Acceleration of the EM algorithm

Shiro Ikeda

The Institute of Physical and Chemical Research (RIKEN)
Saitama, 351-01 Japan, Shiro.Ikeda@brain.riken.go.jp

Abstract— The EM algorithm is widely used to estimate the parameters of many applications. It is simple but the convergence speed is slow. There is another algorithm called the scoring method which is faster but complicated. We show these two methods can be connected by using the EM algorithm recursively.

I. INTRODUCTION

The EM (Expectation Maximization) algorithm[1] was originally proposed by Dempster et al.[2] for estimating the MLE (Maximum Likelihood Estimator) of stochastic models which have hidden random variables. The algorithm is now used in many applications such as Boltzmann machine[3], Mixture of Expert networks[4][5][6] and also in HMM (Hidden Markov Model)[7].

This algorithm gives us an iterative procedure and the practical form is usually very simple. However, the convergence speed is slow compared to the scoring method which is also used to estimate the MLE of these models. There are some works to accelerate the convergence speed of the EM algorithm[8][9], but the procedure is usually not easy and need a lot of calculations.

In this paper, we show that we can accelerate the EM algorithm by using it in a recursive way. The algorithm consists of two stages. In the first stage, we do one EM step with the given data set. In the second stage, we do another EM step not with the given data but with the data drawn from the model. Through these stages, we can have better estimator. We show the theoretical derivation of the algorithm in connection with the scoring method. We also show some results of computer simulations. They show the algorithm gives faster convergence speed.

II. THE EM ALGORITHM AND THE SCORING METHOD

Think about the cases we want to estimate the parameters of a Boltzmann machine[3] or a stochastic Perceptron[10]. These models can be denoted as, $p(x|\boldsymbol{\theta})$ as the probabilistic distribution, where $x = (y, z)$ is the output of cells, y is for visible cells and z is for hidden cells. If a model can be formulated in this form, we can use the EM algorithm for estimating the MLE. The hidden random variable z makes it hard to find the MLE.

For the training of these models, we can only have the sampled data on y generated by a fixed probability distribution. Let us define the observed empirical distribution as $\hat{q}(y) = \sum_{s=1}^{N} \delta(y_s)/N$, where the set of data is $\{y_1, \cdots, y_N\}$. The log-likelihood function is,

$$\begin{aligned} L(Y^N|\boldsymbol{\theta}) &= \frac{1}{N}\sum_{s=1}^{N} \log p(y_s|\boldsymbol{\theta}) = \frac{1}{N}\sum_{s=1}^{N} l(y_s|\boldsymbol{\theta}) \\ &= E_{\hat{q}(y)}\left[l(y|\boldsymbol{\theta})\right]. \end{aligned} \tag{1}$$

We want the MLE $\hat{\boldsymbol{\theta}}$ which maximizes $L(Y^N|\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, L(Y^N|\boldsymbol{\theta})$. Note that in (1), we are trying to fit $\boldsymbol{\theta}$ to $\hat{q}(y)$ but we can also do this for any other distribution $r(y)$ on y, then the likelihood function to be maximized will be $E_{r(y)}\left[l(y|\boldsymbol{\theta})\right]$.

In this paper, we only treat $p(x|\boldsymbol{\theta})$ which is an exponential family. This means the probability density function can be written as

$$p(x|\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{p} \theta^i r_i(x) - k(\boldsymbol{r}(x)) - \psi(\boldsymbol{\theta})\right), \tag{2}$$

where $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^p)$ is the natural parameter and $\boldsymbol{r}(x) = (r_1(x), \cdots, r_p(x))$. The Boltzmann machine and the stochastic Perceptron are also exponential families [10][3]. Note that the marginal distribution $p(y|\boldsymbol{\theta})$, which is defined as $p(y|\boldsymbol{\theta}) = E_{p(z|\boldsymbol{\theta})}\left[p(y|z, \boldsymbol{\theta})\right]$, is not always an exponential family.

The EM algorithm is an iterative algorithm generating a sequence $\{\boldsymbol{\theta}_t\}$ $(t = 1, 2, 3, \cdots)$ of estimates from an initial point $\boldsymbol{\theta}_0$. Each iteration consists of the following two sub-steps:

- E-step: $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = E_{\hat{q}(y)p(z|y,\boldsymbol{\theta}_t)}\left[l(y, z|\boldsymbol{\theta})\right]$
- M-step: $\boldsymbol{\theta}_{t+1} = \mathrm{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$.

After a cycle of E- and M-step, we obtain $\boldsymbol{\theta}_{t+1}$ and it is shown[2] that $L(Y^N|\boldsymbol{\theta}_{t+1}) \geq L(Y^N|\boldsymbol{\theta}_t)$. By iterating E- and M-steps, the algorithm converges to the parameters which should be the MLE. And we have an approximation of the one EM step as the following. For the proof, see [11],

$$\boldsymbol{\theta}_{t+1} \simeq \boldsymbol{\theta}_t + {G_X}^{-1}(\boldsymbol{\theta}_t)\partial L(Y^N|\boldsymbol{\theta}_t). \quad (3)$$

Here, $\partial = (\partial_1, \cdots, \partial_p)^T = (\partial/\partial\theta^1, \cdots, \partial/\partial\theta^p)^T$ and $G_X(\boldsymbol{\theta}) = ({g_X}_{ij}(\boldsymbol{\theta}))$ is the Fisher information matrix of $p(x|\boldsymbol{\theta})$ defined as,

$$\begin{aligned} {g_X}_{ij}(\boldsymbol{\theta}) &= E_{p(x|\boldsymbol{\theta})}\left[\partial_i l(x|\boldsymbol{\theta})\partial_j l(x|\boldsymbol{\theta})\right] \\ &= -E_{p(x|\boldsymbol{\theta})}\left[\partial_i\partial_j l(x|\boldsymbol{\theta})\right]. \quad (4) \end{aligned}$$

(3) shows that the EM algorithm is updating the parameter into the steepest direction of the likelihood function based on the Fisher metric G_X of the model. Note that this relation only holds for the natural parameter $\boldsymbol{\theta}$.

(3) looks similar to what is called the scoring method in statistics. The updating rule of the scoring method is,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + {G_Y}^{-1}(\boldsymbol{\theta}_t)\partial L(Y^N|\boldsymbol{\theta}_t). \quad (5)$$

It is known that the scoring method is an efficient method of calculating the MLE and the convergence speed of it is usually faster than the EM algorithm. This is caused by the difference of the coefficient matrices, $G_X(\boldsymbol{\theta})$ and $G_Y(\boldsymbol{\theta})$. $G_Y(\boldsymbol{\theta}) = ({g_Y}_{ij}(\boldsymbol{\theta}))$ is also the Fisher information matrix of the model $p(y|\boldsymbol{\theta})$ where a hidden random variable z is eliminated,

$$\begin{aligned} {g_Y}_{ij}(\boldsymbol{\theta}) &= E_{p(y|\boldsymbol{\theta})}\left[\partial_i l(y|\boldsymbol{\theta})\partial_j l(y|\boldsymbol{\theta})\right] \\ &= -E_{p(y|\boldsymbol{\theta})}\left[\partial_i\partial_j l(y|\boldsymbol{\theta})\right]. \quad (6) \end{aligned}$$

Both matrices are Fisher information matrices. It is know that $G_X(\boldsymbol{\theta})$ and $G_Y(\boldsymbol{\theta})$ have a relation as the following.

$$\begin{aligned} -E_{p(y|\boldsymbol{\theta})}\left[\partial_i\partial_j l(y|\boldsymbol{\theta})\right] &= -E_{p(x|\boldsymbol{\theta})}\left[\partial_i\partial_j l(x|\boldsymbol{\theta})\right] \\ &\quad +E_{p(x|\boldsymbol{\theta})}\left[\partial_i\partial_j l(z|y,\boldsymbol{\theta})\right] \\ G_Y(\boldsymbol{\theta}) &= G_X(\boldsymbol{\theta}) - G_{Z|Y}(\boldsymbol{\theta}) \quad (7) \end{aligned}$$

where $G_{Z|Y} = ({g_{Z|Y}}_{ij}(\boldsymbol{\theta}))$ is the conditional Fisher information matrix defined as,

$$\begin{aligned} {g_{Z|Y}}_{ij}(\boldsymbol{\theta}) &= -E_{p(y|\boldsymbol{\theta})}\left[E_{p(z|y,\boldsymbol{\theta})}\left[\partial_i\partial_j l(z|y,\boldsymbol{\theta})\right]\right] \\ &= E_{p(y|\boldsymbol{\theta})}\left[{g_{Z|y}}_{ij}(\boldsymbol{\theta})\right]. \end{aligned}$$

Because G_Y, G_X, and $G_{Z|Y}$ are positive definite symmetric matrices in regular cases, we can show an interesting result,

$$\begin{aligned} G_Y &= (I - G_{Z|Y}{G_X}^{-1})G_X \\ {G_Y}^{-1} &= {G_X}^{-1}(I - G_{Z|Y}{G_X}^{-1})^{-1} \\ &= {G_X}^{-1}\left(I + \sum_{i=1}^{\infty}(G_{Z|Y}{G_X}^{-1})^i\right) \quad (8) \end{aligned}$$

(8) can be proved easily by diagonalizing G_Y, G_X, and $G_{Z|Y}$ simultaneously. All the eigenvalues of $G_{Z|Y}{G_X}^{-1}$ are real, positive and smaller than 1 and we have, $(I - G_{Z|Y}{G_X}^{-1})^{-1} = (I + \sum_1^\infty (G_{Z|Y}{G_X}^{-1})^i)$.

III. PROPOSED ALGORITHM

One step of the scoring method changes an estimator $\boldsymbol{\theta}$ into the steepest gradient based on G_Y and it usually converges faster than the basic EM algorithm. But to calculate ${G_Y}^{-1}$ is complicated in most of the cases. We propose an algorithm to approximate the scoring method through the use of the EM algorithm in a recursive way.

Suppose the case we have executed one step of the EM algorithm and have $\boldsymbol{\theta}_{t+1}$ from $\boldsymbol{\theta}_t$. Then we do another E-M step from $\boldsymbol{\theta}_t$ to have $\bar{\boldsymbol{\theta}}_{t+1}$ using the data drawn from $p(y|\boldsymbol{\theta}_{t+1})$. Here, we don't use the original data. With $\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_{t+1}$ and $\bar{\boldsymbol{\theta}}_{t+1}$ we can make a better estimator. This is the essence of the proposed algorithm. The obtained parameter $\bar{\boldsymbol{\theta}}_{t+1}$ has the following property.

Theorem 1 *$\bar{\boldsymbol{\theta}}_{t+1}$ is the parameter estimated from $\boldsymbol{\theta}_t$ by one EM step, taking $p(y|\boldsymbol{\theta}_{t+1})$ as the true (teacher) distribution. $\bar{\boldsymbol{\theta}}_{t+1}$ has the property,*

$$\bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t \simeq {G_X}^{-1}G_Y{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t). \quad (9)$$

The proof can be obtained by following the derivation of (3). Replacing $\hat{q}(y)$ with $p(y|\boldsymbol{\theta}_{t+1})$, we have,

$$\bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t \simeq {G_X}^{-1}\int \partial l(y|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} p(y|\boldsymbol{\theta}_{t+1})\mu(y). \quad (10)$$

Because, $p(y|\boldsymbol{\theta}_{t+1}) \simeq p(y|\boldsymbol{\theta}_t) + p(y|\boldsymbol{\theta}_t)(\partial l(y|\boldsymbol{\theta}_t))^T(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$, we have the proof.

From (7) and (9),

$$\begin{aligned} \bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t &\simeq {G_X}^{-1}(G_X - G_{Z|Y}){G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ &\simeq (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) \\ &\quad -{G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ \boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1} &\simeq {G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t). \quad (11) \end{aligned}$$

With (11), we can approximate the scoring method up to second order by

$$\begin{aligned} \boldsymbol{\theta}' &= 2\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}) \\ &\simeq \boldsymbol{\theta}_t + {G_X}^{-1}(I + G_{Z|Y}{G_X}^{-1})\partial L(Y^N|\boldsymbol{\theta}_t) \quad (12) \end{aligned}$$

Also we can approximate the scoring method up to higher order in the following way.

Do one EM step from $\boldsymbol{\theta}_t$ to have $\bar{\boldsymbol{\theta}}_{t+i}$ where $p(y|\bar{\boldsymbol{\theta}}_{t+i-1})$ is the teacher distribution ($i = 0, 1, \cdots$, and $\bar{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_{t+1}$), $\bar{\boldsymbol{\theta}}_{t+i}$ has the following property.

$$\begin{aligned} \bar{\boldsymbol{\theta}}_{t+i} - \boldsymbol{\theta}_t &\simeq ({G_X}^{-1}G_Y)^i{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \\ &= (I - {G_X}^{-1}G_{Z|Y})^i{G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t) \end{aligned} \quad (13)$$

From $\bar{\boldsymbol{\theta}}_t$, $\cdots$, $\bar{\boldsymbol{\theta}}_{t+i}$, and $\boldsymbol{\theta}_t$, we can approximate $({G_X}^{-1}G_{Z|Y})^i {G_X}^{-1}\partial L(Y^N|\boldsymbol{\theta}_t)$ and the scoring descendant vector up to ith order. And if $i = p$, we do not have to do this more. We can calculate higher order approximation with linear combination.

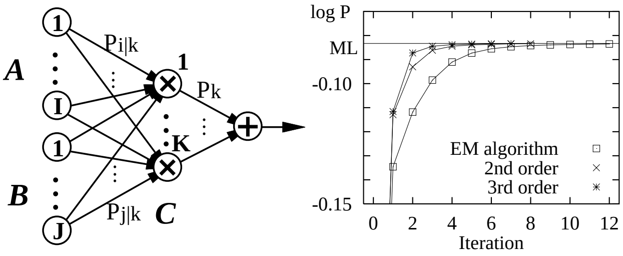
This proposed algorithm shows that after one EM step, we can have better estimator without using the original data. The procedure is very simple, we use the data drawn from the model.

IV. SIMULATIONS

A. Log-Linear model

We first used Log-Linear Model for the simulation. The model (Fig.1) has a triplet of variables (A, B, C), where A, B and C take values on $\{A_i\}$, $\{B_j\}$ and $\{C_k\}$ respectively, $(i = 1, \cdots, I,\ j = 1, \cdots, J,\ k = 1, \cdots, K)$. We can observe two variables A, B of them, but cannot observe C (latent variable). We make a model with the probability distribution, $P(A, B, C) = P(A_i|C_k)P(B_j|C_k)P(C_k)$. The distribution of A and B are independently conditional to C.

When we observe data, we can only know the marginal distribution on A, and B $m_{ij} = n_{ij}/\sum_{i'j'} n_{i'j'}$. Where n_{ij} is the observed number of $(A = A_i, B = B_j)$. From the model, marginal distribution is $P(A_i, B_j) = \sum_k P_{i|k}P_{j|k}P_k$ and we have to estimate the parameters including the hidden probabilistic variable C. We can use the EM algorithm to estimate the parameter, and also the proposed algorithm.



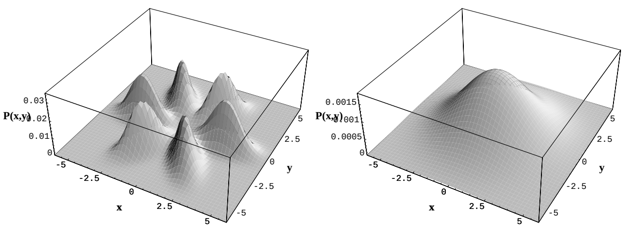
Definition of the model Log likelihood

Figure 1: The definition of the model and the results

The simulation was made with the model in which $I = J = 5$, and $K = 2$. Therefore, $p(A_i, B_j)$ is multinomial distribution of 25 elements. If we have 24 parameters, we can describe the given distribution precisely, but now we only have $(K-1) + K(I-1) + K(J-1) = 17$ parameters. The teacher distribution was made at random, and the problem is to estimate the parameter to fit the teacher distribution.

Fig.1 is the result using the basic EM algorithm, the proposed procedure which approximate the scoring method up to 2nd order and 3rd order. You can see that if we use the 2nd or 3rd order approximation, the convergence speed is much faster than the basic EM algorithm.

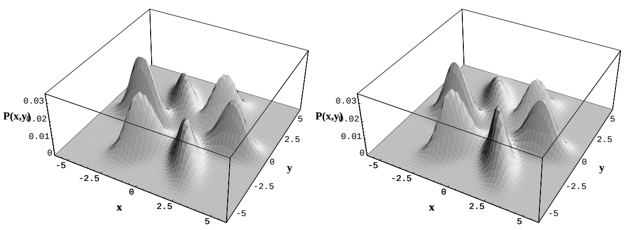
B. Mixture of normal distributions



Teacher model Initial model for learning

Figure 2: Teacher model and Initial model

When the density function of the model is continuous, we cannot use the density function itself to carry out the EM algorithm but we need a sampled data set. We have to create the set from $p(y|\boldsymbol{\theta}_{t+1})$ by drawing data and do one EM step to have $\bar{\boldsymbol{\theta}}_{t+1}$. We did a simulation using the mixture of normal distributions[12]. Fig. 2 shows the teacher model and the initial model for learning. Both models are consists of 6 components, but the components in the initial model are broad and we cannot see each component separately.



The basic EM algorithm The Proposed algorithm

Figure 3: Results of the algorithms

We do not show the form of the EM algorithm, but it is easy to derive the form. In order to show the performance of the higher order approximation of the scoring method, we did simulation as follows.

1. Prepare a sample of 1000 observed y's from the teacher model. Let the parameter of the initial model be $\boldsymbol{\theta}_0$.

2. Using the original data, execute one EM step to have θ'_{t+1} from θ_t.

3. Generate 1000 new data according to $p(y|\theta'_{t+1})$.

4. Using the newly generated data, execute one EM step and calculate $\bar{\theta}'_{t+1}$ from θ_t.

5. Let $\theta_{t+1} = 2\theta'_{t+1} - \bar{\theta}'_{t+1}$, and go to 2.

The final models obtained through the basic EM algorithm and the 2nd order approximation are shown in Fig.3. And their profile of the likelihood function is given in Fig.4. Because the proposed algorithm uses a kind of Monte Carlo method, it does not converge but keep fluctuating. This is also the reason why we did not test higher order approximation. The result shows that the proposed method can accelerate the EM algorithm.

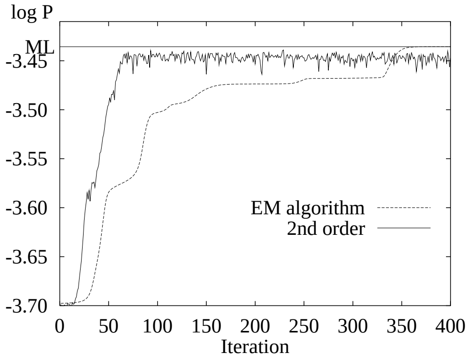


Figure 4: Results of the algorithms

V. DISCUSSION

Through the simulations, we show that this algorithm improves the convergence speed of the EM algorithm. In order to have the second order approximation, we have to use the EM algorithm twice. Therefore we hope that the new algorithm works twice faster than the original EM algorithm. This depends on the problem. In the case of the mixture of normal distributions, it was faster more than two times in our simulation.

For the acceleration of the EM algorithm, there is a work of Louis [8]. He used the same kind of approximation but used the Jacobian matrix J of the map $\theta_{t+1} = EM(\theta)$. J corresponds to $J = (G_{Z|Y} {G_X}^{-1})$ of this paper. Using J and $\theta_{t+1} - \theta_t$, he formulated the acceleration of the EM algorithm as we did. But J is not easy to be calculated for many models. Meng and Rubin gave a method for calculating J using the EM algorithm, but it requires to do EM steps as much times as the number of the parameters. After you have J, you can have the approximation up to any order, but even to have the second order, the method needs many EM steps[9], while our algorithm only needs two EM steps.

We believe that the new algorithm is quite useful for the on-line learning. Suppose the case of on-line learning, but data does not come every time. When we have a new datum, we can update the parameter using the EM algorithm, but when we don't have data for a while, we can continue learning by generating new data from the model. We are now working for the application of Neural Network models and of the on-line learning.

References

[1] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* Wiley series in probability and statistics, John Wiley & Sons, Inc., 1997.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[3] S. Amari, K. Kurata, and H. Nagaoka, "Information geometry of Boltzmann machines," *IEEE Transactions on Neural Networks*, vol. 3, pp. 260–271, March 1992.

[4] R. A. Jacobs and M. I. Jordan, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[5] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, March 1994.

[6] M. I. Jordan and L. Xu, "Convergence results for the EM approach to mixture of experts architectures," *Neural Networks*, vol. 8, no. 9, pp. 1409–1431, 1995.

[7] L. Rabiner, S. Levinson, and M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *The Bell System Technical Journal*, vol. 62, pp. 1075–1105, April 1983.

[8] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. R. Statistical Society, Series B*, vol. 44, no. 2, pp. 226–233, 1982.

[9] M. A. Tanner, *Tools for Statistical Inference – Observed Data and Data Augmentation Methods*, vol. 67 of *Lecture Notes in Statistics.* Springer-Verlag, 1991.

[10] S. Amari, "Dualistic geometry of the manifold of higher-order neurons," *Neural Networks*, vol. 4, no. 4, pp. 443–451, 1991.

[11] D. Titterington, "Recursive parameter estimation using incomplete data," *J. R. Statistical Society, Series B*, vol. 46, no. 2, pp. 257–267, 1984.

[12] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixture." A.I.Memo No.1520, C.B.C.L. Paper No.111, 1995.