The Convergence Property of EM Algorithm

Yu Chen

A thesis presented for a seminar of EM

Mathematical Department
Technical University of Munich
Germany
17.12.2020

Abstract

The thesis reviews Jeff Wu's paper ON THE CONVERGENCE PROPERTIES OF THE EM ALGORITHM[4] which studied: (1) whether EM algorithm finds a local maximum or just a stationary value for the target likelihood over incomplete data? (2) whether the parameter sequence generated from the EM iteration process finally converge?. Jeff Wu dedicated to correcting the error appeared in the paper MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA THE EM ALGORITHM (WITH DISCUSSION)[2] which Jeff Wu's paper is based on via presenting 7 theorems and one corollary. In this thesis, we focus on studying the relationship among the theorems.

1 Introduction

Expectation Maximization (EM) consisting of a E-step and M-step is an iterative algorithm that tries to maximize the likelihood over incomplete data. This algorithm is popular in not only statistics but also optimization, machine learning and computer vision. Due to its wide applications, it has a range of forms describing the E-step and M-step. Dempster, Laird and Rubin (abbreviated DLR) [2] introduced a general form for the EM algorithm and analysed some properties for it, whereas their proof of convergence of an EM sequence is not totally correct. Therefore, Jeff Wu corrects the error in his paper[4]. In this thesis, we inherit the general EM form from the DLR paper and review Jeff Wu's correct proof of EM convergence properties.

Specifically, this thesis answers two questions regarding the convergence of the EM. (1) Does EM finally reach a global maximum or local maximum or stationary value for the likelihood? (2) Whether the sequence generated from the EM iteration converges to a limit? The key to answer the first question is the Global Convergence Theorem[5]. Based on this theorem Jeff Wu drew[4] three theorems answering the first question. In addition, the other four theorems are introduced to obtain the answer for the second question. The EM sequence can converge to the unique maximum when the likelihood is unimodal and a differentiability condition is satisfied[4].

2 The Generalized EM Algorithm

We have 2 sample spaces $\mathbb{X}$ and $\mathbb{Y}$ with corresponding p.d.f. $f(\boldsymbol{x}|\phi)$ and $g(\boldsymbol{y}|\phi)$ satisfying the following relationship:

$$g(\boldsymbol{y}|\phi) = \int_{\{\boldsymbol{x}:\boldsymbol{y}(\boldsymbol{x})=\boldsymbol{y}\}} f(\boldsymbol{x}|\phi)d\boldsymbol{x}$$

where $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ is the observed incomplete data in $\mathbb{Y}$ and $\phi \in \Omega$ the parameter space. The relationship between the two sample space is a many-to-one mapping from $\mathbb{X}$ to $\mathbb{Y}$ which we will talk about in detail in section 3. However, another way to express this relationship is:

$$g(\boldsymbol{y}|\phi) = \int f(\boldsymbol{x}, \boldsymbol{y}|\phi)d\boldsymbol{x}$$

In order to present the generalized EM, we first draw the likelihood of the incomplete observation $\boldsymbol{y}$. To this end, we introduce the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ and ϕ (note: $f(\boldsymbol{x}|\phi) = f(\boldsymbol{x}, \boldsymbol{y}|\phi)$):

$$k(\boldsymbol{x}|\boldsymbol{y}, \phi) = \frac{f(\boldsymbol{x}|\phi)}{g(\boldsymbol{y}|\phi)}$$

So,

$$g(\boldsymbol{y}|\phi) = \frac{f(\boldsymbol{x}|\phi)}{k(\boldsymbol{x}|\boldsymbol{y}, \phi)}$$

.

Then the log likelihood $L(\phi^{'})$ is:

$$\begin{aligned} L(\phi^{'}) &= log(g(\boldsymbol{y}|\phi^{'})) \\ &= E_{\boldsymbol{x}\sim k(\boldsymbol{x}|\boldsymbol{y},\phi)}[log(g(\boldsymbol{y}|\phi^{'}))] \\ &= E_{\boldsymbol{x}\sim k(\boldsymbol{x}|\boldsymbol{y},\phi)}[log(\frac{f(\boldsymbol{x}|\phi^{'})}{k(\boldsymbol{x}|\boldsymbol{y},\phi^{'})})] \\ &= E_{\boldsymbol{x}\sim k(\boldsymbol{x}|\boldsymbol{y},\phi)}[log(f(\boldsymbol{x}|\phi^{'})) - log(k(\boldsymbol{x}|\boldsymbol{y},\phi^{'}))] \\ &= E\{log(f(\boldsymbol{x}|\phi^{'}))|\boldsymbol{y},\phi\} - E\{log(k(\boldsymbol{x}|\boldsymbol{y},\phi^{'}))|\boldsymbol{y},\phi\} \\ &= Q(\phi^{'}|\phi) - H(\phi^{'}|\phi) \end{aligned} \tag{1}$$

Where we assume $Q(\phi^{'}|\phi)$ and $H(\phi^{'}|\phi)$ exist for all pairs of $(\phi^{'},\phi)$

Now, we are interested in maximizing the log likelihood $\max_{\phi^{'}} L(\phi^{'}) = Q(\phi^{'}|\phi) - H(\phi^{'}|\phi)$. The EM algorithm solve this using an iteration process:

$$\phi_p \rightarrow \phi_{p+1} \in M(\phi_p)$$

where M is a point-to-set map consisting of two steps:

- E-STEP Determine $Q(\phi|\phi_p)$ for current ϕ_p
- M-STEP $\phi_{p+1} = argmax_{\phi\in\Omega}Q(\phi|\phi_p)$

However, $Q(\phi|\phi_p)$ can be very complex and may not numerically feasible to maximize it, so we need a more general way to depict the EM algorithm[4]. Since we are only interested in the convergence property of EM algorithm, we do not care the specific methods we use for the M-step. We just need to guarantee the properties that the EM must satisfy in the Generalized EM (GEM).

Dempster, Laird and Rubin (1977) defined the **GEM** algorithm in their DLR paper [2] as an iterative scheme:

$$\phi_{p+1} \in M(\phi_p)$$

where $\phi^{'} \rightarrow M(\phi)$ is a point-to-set map such that:

$$Q(\phi^{'}|\phi) \geq Q(\phi|\phi) \ \forall \ \phi^{'} \in M(\phi) \tag{2}$$

So, we see that EM is special case of the GEM. Moreover, two properties of the GEM have been sumarized in DLR (Theorem 1 and Lemma 1):

$$H(\phi|\phi) \geq H(\phi^{'}|\phi) \ \forall \ \phi^{'} \in \Omega \tag{3}$$

and for any sequence $\{\phi_p\}$ from a GEM algorithm:

$$L(\phi_{p+1}) \ \geq \ L(\phi_p) \tag{4}$$

3 Global Maximum or Local Maximum or Stationary Values?

We have presented the GEM and its properties in the previous section, and it is easy to see from (2) that if $L(\phi_p)$ is bounded, then the sequence $L(\phi_p)$ converges monotonically to some limit L^*. Now, we are interested in whether L^* is the global maximum or local maximum or just stationary value of $L(\phi_p)$ over Ω. To do this, we make the following assumptions:

- 1) $\Omega \subseteq \mathbb{R}^r$

- 2) $\Omega_{\phi_0} = \{\phi \in \Omega : L(\phi) \geq L(\phi_0)\}$ is compact for any $L(\phi_0) > -\infty$
- 3) $L(\phi)$ is continuous in Ω and differentiable in the interior of Ω
- 4) $\{L(\phi_p)\}_{p\geq 0}$ is bounded above for any $\phi_0 \in \Omega$
- 5) ϕ_p is in the interior of Ω, $int(\Omega)$
- 6) ϕ_p converges to some $\phi^* \in int(\Omega)$ such that the Hessian matrices $\nabla^2 Q(\phi^*|\phi^*)$ and $\nabla^2 H(\phi^*|\phi^*)$ exist at the first ϕ^*, and $\nabla^2 Q(\phi^{'}|\phi)$ is continuous in $(\phi^{'}, \phi)$

where assumption 4) is a consequence of the previous three assumptions. Assumption 6) ensures that we have tools to analyze whether L^* is the global maximum or local maximum or just stationary value.

In the M-step of EM, we globally maximize $Q(\phi^{'}|\phi)$ for current ϕ, so based on assumption 6), we have $\nabla^2 Q(\phi^*|\phi^*)$ is non-positive definite (n.p.d.). According to Lemma 2 of DLR, we have $-\nabla^2 H(\phi^*|\phi^*)$ is non-negative definite (n.n.d.). Therefore, the Hessian matrix of the log-likelihood $\nabla^2 L(\phi^*) = \nabla^2 Q(\phi^*|\phi^*) - \nabla^2 H(\phi^*|\phi^*)$ may not be n.p.d. i.e. $L(\phi^*)$ may not be a local maximum. Murray[3] gave an example illustrating that the EM converges to a stationary value rather than a local maximum.

Now, to answer the global or local or stationary question, we introduce point-to-set map M on set X i.e. M maps from points of X to subsets of X. M is called closed at x^* if $x_k \in X$, $\lim_{k\to\infty} x_k = x^*$ and $\lim_{k\to\infty} y_k = y^*$, $y_k \in M(x_k)$ imply $y^* \in M(x^*)$. With this concept, we introduce the following theorem.

Global Convergence Theorem.

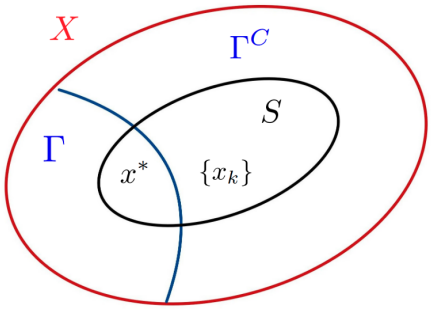

Figure 1: Global Convergence Theorem

- $\{x_k\}_{k=0}^{\infty}$ is generated by $x_{k+1} \in M(x_k)$ where M is p2s map on set X
- Solution set $\Gamma \subset X$

  - 1) $x_k \in S$ where compact set $S \subset X$

  - 2) M is closed over Γ^C

  Then all limit points of $\{x_k\}$ are in Γ

  - 3) function α is continuous on X such that

  a) if $x \in \Gamma^C$, then $\alpha(y) > \alpha(x)\ \forall\ y \in M(x)$;
  b) if $x \in \Gamma$, then $\alpha(y) \geq \alpha(x)\ \forall\ y \in M(x)$.
  Then $\alpha(x_k)$ converges monotonically to $\alpha(x^*)$ for some $x^* \in \Gamma$

PROOF see Appendix. Figure 1 helps understand the relationships among each concept in the theorem and illustrates the first consequence that the limit of the sequence $\{x_k\}$ finally fall into the solution set Γ.

Now, we set M as the point-to-set map in a GEM iteration and set the function α as the log-likelihood function L. Additionally, let the solution set Γ be one of the following:

- M: local maxima in the interior of Ω
- φ: stationary points in the interior of Ω

Then we get Theorem 1 as a special case of the Global Convergence Theorem.

Theorem 1

ϕ_p is a GEM sequence generated by $\phi_{p+1} \in M(\phi_p)$, and suppose:
1) M is a closed p2s map over φ^C (or M^C)
2) $L(\phi_{p+1}) > L(\phi_p)\ \forall \phi_p \notin \varphi$ (or M)
Then, all limit points of $\{\phi_p\}$ are stationary (or local maxima) of L, and $L(\phi_p)$ converges monotonically to $L^* = L(\phi^*)$

Note that if $Q(\psi|\phi)$ is continuous in both ψ and ϕ, then condition 1) in Theorem 1 satisfies[4]. In fact it is also sufficient to imply condition 2) in Theorem 1 such that we Theorem 2.

Theorem 2

If $Q(\psi|\phi)$ is continuous in ψ and ϕ, then all limit points of $\{\phi_p\}$ in an EM are stationary points of L, and $L(\phi_p)$ converges monotonically to $L^* = L(\phi^*)$ for some point ϕ^*.

PROOF

Condition 1) of Theorem 1 has held, we only need to prove condition 2).
$\because \mathrm{H}(\phi|\phi) \geq H(\phi^{'}|\phi)\ \forall \phi^{'} \in \Omega$ (Theorem 1 of DLR)
$\therefore \nabla H(\phi_p|\phi_p) = 0$
$\therefore \nabla L(\phi_p) = \nabla Q(\phi_p|\phi_p) \neq 0\ \forall\ \phi_p \notin \varphi$
$\because \mathrm{Q}(\phi_{p+1}|\phi_p) > Q(\phi_p|\phi_p)$ and $H(\phi_p|\phi_p) \geq H(\phi_{p+1}|\phi_p)$
$\therefore \mathrm{L}(\phi_{p+1}) > L(\phi_p)$ ■

Theorem 2 can be easily applied because the continuity condition is not too strong. For example, if the unobserved data x can be expressed as the curved exponential family , then the continuity holds.

Curved Exponential Family

$\boldsymbol{X}$ is a random vector with p.d.f. $f(\boldsymbol{x}|\phi)$ from a probability space X.

$$f(\boldsymbol{x}|\phi) = A(\phi)\exp\left(\sum_{i=1}^{k} T_i(\boldsymbol{x})\eta_i(\phi)\right)h(\boldsymbol{x})$$

Where $T_i(\boldsymbol{x})$ is a real valued statistics, $\eta_i(\phi)$ is a real valued function on the parameter space $\Omega \subseteq R^q$, and $q < k \in \mathbb{N}$.
If $cov_\phi(\vec{T})$ ($\vec{T} = [T_1, T_2, \cdot, T_k]$) is positive definite, then $\boldsymbol{X}$ belongs to the curved exponential family.

Note that Theorem 2 DOES NOT apply to M (consider some $\phi_p \in \varphi$ whereas $\phi_p \notin M$). So, how to ensure that the sequence $L(\phi_k)$ converges to a local maximum? We need another condition, so we obtain the following theorem.

Theorem 3

If $Q(\psi|\phi)$ is continuous in ψ and ϕ, and $sup_{\phi^{'} \in \Omega} Q(\phi^{'}|\phi) > Q(\phi|\phi)\ \forall\ \phi \in \varphi \backslash M$
then all limit points of $\{\phi_p\}$ in an EM are local maxima of L, and $L(\phi_p)$ converges monotonically to $L^* = L(\phi^*)$ for some local maxima ϕ^*.

However, the new condition in Theorem 3 is hard to verify in real application. Therefore, Theorem 1 is the most general answer to our first question, and Theorem 2 provides a basis in real application.

Now, we have given an answer to the first question that it is not easy to make sure $L(\phi_p)$ converge to a local maximum, whereas we still do not know whether the sequence $\{\phi_p\}$ generated from the EM process converge to a specific point. Even though we have known that $L(\phi_p)$ converges to L^*, this convergence does not imply the convergence of the GEM (EM) sequence $\{\phi_p\}$ because this sequence is generated from a point-to-set map M. We study this question in the next section.

4 Does A EM Sequence $\{\phi_p\}$ Converge?

To better study the convergence of $\{\phi_p\}$, we define two sets as the following:

$$\varphi(a) = \{\phi \in \varphi : L(\phi) \equiv a\}$$

$$M(a) = \{\phi \in M : L(\phi) \equiv a\}$$

According to the definition above, if $L(\phi_p) \to L^*$ then the limit points of ϕ_p are in $\varphi(L^*)$ (or $M(L^*)$). Notice that if $\varphi(L^*)$ (or $M(L^*)$) consists of a single point ϕ^* then $\lim_{p\to\infty} \phi_p = \phi^*$, so we have the following theorem.

Theorem 4

ϕ_p is a GEM sequence generated by $\phi_{p+1} \in M(\phi_p)$, and suppose:
1) M is a closed p2s map over φ^C (or M^C)
2) $L(\phi_{p+1}) > L(\phi_p)\ \forall \phi_p \notin \varphi$ (or M)
If $\varphi(L^*) = \{\phi^*\}$ (or $M(L^*) = \{\phi^*\}$) where $L^* = \lim_{p\to\infty} L(\phi_p)$, then $\lim_{p\to\infty} \phi_p = \phi^*$

Note that condition 1) and 2) in Theorem 4 are the two conditions from Theorem 1, we only introduce an extra condition that $\varphi(L^*) = \{\phi^*\}$ (or $M(L^*) = \{\phi^*\}$). This new condition can be relaxed by $\lim_{p\to\infty} ||\phi_{p+1} - \phi_p|| = 0$ such that we have the next theorem, Theorem 5.

Theorem 5

ϕ_p is a GEM sequence generated by $\phi_{p+1} \in M(\phi_p)$, and suppose:
1) M is a closed p2s map over φ^C (or M^C)
2) $L(\phi_{p+1}) > L(\phi_p)\ \forall \phi_p \notin \varphi$ (or M)
If $\lim_{p\to\infty} ||\phi_{p+1} - \phi_p|| = 0$, then all limit points of $\{\phi_p\}$ are in a connected and compact subset of $\varphi(L^*)$ or $M(L^*)$ where $L^* = \lim_{p\to\infty} L(\phi_p)$.
(Here, a connected subset cannot be represented as an union of two disjoint sets.)
In particular, if $\varphi(L^*)$ or $M(L^*)$ is discrete, then ϕ_p converges to some ϕ^* in $\varphi(L^*)$ or $M(L^*)$.

PROOF

see Theorem 28.1 of Ostrowski (1967) [1] and Theorem 1. ■

Both Theorem 4 and 5 inherit condition 1) and 2) from Theorem 1, but we can obtain a condition that implies these two conditions and at the same time we strengthen the conditions after if in Theorem 4 and 5 to get a new theorem Theorem 6. To do this, we define a new set:

$$\psi(L) = \{\phi \in \Omega : L(\phi) \equiv L\}$$

Theorem 6

ϕ_p is a GEM sequence generated by $\phi_{p+1} \in M(\phi_p)$ with $\nabla Q(\phi_{p+1}|\phi_p) = 0$, and suppose $\nabla Q(\phi^{'}|\phi)$ is continuous in $\phi^{'}$ and ϕ.
If either (a) $\psi(L^*) = \{\phi^*\}$, or (b) $\lim_{p\to\infty} ||\phi_{p+1} - \phi_p|| = 0$ and $\psi(L^*)$ is discrete satisfies, then ϕ_p converges to a stationary point ϕ^* with $L(\phi^*) = L^*$, the limit of $L(\phi_p)$.

PROOF

The continuity of $\nabla Q(\phi^{'}|\phi)$ implies condition 1) and 2) of Theorem 1.
(a) or (b) in Theorem 6 is stronger than the condition after if in Theorem 4 or Theorem 5.
The continuity of $\nabla Q(\phi^{'}|\phi)$ and $\nabla Q(\phi_{p+1}|\phi_p) = 0$ imply $\nabla L(\phi^*|\phi^*) = \nabla Q(\phi^*|\phi^*) = 0$. ■

Condition a) in Theorem 6 can be replaced by that $L(\phi)$ is unimodal in Ω with ϕ^* being the only stationary point, so we get a corollary of Theorem 6 (Theorem 7).

Theorem 7

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

5 Summary

- (1) For an EM sequence $\{\phi_p\}$ that increases the likelihood $L(\phi_p)$, if $L(\phi_p)$ is bounded above, it converges to some L^*.
- (2) If $Q(\psi|\phi)$ is continuous in ψ and ϕ, then all limit points of $\{\phi_p\}$ in an EM are stationary points of L, and $L(\phi_p)$ converges monotonically to $L^* = L(\phi^*)$ for some point ϕ^*. The curved exponential family satisfies the continuity condition of Q. Additionally, if $\{\phi_p\}$ converges to some limit ϕ^*, then ϕ^* is a stationary point under the condition that $\nabla Q(\phi^{'}|\phi)$ is continuous in $\phi^{'}$ and ϕ.
- (3) To ensure that the limit of $L(\phi_p)$ is not a stationary value but a local maximum under the condition of the previous item, we need another condition: $sup_{\phi^{'} \in \Omega} Q(\phi^{'}|\phi) > Q(\phi|\phi)\ \forall\ \phi \in \varphi \backslash M$. However, this is a condition hard to verify in practice. To deal with this issue, Jeff Wu suggests that it is better to set several representative initial points in the parameter space to launch the EM algorithm, because whether EM will trapped into stationary points but not local maxima highly depends on the initializers.
- (4) In addition to item (2), if either (a) $\psi(L^*) = \{\phi^*\}$, or (b) $\lim\limits_{p \to \infty} ||\phi_{p+1} - \phi_p|| = 0$ and $\psi(L^*)$ is discrete satisfies, then ϕ_p converges to a stationary point ϕ^* with $L(\phi^*) = L^*$, the limit of $L(\phi_p)$.
- (5) If $L(\phi)$ is unimodal in Ω with ϕ^* being the only stationary point and that $\nabla Q(\phi^{'}|\phi)$ is continuous in ϕ and $\phi^{'}$, then for any EM sequence $\{\phi_p\}$, ϕ_p converges to the unique maximizer ϕ^* of $L(\phi)$.

6 Appendix

6.1 Proof of Global Convergence Theorem

Firstly, we prove the second consequence that $\alpha(x_k)$ converges monotonically to $\alpha(x^*)$ for some $x^* \in \Gamma$. Assume that x^* is a limit of $\{x_k\}_{k=0}^{\infty}$. Then there is a sub-sequence $\{x_{k_j}\}_{j=0}^{\infty}$ such that $\lim_{j\to\infty} x_{k_j} = x^*$. Since the ascent function $\alpha()$ is continuous, we have $\lim_{j\to\infty} \alpha(x_{k_j}) = \alpha(x^*)$. Additionally, we observe that α is monotonically increasing on $\{x_k\}_{k=0}^{\infty}$, so $\alpha(x^*) \geq \alpha(x_k)\ \forall\ k$. Since $\lim_{j\to\infty} x_{k_j} = x^*$, there exits a j_0 such that for $j > j_0$, $\alpha(x^*) - \alpha(x_{k_j}) < \epsilon \forall \epsilon > 0$. Hence, for all $k \geq k_{j_0}$:

$$\alpha(x^*) - \alpha(x_k) = \alpha(x^*) - \alpha(x_{k_{j_0}}) + \alpha(x_{k_{j_0}}) - \alpha(x_k) < \epsilon$$

which implies that $\alpha(x_k)$ converges monotonically to $\alpha(x^*)$.

Secondly, we prove $x^* \in \Gamma$ by contradiction. Suppose $x^* \notin \Gamma$, and consider the sequence $\{x_{k_j+1}\}_{j=0}^{\infty}$ which satisfies $x_{k_j+1} \in M(x_{k_j})$. Since $x_{k_j+1} \in S$, it has a convergent sequence $\lim_{l\to\infty} x_{(k_j+1)_l} = x^{**}$. Moreover, M is closed on $X\backslash\Gamma$, so $x^{**} \in M(x^*)$. By the previous proof, we have $\lim_{k\to\infty} \alpha(x_k) = \alpha(x^*)$, so $\alpha(x^{**}) = \alpha(x^*)$ which is contradictory to a) of 3) in the theorem.

References

[1] HOUSEHOL. AS. Ostrowski, am-solution of equations and systems of equations, 1967.

[2] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[3] Gordon D Murray. Contribution to discussion of paper by ap dempster, nm laird and db rubin. *J. Roy. Statist. Soc. Ser. B*, 39:27–28, 1977.

[4] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[5] Willard I Zangwill. *Nonlinear programming: a unified approach*, volume 52. Prentice-hall Englewood Cliffs, NJ, 1969.