DOES GEM CONVERGE TO MLE UNDER WU'S CONDITIONS? : A COUNTER EXAMPLE

Nomakuchi, Kentaro
Department of Mathematics, Faculty of Science, Kochi University

https://doi.org/10.5109/12592

出版情報：Bulletin of informatics and cybernetics. 37, pp.65-71, 2005-12. Research Association of Statistical Sciences
バージョン：
権利関係：

DOES GEM CONVERGE TO MLE UNDER WU'S CONDITIONS?
– A COUNTER EXAMPLE –

by

Kentaro Nomakuchi

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.37

FUKUOKA, JAPAN
2005

Bulletin of Informatics and Cybernetics, Vol. 37, 2005

DOES GEM CONVERGE TO MLE UNDER WU'S CONDITIONS? – A COUNTER EXAMPLE –

By

Kentaro Nomakuchi*

Abstract

EM-algorithm users believe that the conditions of Wu (1983) assure the convergence of GEM sequence, but this paper gives a brief counter example which satisfies Wu's conditions but does not converge to MLE or any optimal solutions. It also gives a correction of his proof for the convergence of EM sequence.

1. Introduction

We consider the following statistical model:

$$f(x|\theta),\ \theta \in \Theta,$$

where Θ is a parameter space and f is a density (or probability) function with a parameter $\theta \in \Theta$. When a part of x, say y, is only observed in a sampling, y is called incomplete data and x complete. We assume that y is only available. In such a situation with incomplete data, EM-algorithm or its extension, GEM-algorithm, are frequently used to calculate the maximal likelihood estimate (MLE) of θ (See McLachlan and Krishnan (1997)). Wu (1983) gives conditions for EM and GEM to converge to optimal solutions including MLE. In the EM users, it is believed that his result is fundamental. We also agree that his result for EM is good, but we must say that the one for GEM is not so. His result is a translation of the convergence theorem given by Zangwill (1969), which is called a global convergence theorem by Wu himself. However, we show that he fails the translation for GEM. We will give a counter example that satisfies Wu's conditions but does not converge to MLE, which is a unique optimal solution in our counter example.

For the paper to be self-contained, we give basic terms and results in Section 2 and 3. Section 2 gives a brief review of EM(GEM) algorithm and Section 3 introduces the global convergence theorem and Wu's theorem for GEM. We will give the counter example in Section 4 and a correction of his proof for the convergence of EM sequence in Section 5. Throughout this paper, we use same symbols, x or y, to denote random variables and those realizations.

2. Brief review of EM(GEM) algorithm

Let us denote the observable part of x by x_{obs}. If we have incomplete data y as an observable part of the data x, the set of x to be consistent with y is given by

$$\chi(y) = \{x : x_{\text{obs}} = y\}.$$

* Department of Mathematics, Faculty of Science, Kochi University, Kochi 780-8520, Japan

Given y, the likelihood function of θ (or the marginal density of y) is

$$g(y|\theta) = \int_{x \in \chi(y)} f(x|\theta) dx.$$

To have the MLE concerning this, we must maximize it with respect to $\theta \in \Theta$, which is generally difficult because the objective function $g(y|\theta)$ is of integral form. The EM procedure is then given as follows; The conditional distribution of x given y is defined by

$$f(x|y,\theta) = \frac{f(x|\theta)}{g(y|\theta)}, \quad x \in \chi(y).$$

By using the above, we have a representation of the log-likelihood as follows

$$\log(g(y|\theta)) = \log(f(x|\theta)) - \log(f(x|y,\theta)).$$

Now we assume that a tentative parameter $\eta \in \Theta$ is obtained. Taking expectations of the both sides of the above with respect to $f(x|y,\eta)$, we have

$$\begin{aligned} L(\theta) &= \log(g(y|\theta)) \\ &= \int_{x \in \chi(y)} \log(g(y|\theta)) f(x|y,\eta) dx \\ &= Q(\theta|\eta) - H(\theta|\eta), \end{aligned}$$

where

$$\begin{aligned} Q(\theta|\eta) &= \int_{x \in \chi(y)} \log(f(x|\theta)) f(x|y,\eta) dx, \\ H(\theta|\eta) &= \int_{x \in \chi(y)} \log(f(x|y,\theta)) f(x|y,\eta) dx. \end{aligned}$$

The EM-algorithm, neglecting the function $H(\theta|\eta)$ and using only $Q(\theta|\eta)$, is defined as follows;

Step 1. Select an initial point $\theta_0 \in \Theta$.

Step 2. Repeat the following sub-steps until the θ_n converges $(n = 0, 1, 2, ...)$;

E(xpectation)-step	:	Calculate the $Q(\theta|\theta_n)$,
M(aximization)-step	:	Find a solution $\theta = \theta_{n+1}$ to maximize $Q(\theta|\theta_n)$.

If it is difficult to find the best solution in the M-step, we can substitute a following solution θ_{n+1} for it:

$$Q(\theta_{n+1}|\theta_n) \geq Q(\theta_n|\theta_n),$$

which is called a Generalized EM(GEM)-algorithm. If we set

$$S(\theta) = \{\eta \in \Theta : Q(\eta|\theta) \geq Q(\theta|\theta)\},$$

the definition of GEM becomes $\theta_{n+1} \in S(\theta_n)$. The GEM satisfies the following theorem.

THEOREM 2.1. *(Dempster et al. (1977))*
A GEM instance, $\theta_n, n = 0, 1, 2, ...$, increases the likelihood monotonically, that is,

$$L(\theta_{n+1}) \geq L(\theta_n), \quad n = 0, 1, 2, ...$$

PROOF. The proof follows easily by using the inequality

$$\begin{aligned} H(\theta_{n+1}|\theta_n) &= \int \log(f(x|y, \theta_{n+1}))f(x|y, \theta_n)dx \\ &\leq \int \log(f(x|y, \theta_n))f(x|y, \theta_n)dx \\ &= H(\theta_n|\theta_n). \end{aligned}$$

3. Wu's convergence theorem for GEM

At first, let us introduce Zangwill's convergence theorem. We assume that there is a one-to-many function $S : \Theta \to 2^{\Theta}$ and an algorithm is constructed by $\theta_0 \in \Theta,\ \theta_{n+1} \in S(\theta_n), n = 1, 2, ...$ The function S is said to be closed at $\theta \in \Theta$ if the following condition is satisfied:

$$\theta_n \in \Theta,\ \theta_n \longrightarrow \theta,\ \eta_n \in S(\theta_n),\ \eta_n \longrightarrow \eta \implies \eta \in S(\theta).$$

If S is closed at any $\theta \in A$, S is said to be closed in A. Zangwill (1969) gives the following useful convergence criterion for iterative algorithms.

THEOREM 3.1. *(Zangwill (1969))*
Assume that an algorithm $\theta_0 \in \Theta,\ \theta_{n+1} \in S(\theta_n), n = 1, 2, ...$ satisfies the following conditions;

(i) $\forall n \quad \theta_n \in D$, where D is compact,

(ii) S is closed in Γ^c,

(iii) $\exists \alpha$ on Θ : a continuous real-valued function satisfying

(a) $\theta \notin \Gamma \implies \alpha(\eta) > \alpha(\theta) \quad for \quad \forall \eta \in S(\theta)$,
(b) $\theta \in \Gamma \implies \alpha(\eta) \geq \alpha(\theta) \quad for \quad \forall \eta \in S(\theta)$,

where Γ is a subset of Θ. Then all limit points of θ_n are in Γ, and $\alpha(\theta_n)$ converges monotonically to $\alpha(\theta)$ for some $\theta \in \Gamma$.

PROOF. Let θ be a limit point of θ_n, that is, there exists a subsequence $\theta_{n'}$ such that

$$\theta_{n'} \longrightarrow \theta.$$

Then we define

$$\eta_{n'} = \theta_{n'+1} \in S(\theta_{n'}).$$

For this $\eta_{n'}$, there is a further converging subsequence

$$\eta_{n''} = \theta_{n''+1} \longrightarrow \eta.$$

Hence

$$\theta_{n''} \longrightarrow \theta, \quad \eta_{n''} \longrightarrow \eta, \quad \eta_{n''} \in S(\theta_{n''}).$$

Now, we assume $\theta \notin \Gamma$. Since S is closed at θ, we have $\eta \in S(\theta)$, hence $\alpha(\eta) > \alpha(\theta)$. On the other hand,

$$\alpha(\theta_{n''}) \leq \alpha(\eta_{n''}) \leq \alpha(\theta_{(n+1)''}) \longrightarrow \alpha(\theta).$$

This means $\alpha(\eta) = \alpha(\theta)$, which is a contradiction. Hence $\theta \in \Gamma$.

The function α is usually taken to be an objective function. This theorem is very useful as a tool to show the convergence of iterative algorithms.

In the convergence problem for GEM, the set Γ in Theorem 3.1 would become

$$\Gamma = \{\theta \in \Theta : Q(\theta|\theta) \geq Q(\eta|\theta) \text{ for } \forall \eta \in \Theta\}.$$

Let us say the elements θ in Γ to be GEM-optimal. Taking α to be the likelihood L, Wu (1983) insists that Theorem 3.1 becomes the following theorem.

THEOREM 3.2. *(Theorem 1 in Wu (1983))*
Let $\theta_n, n = 1, 2, ...$ be a GEM instance. Suppose that

(i) $\forall n \quad \theta_n \in D$, where D is compact,

(ii) S is closed in Γ^c,

(iii) $\theta_n \notin \Gamma \implies L(\theta_{n+1}) > L(\theta_n)$.

Then all limit points of θ_n are in Γ, and $\alpha(\theta_n)$ converges monotonically to $\alpha(\theta)$ for some $\theta \in \Gamma$.

Since the condition (b) of (iii) in Theorem 3.1 follows from Theorem 2.1 for GEM, Wu insists that Theorem 3.1 can be translated to Theorem 3.2. As Wu points, if a level set $\{\theta \in \Theta : L(\theta) \geq a\}$ is compact then the condition (i) of Theorem 3.2 automatically follows, and if $Q(\theta, \eta)$ is continuous about (θ, η) then S is closed everywhere. In order that Theorem 3.2 holds true, we have only to check the condition (iii) to show the convergence of GEM in such situations. But, the condition (iii) would be usually a requirement in constructing an algorithm practically. Hence, Theorem 3.2 becomes a very useful tool. Is it true? We will show that it is not true in the next section by constructing a GEM satisfying the conditions of Theorem 3.2 but not converging to MLE or other GEM-optimal solutions.

REMARK. (1) It is essentially difficult to make a direct translation of Theorem 3.1 for GEM because of $\theta \in S(\theta)$, in which the condition (a) of (iii) is never satisfied. If we adopt

$$S'(\theta) = \{\eta \in \Theta : L(\eta) > L(\theta)\}$$

or

$$S''(\theta) = S(\theta) - \{\theta\}$$

instead of $S(\theta)$, the condition (a) of (iii) could be saved, but the condition (ii) would be violated.

(2) If you have a concrete GEM sequence, we recommend to make $S(\theta)$ adjusted to your problem and to apply just Theorem 3.1 to it.

4. A counterexample

Let us consider the following data.

$$\begin{aligned} x &\sim N(\theta, 1), \\ y &= \begin{cases} 1, & x \in [-1, 1] \\ 0, & otherwise, \end{cases} \end{aligned}$$

where y is incomplete data such that x is in [-1,1], or not. The probability function of y is given by

$$g(y|\theta) = \begin{cases} \int_{-1}^{1} \phi(x-\theta)dx, & y = 1 \\ 1 - \int_{-1}^{1} \phi(x-\theta)dx, & y = 0, \end{cases}$$

where ϕ is the standard normal density function. Now, we assume that $y = 1$ is observed, that is, $x \in [-1, 1]$. Then the likelihood for $y = 1$ is given by

$$L(\theta) = \int_{-1}^{1} \phi(x-\theta)dx.$$

This likelihood function is continuous, symmetric and unimodal at $\theta = 0$. Hence, the MLE θ_{MLE} is 0 and the level sets $\{\theta : L(\theta) \geq a\}$ are compact if $a > 0$. To apply the EM approach, we have also

$$\begin{aligned} f(x|y,\theta) &= \frac{1}{L(\theta)}\phi(x-\theta), \quad x \in \chi(y=1) = [-1,1], \\ \log(f(x|\theta)) &= -\frac{1}{2}\theta^2 + \theta x + c, \\ Q(\theta|\theta_n) &= -\frac{1}{2}\theta^2 + \theta E(x|y=1,\theta_n) + c' \\ &= -\frac{1}{2}(\theta - E(x|y=1,\theta_n))^2 + c'', \end{aligned}$$

where c, c' and c'' are appropriate constants. The EM sequence is given by iterating $\theta_{n+1} = E(x|y = 1, \theta_n)$. It is easy to check that if $\theta_0 > 0$ then θ_n converges decreasingly to $\theta_{MLE} = 0$, and if $\theta_0 < 0$ then it does increasingly to $\theta_{MLE} = 0$. Hence this EM instance converges to the MLE $\theta_{MLE} = 0$. On the other hand, the one-to-many function $S(\theta)$ for GEM is given by the following intervals

$$S(\theta) = \begin{cases} [\theta, 2E(x|y=1,\theta) - \theta], & \theta \leq 0 \\ [2E(x|y=1,\theta) - \theta, \theta], & \theta > 0. \end{cases}$$

Since $Q(\theta|\eta)$ is clearly continuous in both θ and η, S becomes closed everywhere. Hence we can consider a following instance of GEM;

$$\theta_0 = 1, \quad \theta_n = \mathrm{Max}\{E(x|y,\theta_{n-1}), \theta_{n-1} - \frac{1}{2^{n+1}}\}, \ n = 1, 2, 3, ...$$

Let $\Gamma = \{0\} = \{\theta_{MLE}\}$. Then, all conditions of Theorem 3.2 are satisfied because of $L(\theta_{n+1}) > L(\theta_n)$, $n = 0, 1, ...$ Since $\theta_n > 0$ is non-increasing and $\theta_{n-1} - \theta_n \leq \frac{1}{2^{n+1}}$, we

have

$$\begin{aligned}\theta_n &= \theta_0 - \sum_{i=1}^{n}(\theta_{i-1} - \theta_i) \\ &\geq 1 - \sum_{i=1}^{n}\frac{1}{2^{i+1}} \\ &> \frac{1}{2}.\end{aligned}$$

Hence, our GEM instance does not converge to a unique GEM optimal solution θ_{MLE} in Γ.

5. Convergence of EM

We think that the result about the convergence of EM by Wu (1983) is sound. He, however, uses Theorem 3.2 to show it, and his proof is a little funny because of using differentiation of $Q(\eta|\theta)$ without assumption. We correct his proof and utilize just Theorem 3.1. To do that, let us define

$$\begin{aligned}S(\theta) &= \{\bar{\eta} \in \Theta \ : \ Q(\bar{\eta}|\theta) = \max_{\eta\in\Theta} Q(\eta|\theta)\}, \\ \Gamma &= \{\theta^* \in \Theta \ : \ Q(\theta^*|\theta^*) = \max_{\theta\in\Theta} Q(\theta|\theta^*)\}.\end{aligned}$$

Let us say the elements $\theta \in \Gamma$ to be EM-optimal. The following theorem assures that EM sequence converges essentially to an EM-optimal solution.

THEOREM 5.1. *Let $\theta_n, n = 0, 1, 2, ...$ be an EM instance such that*

$$\theta_{n+1} \in S(\theta_n), n = 0, 1, 2, ...$$

Suppose that

(1) $\Theta_0 = \{\theta \in \Theta \ : \ L(\theta) \geq L(\theta_0)\}$ is compact,

(2) $Q(\eta|\theta)$ is continuous in both η and θ.

Then all limit points of θ_n are in Γ, and $L(\theta_n)$ is non-decreasing and converges to $L(\theta)$ for some θ in Γ.

PROOF. Setting $\alpha(\theta) = L(\theta)$, we show that all the conditions of Theorem 3.1 are satisfied. The condition (i) is clear from (1), and the (b) of condition (iii) is also clear from Theorem 2.1. To check the condition (ii), we assume that

$$\bar{\theta} \in \Theta, \ \theta_n \to \bar{\theta}, \ \eta_n \to \bar{\eta}, \ \eta_n \in S(\theta_n).$$

It follows from the definition of S that

$$Q(\eta_n|\theta_n) = \max_{\eta\in\Theta} Q(\eta|\theta_n),$$

that is, for any $\eta \in \Theta$,

$$Q(\eta_n|\theta_n) \ \geq \ Q(\eta|\theta_n).$$

Hence, we have from the condition (2)

$$Q(\bar{\eta}|\bar{\theta}) \ \geq \ Q(\eta|\bar{\theta}).$$

Since $\eta \in \Theta$ is arbitrary, we have $\bar{\eta} \in S(\bar{\theta})$, which implies that S is closed at any point $\bar{\theta} \in \Theta$. To check the condition (a) of (iii), we set $\theta \in \Gamma^c$. Then, we have

$$\max_{\eta \in \Theta} Q(\eta|\theta) > Q(\theta|\theta),$$

from which it is easy to see that $L(\eta) > L(\theta)$ for any $\eta \in S(\theta)$.

References

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm(with discussion), *Journal of Royal Statistical Society, Ser. B*, **39**, 1–38.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley.

Wu, C.F.J. (1983). On the Convergence Properties of the EM Algorithm, *Annals of Statistics*, **11**, 95–103.

Zangwill, W.I. (1969). *Nonlinear Programming : A Unified Approach*, Prentice-Hall.

Received October 10, 2003
Revised October 4, 2004