ACCELERATING EM BY TARGETED AGGRESSIVE DOUBLE EXTRAPOLATION

Han-Shen Huang[1], *Bo-Hou Yang*[1,2], *Ren-Yuan Lyu*[2], *Chun-Nan Hsu*[1]

[1]Institute of Information Science, Academia Sinica, Taipei,Taiwan
[2]Department of Electrical Engineering, Chang Gung University,Taoyuan,Taiwan
chunnan@iis.sinica.edu.tw

ABSTRACT

The Expectation-Maximization (EM) algorithm is one of the most popular algorithms for parameter estimation from incomplete data, but its convergence can be slow for some large-scale or complex problems. Extrapolation methods can effectively accelerate EM, but to ensure stability, the learning rate of extrapolation must be compromised. This paper describes the TJ^{2}aEM method, a targeted extrapolation method that can extrapolate much more aggressively than competing methods without causing instability problems. We analyze its convergence properties and report experimental results.

Index Terms— Eigenvalues and eigenfunctions, Parameter estimation, Extrapolation, Convergence of numerical methods, Acceleration

1. INTRODUCTION

Let $\theta \in \mathbb{R}^d$ be a d-dimensional parameter vector for a probabilistic model to be trained by an EM mapping $M : \mathbb{R}^d \rightarrow \mathbb{R}^d$ which ensures that $L(M(\theta)) \geq L(\theta)$, L is the data likelihood. Starting from $\theta^{(0)}$, the EM algorithm applies M to $\theta^{(0)}$ iteratively until convergence to a local optimum θ^* that satisfies $\theta^* = M(\theta^*)$. One of the effective approach to accelerating the EM algorithm is the *parameterized EM* (pEM) algorithm, which extrapolate along the direction to the EM estimate with a learning rate η. Let $M_\eta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be the pEM mapping defined by:

$$M_\eta(\theta) \equiv \theta + \eta(M(\theta) - \theta). \quad (1)$$

It can be shown that with a constant $\eta \in (1, 2)$, pEM is guaranteed to converge, but it is usually too conservative to gain sufficient speedup. An optimal learning rate can be derived for pEM-like extrapolation [1] but in practice it is difficult to obtain this learning rate because it depends on the maximal and minimal eigenvalues of the Jacobian J of the EM mapping. In fact, the extrapolation can be made more aggressive to further accelerate the EM algorithm. *Adaptive overrelaxed EM* (aEM) [2] increases η by a constant ratio at every iteration if the pEM extrapolation increases the likelihood and resets η to one otherwise. aEM is guaranteed to converge because at each iteration, when an aEM step fails to improve the likelihood, it will backtrack to a plain EM step to guarantee monotone improvement of the likelihood. *Staggered EM* [3] estimates the maximal eigenvalue of J to obtain the upper bound of η and then rotates among learning rates within the bounded range in a predefined order or at random. Since these methods confine the range of the adjustment, their extrapolation may not be aggressive enough to achieve substantial speedup in some cases. More recently, the ϵ-accelerated EM [4] was proposed based on the vector ϵ algorithm [5], which was originally designed to accelerate a slowly convergent sequence. Varadhan and Roland proposed the SQUAREM algorithm [6], which extrapolates to a parameter vector on the straight line across two consecutive EM estimates in the parameter space such that this parameter vector is estimated to be the closest to the local optimum. However, they share a common disadvantage that they focus too much on accelerating slowly converged dimensions only.

We identified a key issue of pEM that prevents it from using an aggressive learning rate. The issue is that when we apply a learning rate that exceeds the proper range, the eigenvalue for a fast converging dimension of the Jacobian of the mapping will become negative. When the eigenvalue is less than -1, the extrapolation may bring the search away from the local optimum. In this paper, we describe a simple but effective solution to ensure that all eigenvalues including the minimal one are non-negative. We integrated this solution to aEM and *the triple jump EM method* (TJEM) [7] to derive a new method called TJ^{2}aEM, which has the advantage of aEM but can extrapolate much more aggressively to achieve a much higher acceleration. In the remainder of this paper, we present the derivation of this method, analyze its convergence properties, and empirically demonstrate its effectiveness in comparison with aEM. Due to the page limit, proofs of lemmas and propositions can be found in [8].

2. THE TRIPLE JUMP METHOD

Assuming that the EM mapping M is differentiable. We can apply a linear Taylor expansion of M around θ^* so that

$$\theta^{(t+1)} = M(\theta^{(t)}) \approx \theta^* + M'(\theta^*)(\theta^{(t)} - \theta^*) \quad (2)$$
$$= \theta^* + J(\theta^{(t)} - \theta^*), \quad (3)$$

where J abbreviates $M'(\theta^*)$, the Jacobian of EM at θ^*. From Eq. 3, we can derive Aitken's acceleration, which is the fundamental principle of pEM:

$$\theta^{(t+1)} = \theta^{(t)} + (I - J)^{-1}(M(\theta^{(t)}) - \theta^{(t)}).$$

Comparing this with Eq. 1, we can see that pEM is optimal when $\eta = (I - J)^{-1}$. However, it is usually prohibitively expensive to exactly compute J. Instead, a well-known method is to replace J with its largest eigenvalue λ_{max} to approximate J. We can use two consecutive EM estimates [9] to estimate λ_{max} by

$$\gamma^{(t)} \equiv \frac{\|M(\theta^{(t)}) - \theta^{(t)}\|}{\|\theta^{(t)} - \theta^{(t-1)}\|}. \quad (4)$$

Fraley [10] empirically assessed the accuracy of this estimate and showed that it is reasonably good, especially when there are high percentages of missing data. It can also be shown that as $t \to \infty$, $\gamma^{(t)} \leq \lambda_{max}$ asymptotically in the neighborhood of θ^* [3]. With the estimated eigenvalue, Aitken's acceleration becomes:

$$\theta^{(t+1)} = \theta^{(t)} + (1 - \gamma^{(t)})^{-1}(M(\theta^{(t)}) - \theta^{(t)}). \quad (5)$$

We named this method as *triple jump EM* (TJEM) because its search path is similar to the hop, step and jump phases in triple jump [7]. We can accelerate TJEM further by replacing the EM mapping with a pEM mapping M_η in Eq. 4 and 5. Let $\theta_\eta^{(t)} \equiv M_\eta(\theta^{(t)})$, we have

$$\gamma_\eta^{(t)} \equiv \frac{\|\theta_\eta^{(t)} - \theta^{(t)}\|}{\|\theta^{(t)} - \theta^{(t-1)}\|} \quad (6)$$

for the eigenvalue estimation and

$$\theta^{(t+1)} = \theta^{(t)} + (1 - \gamma_\eta^{(t)})^{-1}(\theta_\eta^{(t)} - \theta^{(t)}) \quad (7)$$

for extrapolation. We will refer to this mapping as the TJpEM mapping.

3. CONVERGENCE PROPERTY ANALYSIS

The rate of convergence of a fixed-point iteration mapping M is determined by the spectral radius ρ of its Jacobian J. For a pEM mapping M_η, the i-th eigenvalue, denoted by $\lambda_{\eta i}$, can be expressed as

$$\lambda_{\eta i} = (1 - \eta) * 1.0 + \eta\lambda_i, \quad (8)$$

where λ_i is the i-th eigenvalue of J of the EM mapping. For TJpEM, its rate of convergence is determined by the spectral radius of the Jacobian of the composition of the two mappings at θ^*:

$$M'_{\gamma_\eta}(M_\eta(\theta^*))M'_\eta(\theta^*) = M'_{\gamma_\eta}(\theta^*)M'_\eta(\theta^*) = J_{\gamma_\eta}J_\eta.$$

Lemma 1 gives the eigenvalues of $J_{\gamma_\eta}J_\eta$.

Lemma 1 *The i-th eigenvalue of the Jacobian of $M_{\gamma_\eta} \circ M_\eta$ at θ^* with estimated spectral radius $\gamma_\eta^{(t)}$ is*

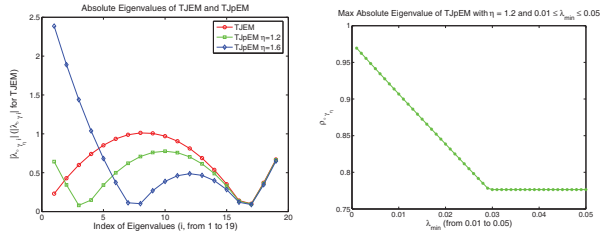
$$\lambda_{\eta i}\frac{\lambda_{\eta i} - \gamma_\eta^{(t)}}{1 - \gamma_\eta^{(t)}}.$$

To compare the spectral radii of TJEM and TJpEM, we assume that Eq. 8, the relation between the eigenvalues for the Jacobians of the EM and pEM mappings, holds for the estimated spectral radii $\gamma^{(t)}$ and $\gamma_\eta^{(t)}$:

$$\gamma_\eta^{(t)} = 1 - \eta + \eta\gamma^{(t)}.$$

Now, consider a Jacobian of the EM mapping with 19 distinct eigenvalues λ_i, $i = 1, \ldots, 19$. Assume further that $\lambda_i = 0.05 * i$. It follows that $\lambda_{min} = 0.05$ and $\lambda_{max} = 0.95$. Suppose TJEM estimates λ_{max} as $\gamma^{(t)} = 0.83$, an inaccurate underestimate. Then according to Eq. 8 and Lemma 1, if we choose $\eta = 1.2$ for TJpEM, we will have $\lambda_{\eta min}$, $\lambda_{\eta max}$, and $\gamma^{(t)}$ to be -0.14, 0.94, and 0.796, respectively, and if we choose $\eta = 1.6$, -0.52, 0.92, and 0.728, respectively. Fig. 1(a) illustrates the absolute eigenvalues of TJEM and TJpEM with these different learning rates. We can clearly observe the tendency that, with the growth of η, the peak of the concave curves in the middle in Fig. 1(a) decreases gradually, but the end of the left tails increases drastically. The figure illustrates that TJpEM can converge faster TJEM with a proper learning rate (e.g. $\eta = 1.2$), while slower or even diverge with a large (e.g. $\eta = 1.6$).

Then, we change λ_{min} and keep the other eigenvalues unchanged to see how sensitive the spectral radius is to λ_{min} and plot the result in Fig. 1(b), which shows that when $\lambda_{min} < 0.03$, the spectral radius increases linearly as λ_{min} decreases. The result shows that the spectral radius may be influenced by slight changes to λ_{min}. We can thus derive an upper bound of η for TJpEM guaranteed to converge faster than TJEM.



(a) Composite absolute eigenvalues of TJEM and TJpEM with $\eta = 1.2$ and 1.6.

(b) Composite spectral radii of TJpEM with $\eta = 1.2$ and different λ_{min}.

Fig. 1. Impact of Negative Eigenvalues

Proposition 2 *Within the neighborhood of θ^*, TJpEM with $\eta < \frac{1+\frac{\gamma}{4}}{1-\lambda_{min}}$ can converge faster than TJEM, under the assumption that $\gamma_\eta^{(t)} = 1 - \eta + \eta\gamma^{(t)}$.*

4. DOUBLE EXTRAPOLATION

Proposition 2 implies that TJpEM converges faster than TJEM with a proper learning rate, but when the learning rate exceeds the proper range, TJpEM might converge slower due to a large-sized negative eigenvalue. We now present a simple solution that can constrain the size of the spectral radius. The key idea is to apply double extrapolation that combines two pEM extrapolations into one. Then our mapping will be M_η^2, whose Jacobian $J_\eta^2 = Q\Lambda_\eta^2 Q^{-1}$, where Λ_η^2 will contain no negative eigenvalue. We will call this method as TJ^{2}pEM. Its derivation is as follows:

$$\begin{aligned}\theta^* &= \theta^{(t-1)} + \sum_{h'=0}^{\infty}(\theta^{(t+2h'+1)} - \theta^{(t+2h'-1)}) \quad (9)\\ &\approx \theta^{(t-1)} + \sum_{h'=0}^{\infty} J^{2h'}(\theta^{(t+1)} - \theta^{(t-1)})\\ &= \theta^{(t-1)} + (I - J_\eta^2)^{-1}(\theta^{(t+1)} - \theta^{(t-1)})\end{aligned}$$

Again, we replace $\theta^{(t+1)}$ with $\theta_\eta^{(t)}$ and use $\gamma_\eta^{(t)}$ as in TJpEM to replace J_η to obtain the extrapolation mapping of TJ^{2}pEM:

$$\theta^{(t+1)} = \theta^{(t-1)} + (1 - (\gamma_\eta^{(t)})^2)^{-1}(\theta_\eta^{(t)} - \theta^{(t-1)}). \quad (10)$$

Note that instead of extrapolating from $\theta^{(t)}$, TJ^{2}pEM extrapolates from $\theta^{(t-1)}$ at the t-th iteration. The next lemma gives the eigenvalues of the Jacobian of the TJ^{2}pEM mapping.

Lemma 3 *The i-th eigenvalue $\lambda_{\gamma_\eta^2 i}$ of the Jacobian of the TJ^{2}pEM mapping is:*

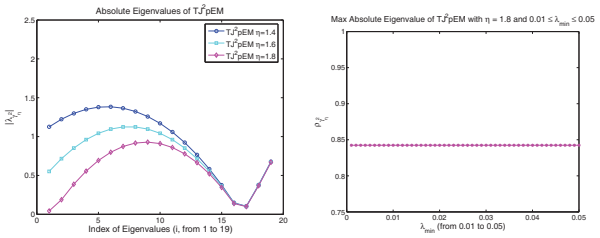
$$\lambda_{\gamma_\eta^2 i} = \frac{(\lambda_{\eta i})^2 - (\gamma_\eta^{(t)})^2}{1 - (\gamma_\eta^{(t)})^2}.$$

Proposition 4 *Given the same η, TJ^{2}pEM can converge faster than TJpEM if $\lambda_{\eta min} < -\frac{1}{2}$ and $\lambda_{\eta max} \leq \frac{1+\sqrt{2}}{2}\gamma_\eta^{(t)}$.*

Proposition 4 suggest that TJ^{2}pEM will successfully alleviate the impact of negative eigenvalues $\lambda_{\eta min}$ of TJpEM due to a large learning rate η. Another factor that influences $\lambda_{\eta min}$ is λ_{min}. Fig. 2 illustrates how λ_{min} may affect $\lambda_{\eta min}$. With the same example as in Fig. 1, Fig. 2(a) shows the absolute eigenvalues of TJ^{2}pEM with $\eta = 1.4$, 1.6, and 1.8. The curves have no left tail and a large η produces a smaller spectral radius. Fig. 2(b) shows that changes to λ_{min} will not affect the spectral radius here, suggesting that TJ^{2}pEM is barely affected by λ_{min}.

5. THE TJ^{2}AEM METHOD

Previously, Salakhutdinov et al. [2] showed that dynamically adjusting the learning rate for pEM will achieve a higher speedup than using the optimal learning rate at all iterations. However, the range of adjustment can only be confined in a proper range, otherwise, as discussed in the previous sections, a large sized negative eigenvalue may appear in the Jacobian. Now that we have solved this issue, we can replace the pEM mapping in TJ^{2}pEM with aEM to take advantage of both targetted aggressive extrapolation by the triple jump method and dynamic adjustment of the learning rate. We can establish that TJ^{2}aEM converges faster than TJ^{2}pEM in the neighborhood of the local optimum.



(a) Absolute eigenvalues of TJ^{2}pEM

(b) Spectral radii of TJ^{2}pEM with $\eta = 1.8$ and different λ_{min}

Fig. 2. Effect of Double Extrapolation

Proposition 5 *Let $\eta^{(1)} = \eta^* + \Delta$ and $\eta^{(2)} = \eta^* - \Delta$, where Δ is an arbitrary constant between 0 and 1. Let λ be the eigenvalues of the Jacobians of the mappings indicated by its subscripts. The spectral radius of TJ^{2}aEM with alternating learning rates $\eta^{(1)}$ and $\eta^{(2)}$ will be smaller than that of TJ^{2}pEM with η^*.*

Proof The spectral radius of TJ^{2}aEM is

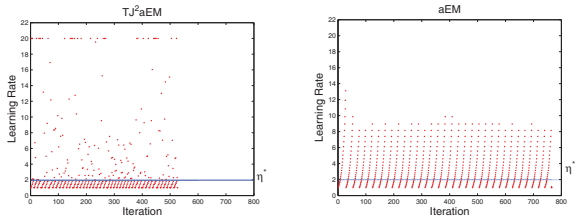
$$\begin{aligned}&|\lambda_{\gamma^2_{\eta^{(1)}} i}\lambda_{\gamma^2_{\eta^{(2)}} i}|\\ =\ & \frac{(\lambda_{\eta^{(1)} i})^2 - (\lambda_{\eta^{(1)}\max})^2}{1 - (\lambda_{\eta^{(1)}\max})^2} \cdot \frac{(\lambda_{\eta^{(2)} i})^2 - (\lambda_{\eta^{(2)}\max})^2}{1 - (\lambda_{\eta^{(2)}\max})^2}\\ \propto\ & \frac{(\lambda_{\eta^{(1)} i} + \lambda_{\eta^{(1)}\max})(\lambda_{\eta^{(2)} i} + \lambda_{\eta^{(2)}\max})}{(1 + \lambda_{\eta^{(1)}\max})(1 + \lambda_{\eta^{(2)}\max})}\\ =\ & \frac{(\lambda_{\eta^* i} + \lambda_{\eta^*\max})^2 - (\Delta(2 - \lambda_i - \lambda_{max}))^2}{(1 + \lambda_{\eta^*\max})^2 - (\Delta(1 - \lambda_{max}))^2}\\ \leq\ & \frac{(\lambda_{\eta^* i} + \lambda_{\eta^*\max})^2 - (\Delta(1 - \lambda_{max}))^2}{(1 + \lambda_{\eta^*\max})^2 - (\Delta(1 - \lambda_{max}))^2}\\ \leq\ & \frac{(\lambda_{\eta^* i} + \lambda_{\eta^*\max})^2}{(1 + \lambda_{\eta^*\max})^2}.\end{aligned}$$

When $\Delta = 0$, $|\lambda_{\gamma^2_{\eta^{(1)}} i}\lambda_{\gamma^2_{\eta^{(2)}} i}| = |\lambda_{\gamma^2_{\eta^*} i}|^2$, the spectral radius of TJ^{2}pEM. The above inequality implies that $|\lambda_{\gamma^2_{\eta^{(1)}} i}\lambda_{\gamma^2_{\eta^{(2)}} i}|$ with $\Delta \neq 0$ is smaller than $|\lambda_{\gamma^2_{\eta^*} i}|^2$. ■

6. EXPERIMENT

We compared the acceleration performance of pEM with an optimal learning rate and two algorithms that dynamically adjust their learning rates, aEM and TJ^{2}aEM, for training a mixture-of-Gaussian (MoG) model, where we have five equally-weighted 2D Gaussians with means at $\{(0,0), (0,1), (1,0), (0,-1), (-1,0)\}$ and variances 0.8. We randomly sampled 2,000 cases to form the experimental data set. We empirically determined the optimal learning rate η^* for the MoG as follows. First, we ran EM with a tiny threshold $(1.0e-11)$ and kept track of the parameter vectors searched and their likelihood. It took the EM algorithm 4,885 iterations to converge. We chose $\theta^{(501)}$ obtained by EM as the initial value because it is near the local optimum θ^*. Then, we tried pEM with various learning rates η and found that $\eta^* = 1.96$ is the optimal learning rate.

After that, we ran both aEM and TJ^{2}aEM from $\theta^{(501)}$. At each iteration, they dynamically adjust their learning rates. For aEM, its learning rate is adjusted by $\eta^{(t+1)} = 1.1\eta^{(t)}$, while for TJ^{2}aEM, η is dynamically assigned to 1.2, 1.4, 1.6, or 1.8 in a zigzag manner. With a different η, TJ^{2}aEM will come up with a different estimate γ_η at each iteration, and use the *effective learning rate* $\frac{1}{1-\gamma_\eta}$ to perform double extrapolation (see Eq. 5)We compared the effective learning rate of TJ^{2}aEM and the learning rate of aEM at each iteration, as shown in Fig 3. We can see that aEM increases its learning rate linearly until it reaches a point where it cannot satisfactorily improve the likelihood, while TJ^{2}aEM adjusts its effective learning rate irregularly and much aggressively. TJ^{2}aEM may adjust its learning rate to up to our predefined upper bound many times while aEM only reaches as high as 14 once and usually stops at 9. In the end, the elapsed iterations for TJ^{2}aEM and aEM are 527 and 766, respectively. Both outperform pEM with a fixed optimal learning rate, which required 1,327 iterations to converge.



(a) Learning rates used by TJ^{2}aEM (b) Learning rates used by aEM

Fig. 3. Trace of learning rates used by TJ^{2}aEM and aEM as a function of iterations.

7. CONCLUSION

We have presented TJ^{2}aEM, a targetted extrapolation method to accelerate the convergence of EM. TJ^{2}aEM extrapolates aggressively along with a dynamic learning rate. We contribute new ideas to explore for further acceleration. The first is that a mapping whose Jacobian contains negative eigenvalues, like pEM, can still achieve speedup. Traditionally, only mappings with semi-positive definite Jacobians are considered. The second is that negative eigenvalues can be handled by double extrapolation.

8. REFERENCES

[1] Gunther Hammerlin and Karl-Heinz Hoffmann, *Numerical Mathematics*, Springer-Verlag: New York, 1991.

[2] Ruslan Salakhutdinov and Sam Roweis, "Adaptive overrelaxed bound optimization methods," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 2003, pp. 664–671.

[3] Tim Hesterberg, "Staggered Aitken acceleration for EM," in *Proceedings of the Statistical Computing Section of the American Statistical Association*, Minneapolis, Minnesota, USA, August 2005.

[4] Masahiro Kuroda and Michio Sakakihara, "Accelerating the convergence of the EM algorithm using the vector ϵ algorithm," *Computational Statistics and Data Analysis*, vol. 51, pp. 1549–1561, 2006.

[5] P. Wynn, "Acceleration techniques for iterated vector and matrix problems," *Mathematics of Computation*, vol. 16, 1962.

[6] R. Varadhan and Ch. Roland, "Squared extrapolation methods (SQUAREM)," Tech. Rep. Paper 63, Department of Biostatistics Working Paper, Johns Hopkins University, 2004.

[7] Han-Shen Huang, Bo-Hou Yang, and Chun-Nan Hsu, "Triple-jump acceleration for the EM algorithm," in *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM-2005)*, Houston, TX, USA, November 2005, pp. 649–652.

[8] Han-Shen Huang, Bo-Hou Yang, and Chun-Nan Hsu, "TJ^{2}aEM: targeted aggressive extrapolation method for accelerating the em algorithm.," Tech. Rep. TR-IIS-07-012, Institute of Information Science, Academia Sinica, Taiwan, 2007.

[9] Joseph L. Schafer, *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall / CRC Press, 1997.

[10] Chris Fraley, "On computing the largest fraction of missing information for the em algorithm and the worst linear function for data augmentation," *Computational Statistics and Data Analysis*, vol. 31, no. 1, pp. 13–26, 1999.