM are equivalent in their local convergence and rate. So, our following results should hold for both of them.

Theorem 2 *For Gaussian mixture, we have*

(1) The Momentum EM and the Momentum VEM share the same local convergence rate r, which is up-bounded by

$$r \leq \| E^T(I + \eta P(\Theta^*)H(\Theta^*)) \| \leq \sqrt{1 + \eta^2\lambda_M^2 - 2\eta\lambda_m} = r_\eta^u. \tag{11}$$

where E, λ_M, λ_m are the same as that given in Theorem 1.

(2) $r_\eta^{u*} = \min_\eta r_\eta^u = \sqrt{1 - (\lambda_M/\lambda_m)^2}$ at $\eta^* = \lambda_m/\lambda_M^2$.

(3) When $r^u < 1$ or equivalently $\lambda_M < \sqrt{2\lambda_m}$, we further have $r_\eta^{u*} < r^u$, $\eta^* = \lambda_m/\lambda_M^2 > 0.5$ where r^u given by Equation (8)

This theorem justifies that the MEM can speed up convergence because $r_\eta^{u*} < r^u$ when $\lambda_M < \sqrt{2\lambda_m}$, which holds usually due to the function of P matrix, as shown in Xu and Jordan [7]. Moreover, it also provides a theoretical guide to select the momentum amount via η, which should be at least $\eta^* > 0.5$, and reaches its best at $\eta^* = \lambda_m/\lambda_M^2$.

Since it is possible to get P matrix by Equation (4) and the Hessian H by the formulae given in Ma, Xu and Jordan [2] (although they need quite complicated computation), we suggest that some further speed-up of the EM can be made along the following two directions:

(1). To find a fast way to estimate λ_m and λ_M^2 adaptively, such that we can adaptively change η to approximate η^*.

(2). To find a fast way to estimate Q as an approximation of the pseudo-inverse matrix of PH, such that we can further have an Quasi-Newton algorithm by

$$\Theta^{(k+1)} = \Theta^{(k)} + \eta Q \Delta\Theta, \tag{12}$$

where the incremental $\Delta\Theta = \Theta^{(k+1)} - \Theta^{(k)}$ given by the EM algorithm.

4. The Proof of Theorems

Since the local convergence rates of the EM and VEM are the same, and also the EM and VEM are the special case $\eta = 1$ of the Momentum EM and the Momentum VEM respectively, in the following we only need to prove the case for the Momentum VEM.

(1) During the iteration of the Momentum VEM algorithm, we have always $\Theta \in \mathcal{R}$ with $\mathcal{R} = \{\Theta : \sum_{j=1}^K \alpha_j = 1\}$ for the case that $\text{vec}[\Sigma_j]^T = [\sigma_{11}^{(j)}, \cdots, \sigma_{nn}^{(j)}]$ with $\Sigma_j = diag[\sigma_{11}^{(j)}, \cdots, \sigma_{nn}^{(j)}]$ or $\mathcal{R} = \{\Theta : \sum_{j=1}^K \alpha_j = 1,\ \sigma_{pq}^{(j)} = \sigma_{qp}^{(j)},\ for\, all\, q, p, j\}$ for a general covariance matrix Σ_j. R can be translated into a subspace D, with its spanning unit orthogonal basis vectors $e_1, \cdots, e_{m-1}$ that forms E as given in Theorem 2, by a constant shift Θ_0. For each Θ, we have $\Theta' = \Theta - \Theta_0 \in \mathcal{D}$, and

we further let the coordinates of Θ' under the bases $e_1, \cdots, e_m$ to be denoted by Θ_c, we have

$$\Theta - \Theta_0 = E\Theta_c, \qquad or \qquad \Theta = E\Theta_c + \Theta_0. \tag{13}$$

Multiplying its both sides by E^T, it follows from $E^T E = I$ that

$$E^T\Theta = E^T E\Theta_c + E^T\Theta_0, \qquad or \qquad \Theta_c = E^T\Theta - E^T\Theta_0. \tag{14}$$

Putting it into Equation (13), we have

$$\Theta' = \Theta - \Theta_0 = E^T E(\Theta - \Theta_0) = EE^T\Theta', \quad for \quad \Theta' \in \mathcal{D}, \tag{15}$$
$$\|\Theta'\|^2 = \Theta'^T\Theta' = \Theta'^T E^T E\Theta' = \|E^T\Theta'\|^2, \quad for \quad \Theta' \in \mathcal{D}, \tag{16}$$

From Equation (10), by Taylor expansion of $P(\Theta^{(k)})\frac{\partial l}{\partial \Theta}|_{\Theta=\Theta^{(k)}}$ around the converged point Θ^* and noticing that $P(\Theta^{(k)})\frac{\partial l}{\partial \Theta}|_{\Theta=\Theta^*} = 0$, we have

$$\begin{aligned}
\Theta^{(k+1)} &= \Theta^{(k)} + \eta P(\Theta^*)H(\Theta^*)(\Theta^{(k)} - \Theta^*), \\
\Theta^{(k+1)} - \Theta^* &= \Theta^{(k)} - \Theta^* + \eta P(\Theta^*)H(\Theta^*)(\Theta^{(k)} - \Theta^*) \\
&= (I + \eta P(\Theta^*)H(\Theta^*))(\Theta^{(k)} - \Theta^*), \\
E^T(\Theta^{(k+1)} - \Theta^*) &= E^T(I + \eta P(\Theta^*)H(\Theta^*))(\Theta^{(k)} - \Theta^*), \\
\|\Theta^{(k+1)} - \Theta^*\| &= \|E^T(\Theta^{(k+1)} - \Theta^*)\| \\
&\leq \|E^T(I + \eta P(\Theta^*)H(\Theta^*))\|\|\Theta^{(k)} - \Theta^*\|,
\end{aligned} \tag{17}$$

since $\Theta^{(k+1)} - \Theta^* \in D$ and $\Theta^{(k)} - \Theta^* \in D$ and thus Equation (15) and Equation (16) hold. It further follows that the local convergence rate of the EM algorithm is up-bounded by

$$\begin{aligned}
&r \leq \|E^T(I + \eta P(\Theta^*)H(\Theta^*))\| = \\
&\sqrt{\max_{\|x\|=1} x^T\{E^T(I + \eta H(\Theta^*)P(\Theta^*))(I + \eta P(\Theta^*)H(\Theta^*))E\}x} \leq \\
&\sqrt{1 + \eta^2 \max_{\|x\|=1} x^T E^T H(\Theta^*)P(\Theta^*)P(\Theta^*)H(\Theta^*)Ex - T_3} \\
&T_3 = \sqrt{\eta \min_{\|x\|=1} x^T E^T[-H(\Theta^*)P(\Theta^*) - P(\Theta^*)H(\Theta^*)]Ex}
\end{aligned}$$

Since $-H$ is positive definite and P is positive definite on R and thus that $E^T H(\Theta^*)P(\Theta^*)P(\Theta^*)H(\Theta^*)E$ and $-E^T[H(\Theta^*)P(\Theta^*)) + P(\Theta^*)H(\Theta^*)]E$ are positive definite matrices. Moreover, we have

$\max_{\|x\|=1} x^T E^T H(\Theta^*)P(\Theta^*)P(\Theta^*)H(\Theta^*)Ex = \lambda_M$

$\min_{\|x\|=1} x^T E^T[-H(\Theta^*)P(\Theta^*) - P(\Theta^*)H(\Theta^*)]E\}x = \lambda_m$

with λ_M, λ_m are defined in Equation (8). Therefore, we have

$r \leq \|E^T(I + \eta P(\Theta^*)H(\Theta^*))\| \leq \sqrt{1 + \eta^2\lambda_M^2 - 2\eta\lambda_m} = r_\eta^u$.

(2) It is obvious to get $r_\eta^{u*} = \min_\eta r_\eta^u = \sqrt{1 - (\lambda_M/\lambda_m)^2}$ at $\eta^* = \lambda_m/\lambda_M^2$.

(3) It is easy to get from (2).

5. Conclusion

The EM algorithm and VEM are equivalent in its local convergence rate. The momentum term can speed up the convergence but needs to be chosen at least with $\eta > 0.5$, which can be further improved if we can adaptively estimate the optimal η^* given in Theorem 3.

References

1. A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. of Royal Statistical Society, B39, pp. 1–38, 1977.
2. J. Ma, L. Xu and M.I. Jordan, "Asymptotic Convergence Rate of the EM Algorithm for Gaussian Mixtures", preprint, submitted to a journal, 1996.
3. X.L. Meng, "On the rate of convergence of the ECM algorithm", Ann. Statist., Vol. 22, pp. 326–339, 1994.
4. I. Meilijson, "A fast improvement to the EM algorithm on its own terms", J. of Royal Statistical Society, B51, pp. 127–138, 1989.
5. B.C. Peters H.F. and Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters foe a mixture of normal distributions", SIAM J. Applied Mathematics, Vol. 35, pp. 362–378, 1978.
6. R.A. Redner and H.F. Walker, "Mixture densities, maximum likelihood, and the EM algorithm", SIAM Review, Vol. 26, pp. 195–239, 1984.
7. L. Xu and M.I. Jordan, "On convergence properties of the EM algorithm for gaussian mixtures", Neural Computation, Vol. 8, No. 1, Jan, 1996, pp. 129–151, 1996.