LETTER **Communicated by Shu-Kay Ng**

Asymptotic Convergence Properties of the EM Algorithm for Mixture of Experts

Yan Yang
Jinwen Ma
jwma@math.pku.edu.cn
Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China

Mixture of experts (ME) is a modular neural network architecture for supervised classification. The double-loop expectation-maximization (EM) algorithm has been developed for learning the parameters of the ME architecture, and the iteratively reweighted least squares (IRLS) algorithm and the Newton-Raphson algorithm are two popular schemes for learning the parameters in the inner loop or gating network. In this letter, we investigate asymptotic convergence properties of the EM algorithm for ME using either the IRLS or Newton-Raphson approach. With the help of an overlap measure for the ME model, we obtain an upper bound of the asymptotic convergence rate of the EM algorithm in each case. Moreover, we find that for the Newton approach as a specific Newton-Raphson approach to learning the parameters in the inner loop, the upper bound of asymptotic convergence rate of the EM algorithm locally around the 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 the ME model. That is, as the average overlap of the true ME model with large sample tends to zero, the EM algorithm with the Newton approach to learning the parameters in the inner loop tends to be asymptotically superlinear. Finally, we substantiate our theoretical results by simulation experiments.

1 Introduction

For a complex problem with different subtasks on different occasions, it is often efficient to use the divide-and-conquer principle that divides a single problem into simpler ones whose separate solutions can be combined to yield a final solution. According to this principle, Jacobs, Jordan, Nowlan, and Hinton (1991) proposed the mixture of experts (ME) in which the data are assumed to be summarized by a collection of networks, each

Color versions of figures in this letter are presented in the online supplement available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00154.

defined over a local region of the input space. That is, the ME architecture is composed of several different expert networks plus a gating network that decides which of the experts should be used for each training case. Jordan and Jacobs (1992) further proposed the hierarchical mixture of experts (HME) model, which is a tree-structured model with a tree of gating networks combining the expert networks in a larger group.

For learning the parameters in the ME architecture with a given data set, Jordan and Jacobs (1994) established an expectation-maximization (EM) algorithm under the EM framework (Dempster, Laird, & Rubin, 1977). A further theoretical investigation on this EM algorithm was made by Jordan and Xu (1995) to show its relation with the gradient ascent method. The EM algorithm is a general technique for maximum likelihood estimation, consisting of the expectation (E) step and the maximization (M) step. In the E-step, using the given observed data and the current estimates of the parameters, the expectation of the log-likelihood function over the complete data space defined by the so-called Q function is computed. In the M-step, the parameters are updated to maximize the Q-function. Particularly for the ME architecture, the M-step needs to solve two maximization problems—one associated with the parameters in the gating network and the other with the parameters in the expert networks. Fortunately, the latter maximization problem can be solved directly, and the former one can be solved in a double-loop way by the iteratively reweighted least squares (IRLS) algorithm (Jordan & Jacobs, 1994). Later, Chen, Xu, and Chi (1999) suggested the Newton-Raphson approximation approach for implementing inner-loop learning in the EM algorithm instead of the IRLS approach in order to improve stability. Moreover, Ng and McLachlan (2004) proposed the use of an expectation-conditional maximization (ECM) algorithm (Meng, 1994) to train the ME network. The single-loop EM algorithms for ME were also proposed to speed up the convergence (Xu, Jordan, & Hinton, 1994; Yang & Ma, 2009).

Although the EM algorithm for ME has been widely used in pattern recognition and signal processing, its convergence properties have not been investigated in depth. Jordan and Xu (1995) provided a good theoretical analysis on the EM algorithm for ME through the IRLS approach showing that the EM algorithm outperforms the gradient ascent algorithm by having a positive projection on the gradient of the log likelihood. However, so far there has not been any further theoretical result on the convergence of the EM algorithm for ME. Recent theoretical investigations on the asymptotic convergence properties of the EM algorithms for gaussian mixtures and the mixtures of densities from exponential families (Xu & Jordan, 1996; Xu, 1997; Ma, Xu, & Jordan, 2000; Ma & Xu, 2005; Ma & Fu, 2005) provide a new view on the EM algorithm for ME. In fact, these investigations found that the asymptotic convergence behavior of the EM algorithm is closely related to the overlap measure of the components in the true mixture to generate the sample data. Since the input data of the ME can be considered from a finite mixture with a certain degree of overlap, we can apply this overlap

measure analysis to the EM algorithm for ME to investigate its asymptotic properties.

In this letter, after defining a measure of average overlap of experts in the ME architecture (in a similar way as that of gaussian mixtures), we investigate the asymptotic convergence rate of the EM algorithm for ME using either the IRLS or Newton-Raphson scheme. In each case, we obtain an upper bound of asymptotic convergence rate of the EM algorithm. Moreover, we found that for the Newton method as a specific Newton-Raphson approach to learning the parameters in the inner loop, the asymptotic convergence rate of the EM algorithm locally around the true solution Θ^* tends to be zero, as the measure of average overlap in the true ME architecture tends to zero. As the average overlap of the true ME architecture using a large sample tends to zero, the EM algorithm for ME with the Newton approach to learning the parameters in the inner loop tends to be asymptotically superlinear.

The rest of the letter is organized as follows. In section 2, we introduce the EM algorithm for ME, as well as a general upper bound of its asymptotic convergence rate. We then present several definitions, conditions, and lemmas in section 3. Section 4 contains the main theorems. We further substantiate them by simulation experiments in section 5. A brief conclusion is given in section 6.

2 The EM Algorithm and Its Asymptotic Convergence Rate

We consider the following ME model,

$$P(y|x,\Theta) = \sum_{j=1}^{K} P(j|x)P(y|x,\theta_j) = \sum_{j=1}^{K} g_j(x,\theta_0)P(y|x,\theta_j), \tag{2.1}$$

where

$$g_j(x,\theta_0) = \begin{cases} \dfrac{e^{s_j(x,\theta_0)}}{1+\sum_{i=1}^{K-1} e^{s_i(x,\theta_0)}}, & j = 1,\ldots,K-1, \\ \dfrac{1}{1+\sum_{i=1}^{K-1} e^{s_i(x,\theta_0)}}, & j = K, \end{cases} \tag{2.2}$$

$$\begin{aligned} P(y|x,\theta_j) &= \frac{1}{(2\pi)^{m/2}|\Sigma_j|^{1/2}} \\ &\quad \times \exp\left\{-\frac{1}{2}[y-f_j(x,\theta_j)]^T\Sigma_j^{-1}[y-f_j(x,\theta_j)]\right\}. \end{aligned} \tag{2.3}$$

K is the number of experts in the mixture, $x \in \mathbb{R}^n$ denotes the input vector, and $y \in \mathbb{R}^m$ is the output vector. Θ is the set of all the parameters, including $\theta_0 = [\theta_{01}^T, \cdots, \theta_{0(K-1)}^T]^T$ and component parameter vectors θ_j with the

corresponding covariance matrices Σ_j, which are assumed positive definite or diagonal. All of the functions f_j and s_j are assumed to be linear in the parameters

$$f_j(x, \theta_j) = X^T \theta_j, \quad j = 1, \ldots, K,$$
$$s_j(x, \theta_0) = [x^T, 1]\theta_{0j}, \quad j = 1, \ldots, K-1,$$

where

$$X^T = \left\{ \begin{array}{ccccc|ccccc} x^T & 0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & x^T & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots & \vdots & & & & \vdots \\ 0 & \cdots & \cdots & 0 & x^T & 0 & \cdots & \cdots & 0 & 1 \end{array} \right\}.$$

For convenience, we let $s_K(x, \theta_0) = 0$.

Given a training set $\mathcal{S} = \{x^{(t)}, y^{(t)}\}_{t=1}^N$ from the mixture of experts described by equation 2.1, the log-likelihood function can be written as

$$l(\Theta, \mathcal{S}) = \sum_{t=1}^{N} \ln \sum_{j=1}^{K} g_j(x^{(t)}, \theta_0) P(y^{(t)} | x^{(t)}, \theta_j). \tag{2.4}$$

To maximize $l(\Theta, \mathcal{S})$, the EM algorithm was established by Jordan and Jacobs (1994) and further detailed by Jordan and Xu (1995), which can be given iteratively in two steps as follows. In the E-step, we compute the posterior probabilities $h_j^{(k)}(t)$ by

$$h_j^{(k)}(t) = P(j | x^{(t)}, y^{(t)}, \Theta^{(k)}) = \frac{g_j(x^{(t)}, \theta_0^{(k)}) P(y^{(t)} | x^{(t)}, \theta_j^{(k)})}{\sum_{i=1}^{K} g_i(x^{(t)}, \theta_0^{(k)}) P(y^{(t)} | x^{(t)}, \theta_i^{(k)})}. \tag{2.5}$$

In the M-step, we update the parameters as follows:

$$\theta_0^{(k+1)} = \theta_0^{(k)} + \gamma_g (R_g^{(k)})^{-1} e_g^{(k)}, \tag{2.6}$$

$$\theta_j^{(k+1)} = (R_j^{(k)})^{-1} c_j^{(k)}, \tag{2.7}$$

$$\begin{aligned} \Sigma_j^{(k+1)} &= \frac{1}{\sum_{t=1}^{N} h_j^{(k)}(t)} \sum_{t=1}^{N} h_j^{(k)}(t) \left[y^{(t)} - f_j(x^{(t)}, \theta_j^{(k)}) \right] \\ &\quad \times \left[y^{(t)} - f_j(x^{(t)}, \theta_j^{(k)}) \right]^T, \end{aligned} \tag{2.8}$$

where

$$e_g^{(k)} = \sum_{t=1}^{N} \sum_{j=1}^{K-1} \left[h_j^{(k)}(t) - g_j(x^{(t)}, \theta_0^{(k)}) \right] \frac{\partial s_j\left(x^{(t)}, \theta_0^{(k)}\right)}{\partial \theta_0^{(k)}},$$

$$R_g^{(k)} = \sum_{t=1}^{N} \sum_{j=1}^{K-1} g_j\left(x^{(t)}, \theta_0^{(k)}\right)\left(1 - g_j\left(x^{(t)}, \theta_0^{(k)}\right)\right) \frac{\partial s_j\left(x^{(t)}, \theta_0^{(k)}\right)}{\partial \theta_0^{(k)}} \frac{\partial s_j\left(x^{(t)}, \theta_0^{(k)}\right)}{\partial \theta_0^{(k)T}},$$

$$c_j^{(k)} = \sum_{t=1}^{N} h_j^{(k)}(t) X_t \left(\Sigma_j^{(k)}\right)^{-1} y^{(t)},$$

$$R_j^{(k)} = \sum_{t=1}^{N} h_j^{(k)}(t) X_t \left(\Sigma_j^{(k)}\right)^{-1} X_t^T,$$

and γ_g is the learning rate. It is clear that $R_j^{(k)}$ is nonsingular with probability one when the sample size N becomes sufficiently large.

If we consider the above iteration paradigm as a global loop of parameter learning, it actually contains an inner loop for learning the parameters θ_0 in the gating network by equation 2.6. Actually, this inner loop needs to implement the IRLS algorithm for a number of iterations. In order to improve the performance of the EM algorithm, Chen et al. (1999) suggested implementing the Newton-Raphson method in the inner loop, which can be given iteratively as

$$\theta_0^{(k+1)} = \theta_0^{(k)} - \alpha H_g^{-1}\left(\theta_0^{(k)}, \mathcal{S}\right) J\left(\theta_0^{(k)}, \mathcal{S}\right), \tag{2.9}$$

where α is the learning rate and $0 < \alpha \leq 1$. H_g is the Hessian matrix of the log-likelihood function associated with the parameters θ_0:

$$l_g(\theta_0, \mathcal{S}) = \sum_{t=1}^{N} \sum_{j=1}^{K} h_j(t) \log g_j(x^{(t)}, \theta_0), \tag{2.10}$$

and J is the first derivative of $l_g(\theta_0, \mathcal{S})$ with respect to θ_0. Specifically, the Newton-Raphson method becomes the Newton method when $\alpha = 1$.

For theoretical analysis, Jordan and Xu (1995) established the following relationship between the EM update of the IRLS approach and the gradient of the log likelihood:

$$\theta_0^{(k+1)} - \theta_0^{(k)} = P_g^{(k)} \frac{\partial l}{\partial \theta_0}|_{\theta_0 = \theta_0^{(k)}}, \tag{2.11}$$

$$\theta_j^{(k+1)} - \theta_j^{(k)} = P_j^{(k)} \frac{\partial l}{\partial \theta_j}|_{\theta_j=\theta_j^{(k)}}, \tag{2.12}$$

$$\mathrm{vec}\left[\Sigma_j^{(k+1)}\right] - \mathrm{vec}\left[\Sigma_j^{(k)}\right] = P_{\Sigma_j}^{(k)} \frac{\partial l}{\partial \mathrm{vec}[\Sigma_j]}|_{\Sigma_j=\Sigma_j^{(k)}}, \tag{2.13}$$

where

$$P_g^{(k)} = \gamma_g \left(R_g^{(k)}\right)^{-1},$$
$$P_j^{(k)} = \left(R_j^{(k)}\right)^{-1},$$
$$P_{\Sigma_j}^{(k)} = \frac{2}{\sum_{t=1}^{N} h_j^{(k)}(t)} \Sigma_j^{(k)} \otimes \Sigma_j^{(k)},$$

$\mathrm{vec}[B]$ denotes the vector obtained by stacking the column vectors of the matrix B and $\otimes$ denotes the Kronecker product. For convenience, we set $P(\Theta^{(k)}) = \mathrm{diag}[P_g^{(k)}, P_1^{(k)}, \ldots, P_K^{(k)}, P_{\Sigma_1}^{(k)}, \ldots, P_{\Sigma_K}^{(k)}]$ be the projection matrix.

Jordan and Xu (1995) further found that $P(\Theta^{(k)})$ makes the EM algorithm superior to the gradient ascent algorithm by implementing a positive projection on the gradient of the log likelihood. Moreover, this relation constructs a theoretical foundation for our analysis on the convergence rate of the EM algorithm. Theoretically, the EM iterative procedure converges to a local maximum of the log-likelihood (Dempster et al., 1977). We suppose that $\hat{\Theta}$ is a local solution to maximizing the log-likelihood function given equation 2.4 and the EM algorithm converges to it. Furthermore, we assume that the sample data $\{x^{(t)}, y^{(t)}\}_{t=1}^{N}$ (as the training data) are generated from the mixture of experts of the parameters Θ^* in an independent and identically distributed (i.i.d.) manner with the help of a given probability density function $P(x)$ for generating the component sample data $x^{(t)}$, and that the EM algorithm asymptotically correctly converges to this true parameter (i.e., when N is large, the EM algorithm converges to $\hat{\Theta}$ with $\lim_{N\to\infty} \hat{\Theta} = \Theta^*$). We now analyze the local convergence rate around this consistent solution in the limit form. Following the same analysis of the EM algorithm for gaussian mixtures (Ma et al., 2000) by simplifying the inner-loop learning as a one-step iteration, we found that the local convergence rate of the EM algorithm for ME around $\hat{\Theta}$ is bounded by

$$r = \lim_{k\to\infty} \frac{||\Theta^{(k+1)} - \hat{\Theta}||}{||\Theta^{(k)} - \hat{\Theta}||} \leq ||I + \lim_{N\to\infty} P(\Theta^*)H(\Theta^*)||, \tag{2.14}$$

where $H(\Theta^*)$ is the Hessian matrix of $l(\Theta, \mathcal{S})$ at $\Theta = \Theta^*$ under the sample data set. It should be noted that as compared with equation 2.21 given in

Ma et al. (2000) for the EM algorithm for gaussian mixtures, there is no E matrix because there is no constraint on θ_0.

According to equation 2.14, the local convergence rate of the EM algorithm for ME around the true solution Θ^* is dominated by the convergence result of the matrix product $P(\Theta^*)H(\Theta^*)$. Thus, we will try to analyze the convergence behavior of $P(\Theta^*)H(\Theta^*)$ as N increases to infinity in the following sections.

3 Definitions and Lemmas

Inspired by the definition of the average overlap measure for gaussian mixtures with the true parameters (Ma et al., 2000), we can also utilize $\gamma_{ij}(t) = h_i(t)(\delta_{ij} - h_j(t))$ to measure the overlap between experts i and j per sample $(x^{(t)}, y^{(t)})$, where δ_{ij} is the Kronecker function and

$$h_j(t) = P(j|x^{(t)}, y^{(t)}, \Theta^*) = \frac{g_j(x^{(t)}, \theta_0^*)P(y^{(t)}|x^{(t)}, \theta_j^*)}{\sum_{i=1}^{K} g_i(x^{(t)}, \theta_0^*)P(y^{(t)}|x^{(t)}, \theta_i^*)}. \tag{3.1}$$

With a training set $\mathcal{S} = \{(x^{(t)}, y^{(t)})\}_{t=1}^{N}$ from the mixture of K experts of the parameters Θ^*, we can asymptotically define a set of quantities on the overlap of any two experts (i.e., expert distributions), including one and itself, as follows:[1]

$$e_{ij}(\Theta^*) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} |\gamma_{ij}(t)| = \int |\gamma_{ij}(x, y)| P(x, y|\Theta^*)\, dx dy,$$

where $\gamma_{ij}(x, y) = h_i(x, y)(\delta_{ij} - h_j(x, y))$, $h_j(x, y) = P(j|x, y, \Theta^*)$ and $P(x, y|\Theta^*) = P(x)P(y|x, \Theta^*)$.

As for the average overlap measure for the ME model with the true parameters Θ^*, we consider the worst case and define

$$e(\Theta^*) = \max_{i,j} e_{ij}(\Theta^*), \qquad \text{for } i, j = 1, \ldots, K.$$

For further analysis, we also define

$$e_{ij}(x, \Theta^*) = \int |\gamma_{ij}(x, y)| P(y|x, \Theta^*)\, dy$$

[1] Here the overlap measure between one and itself means the sum of the overlap measures of this expert to all the other experts.

and

$$e(x, \Theta^*) = \max_{i,j} e_{ij}(x, \Theta^*), \qquad \text{for } i, j = 1, \ldots, K.$$

Just as in the case of gaussian mixtures, $e(\Theta^*)$ can tend to zero when the experts in the ME model can be well separated. Actually, the ME model is simplified to a gaussian mixture if x is fixed. For convenience of the analysis, we make some assumptions that regularize the manner of $e(\Theta^*)$ tending to zero:

$$\text{Condition 1:} \qquad g_i(x, \theta_0^*) \geq \omega, \qquad \text{for } i = 1, \ldots, K,$$

where ω is a positive constant. Our second assumption is that the eigenvalues of all the covariance matrices satisfy

$$\text{Condition 2:} \qquad \beta\lambda(\Theta^*) \leq \lambda_{ik} \leq \lambda(\Theta^*), \text{ for } i = 1, \ldots, K, k = 1, \ldots, m,$$

where β is also a positive constant and $\lambda(\Theta^*)$ is defined to be the maximum eigenvalue of the covariance matrices $\Sigma_1^*, \ldots, \Sigma_K^*$, that is,

$$\lambda(\Theta^*) = \max_{i,k} \lambda_{ik}.$$

The third assumption is that

$$\text{Condition 3:} \qquad \nu D_{\max}(\Theta^*, x) \leq D_{\min}(\Theta^*, x) \leq ||f_i(x, \theta_i^*) - f_j(x, \theta_j^*)|| \leq D_{\max}(\Theta^*, x), \text{ for } i \neq j,$$

where $D_{\max}(\Theta^*, x) = \max_{i \neq j} ||f_i(x, \theta_i^*) - f_j(x, \theta_j^*)||$, $D_{\min}(\Theta^*, x) = \min_{i \neq j} ||f_i(x, \theta_i^*) - f_j(x, \theta_j^*)||$, ν is still a positive constant.

We then define three kinds of special polynomial functions that we often meet in the further analyses.

Definition 1. *$q(y, x, \Theta^*)$ is called a regular function if it satisfies:*

i. If both Θ^* and x are fixed, $q(y, x, \Theta^*)$ is a polynomial function of the component variables $y_1, \ldots, y_m$ of y.

ii. If y is fixed, $q(y, x, \Theta^*)$ is a polynomial function of the elements of $f_1(x, \theta_1^*), \ldots, f_K(x, \theta_K^*), g_1(x, \theta_0^*), \ldots, g_K(x, \theta_0^*), g_1(x, \theta_0^*)^{-1}, \ldots, g_K(x, \theta_0^*)^{-1}$, as well as $\Sigma_1^*, \ldots, \Sigma_K^*, \Sigma_1^{*-1}, \ldots, \Sigma_K^{*-1}$.

Definition 2. *$q(y, x, \Theta^*)$ is called a balanced function if it satisfies (i) and*

iii. If y is fixed, $q(y, x, \Theta^*)$ is a polynomial function of the elements of $f_1(x, \theta_1^*), \ldots, f_K(x, \theta_K^*), g_1(x, \theta_0^*), \ldots, g_K(x, \theta_0^*), g_1(x, \theta_0^*)^{-1}, \ldots, g_K(x, \theta_0^*)^{-1}, \Sigma_1^*, \ldots, \Sigma_K^*, \lambda(\Theta^*)\Sigma_1^{*-1}, \ldots, \lambda(\Theta^*)\Sigma_K^{*-1}$.

Definition 3. *For a regular function $q(y, x, \Theta^*)$, if there is a positive number s such that $\lambda^s(\Theta^*)q(y, x, \Theta^*)$ is converted into a balanced function, then $q(y, x, \Theta^*)$ is called a regular and convertible function.*

We now describe the four lemmas we will use in the proofs of the main theorems.

Lemma 1. *Let $\mathcal{D} \subset \mathbb{R}^n$ be a bounded closed set such that $\int_{\mathcal{D}} P(x)dx = 1$, for any x in $\mathcal{D}$, $P(x) > 0$, and Θ^* as well as $x \in \mathcal{D}$ satisfy conditions 1 to 3. As $e(\Theta^*) \to 0$ is considered an infinitesimal,*

i. Letting $\widetilde{\mathcal{D}}$ be a subset of $\mathcal{D}$ such that $x \in \widetilde{\mathcal{D}}$ iff $e(x, \Theta^*) \to 0$, the complementary set $\mathcal{D} - \widetilde{\mathcal{D}}$ is a zero measure set.

ii. Letting $e(\Theta^*) = e_{pq'}(\Theta^*)$, and $||1/e_{pq'}(\cdot, \Theta^*)||_\infty$ denote the supremum of $|1/e_{pq'}(x, \Theta^*)|$ on $\mathcal{D}$, there exists a positive number M such that $\lim_{e(\Theta^*)\to 0} ||1/e_{pq'}(\cdot, \Theta^*)||_\infty e(\Theta^*) \le M$.

Proof. Let $\lambda^i_{\max} = \max_j\{\lambda_{ij}\}$, $\eta_{ij}(x, \Theta^*) = (\lambda^i_{\max})^{\frac{1}{2}}(\lambda^j_{\max})^{\frac{1}{2}}/||f_i(x, \theta^*_i) - f_j(x, \theta^*_j)||$, and $\eta(x, \Theta^*) = \max_{i\neq j} \eta_{ij}(x, \Theta^*)$. As proved by Ma et al. (2000), $e(x, \Theta^*) \to 0$ is equivalent to $\eta(x, \Theta^*) \to 0$, and as $\eta(x, \Theta^*) \to 0$, $\eta(x, \Theta^*)$, $\eta_{ij}(x, \Theta^*)$, and $\zeta_{ij}(x, \Theta^*) \triangleq \lambda^i_{\max}/||f_i(x, \theta^*_i) - f_j(x, \theta^*_j)||$ are all equivalent infinitesimals.

Let $\delta_{\mathcal{D}} = \max_{x\neq x'\in\mathcal{D}} ||x - x'||$. Let $e(\Theta^*) = e_{pq'}(\Theta^*)$ and x_0 be a point in $\widetilde{\mathcal{D}}$, noting that zero point cannot be in $\widetilde{\mathcal{D}}$. Obviously there exists such an x_0; otherwise, $e(\Theta^*)$ cannot tend to 0. Let $q = q'$ if $q' \neq p$; otherwise, q may be any other component index except for p. For $e_{pp}(x_0, \Theta^*) = \sum_{q\neq p} e_{pq}(x_0, \Theta^*)$, we have $e_{pq}(x_0, \Theta^*) \to 0$ as $e(x_0, \Theta^*) \to 0$. For any other point x in $\mathcal{D}$, we have $||f_p(x, \theta^*_p) - f_q(x, \theta^*_q)|| - ||f_p(x_0, \theta^*_p) - f_q(x_0, \theta^*_q)|| \le ||f_p(x, \theta^*_p) - f_q(x, \theta^*_q) - f_p(x_0, \theta^*_p) + f_q(x_0, \theta^*_q)|| \le (\delta_{\mathcal{D}} + 1)||\theta^*_p - \theta^*_q||$. Therefore, we get

$$\frac{\zeta_{pq}(x_0, \Theta^*)}{\zeta_{pq}(x, \Theta^*)} = \frac{||f_p(x, \theta^*_p) - f_q(x, \theta^*_q)||}{||f_p(x_0, \theta^*_p) - f_q(x_0, \theta^*_q)||} \le \frac{(\delta_{\mathcal{D}} + 1)||\theta^*_p - \theta^*_q||}{||f_p(x_0, \theta^*_p) - f_q(x_0, \theta^*_q)||} + 1.$$

It can be seen from the above inequality that as $\zeta_{pq}(x_0, \Theta^*)$ tends to zero, for any x in $\widetilde{\mathcal{D}}$, $\zeta_{pq}(x, \Theta^*)$ tends to zero with the same or a lower order. Since x_0 can be any point in $\widetilde{\mathcal{D}}$, we further achieve that as $e(\Theta^*) \to 0$, for any x in $\widetilde{\mathcal{D}}$, $e_{pq'}(x, \Theta^*)$ tends to 0 with the same order.

By the definition of $e(\Theta^*)$, we have

$$1 = \lim_{e(\Theta^*)\to 0} \int \frac{e_{pq'}(x, \Theta^*)}{e(\Theta^*)} P(x)\, dx \tag{3.2}$$

$$= \int_{\widetilde{\mathcal{D}}} \lim_{e(\Theta^*)\to 0} \frac{e_{pq'}(x, \Theta^*)}{e(\Theta^*)} P(x)\, dx + \int_{\mathcal{D}-\widetilde{\mathcal{D}}} \lim_{e(\Theta^*)\to 0} \frac{e_{pq'}(x, \Theta^*)}{e(\Theta^*)} P(x)\, dx. \tag{3.3}$$

If the set $\mathcal{D} - \widetilde{\mathcal{D}}$ is not a zero measure set, the second term in the right side of equation 3.3 tends to infinity because $e_{pq'}(x, \Theta^*)/e(\Theta^*)$ tends to infinity as $e(\Theta^*) \to 0$. Thus, the measure of $\mathcal{D} - \widetilde{\mathcal{D}}$ must be zero. So i is proved.

Furthermore, for any x in $\widetilde{\mathcal{D}}$, as $e(\Theta^*) \to 0$, $e_{pq'}(x, \Theta^*)$ and $e(\Theta^*)$ are equivalent infinitesimals. Otherwise, $e_{pq'}(x, \Theta^*)/e(\Theta^*)$ tends to infinity or zero, and the first term on the right side of equation 3.3 cannot be 1. For any x in $\widetilde{\mathcal{D}}$, x' in $\mathcal{D} - \widetilde{\mathcal{D}}$, we have $\lim_{e(\Theta^*)\to 0} ||1/e_{pq'}(x', \Theta^*)|| < \lim_{e(\Theta^*)\to 0} ||1/e_{pq'}(x, \Theta^*)||$. Hence, there exists a positive number M, such that

$$\lim_{e(\Theta^*)\to 0} \left|\left| \frac{1}{e_{qp'}(\cdot, \Theta^*)} \right|\right|_\infty e(\Theta^*) \le M. \tag{3.4}$$

Therefore, ii is also proved.

Lemma 2. *Let $\mathcal{D} \subset \mathbb{R}^n$ be a bounded closed set such that $\int_{\mathcal{D}} P(x)dx = 1$, for any x in $\mathcal{D}$, $P(x) > 0$ and Θ^* as well as x satisfy conditions 1 to 3. Suppose that $q(y, x, \Theta^*)$ is a regular and convertible function and $u(x)$ is a polynomial function of the component variables $x_1, \ldots, x_n$ of x. As $e(\Theta^*) \to 0$ is considered as an infinitesimal, we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} |\gamma_{ij}(t)| q\left(y^{(t)}, x^{(t)}, \Theta^*\right) u\left(x^{(t)}\right)$$
$$= \int |\gamma_{ij}(x, y)| q(y, x, \Theta^*) u(x) P(x, y|\Theta^*) dy dx = o(e^{0.5-\varepsilon}(\Theta^*)),$$

where $\varepsilon > 0$ is an arbitrarily small number.

Proof. Let $e(\Theta^*) = \int_{\mathcal{D}} e_{pq'}(x, \Theta^*) P(x) dx$, where $e_{pq'}(x, \Theta^*) = \int |\gamma_{pq'}(x, y)| P(y|x, \Theta^*) dy$. According to lemma 1, as $e(\Theta^*) \to 0$ is considered as an infinitesimal, there is a positive number M such that $\lim_{e(\Theta^*)\to 0} ||1/e_{pq'}(\cdot, \Theta^*)||_\infty e(\Theta^*) \le M$.

Letting $\mu_{ij}(x, \Theta^*) = \int |\gamma_{ij}(x, y)| q(y, x, \Theta^*) P(y|x, \Theta^*) dy$, according to lemma 4 given in Ma et al. (2000) and the above conditions, as $e(x, \Theta^*) \to 0$ is considered an infinitesimal, $e(x, \Theta^*)$ and $e_{pq'}(x, \Theta^*)$ are equivalent infinitesimals, and $\mu_{ij}(x, \Theta^*) = o(e^{0.5-\varepsilon}(x, \Theta^*))$, where $\varepsilon > 0$ is an arbitrarily small number. Thus, as $e(\Theta^*) \to 0$, for any x in $\widetilde{\mathcal{D}}$, we have $e(x, \Theta^*) \to 0$ and, further, $\mu_{ij}(x, \Theta^*) = o(e_{pq'}^{0.5-\varepsilon}(x, \Theta^*))$.

The polynomial function $u(x)$ is bounded on $\mathcal{D}$, and we can write it as $||u||_\infty \le \delta$, where δ is a positive number. By lemma 1, $\mathcal{D} - \widetilde{\mathcal{D}}$ is a zero measure set, so that integration on $\mathcal{D}$ is equivalent to $\widetilde{\mathcal{D}}$. Recalling Holder's

inequality $\int |v_1(x)v_2(x)|dx \le ||v_1||_\infty \int |v_2(x)|dx$, we have

$$\begin{aligned}
&\lim_{e(\Theta^*)\to 0} \frac{\int |\gamma_{ij}(x,y)|q(y,x,\Theta^*)u(x)P(x,y|\Theta^*)\,dydx}{e^{0.5-\varepsilon}(\Theta^*)} \\
&\quad = \lim_{e(\Theta^*)\to 0} \int_{\mathcal{D}} \mu_{ij}(x,\Theta^*)u(x)e^{-0.5+\varepsilon}(\Theta^*)P(x)\,dx \\
&\quad = \lim_{e(\Theta^*)\to 0} \int_{\widetilde{\mathcal{D}}} u(x)\frac{\mu_{ij}(x,\Theta^*)}{e^{0.5-\varepsilon}_{pq'}(x,\Theta^*)}\frac{1}{e^{-0.5+\varepsilon}_{pq'}(x,\Theta^*)}e^{-0.5+\varepsilon}(\Theta^*)P(x)\,dx \\
&\quad \le \lim_{e(\Theta^*)\to 0} \delta||\frac{\mu_{ij}(\cdot,\Theta^*)}{e^{0.5-\varepsilon}_{pq'}(\cdot,\Theta^*)}||_\infty||\frac{1}{e^{-0.5+\varepsilon}_{pq'}(\cdot,\Theta^*)}||_\infty e^{-0.5+\varepsilon}(\Theta^*) \\
&\quad \le \delta M^{-0.5+\varepsilon}||\lim_{e(\Theta^*)\to 0}\frac{o(e^{0.5-\varepsilon}_{pq'}(\cdot,\Theta^*))}{e^{0.5-\varepsilon}_{pq'}(\cdot,\Theta^*)}||_\infty \\
&\quad = 0.
\end{aligned}$$

Hence, $\int |\gamma_{ij}(x,y)|q(y,x,\Theta^*)u(x)P(x)\,dydx = o(e^{0.5-\varepsilon}(\Theta^*))$. The proof is completed.

Lemma 3. *Let $\mathcal{D} \subset \mathbb{R}^n$ be a bounded closed set such that $\int_{\mathcal{D}} P(x)dx = 1$, for any x in $\mathcal{D}$, $P(x) > 0$, Θ^* as well as x satisfy conditions 1 to 3. Then there exists a positive number δ for all $j \in \{1,\dots,K\}$, $lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N} h_j(t) \ge \delta$.*

Proof. For any $l \ne j$, $l, j \in \{1,\dots,K\}$, $f_l(x)^T\Sigma_l^{*-1} - f_j(x)^T\Sigma_j^{*-1}$ and $f_l(x)^T\Sigma_l^{*-1}f_l(x) - f_j(x)^T\Sigma_j^{*-1}f_j(x)$ are bounded on $\mathcal{D}$ where we use $f_l(x)$ as a short expression of $f_l(x,\theta_l^*)$ in purpose of conciseness. Eyy^T and Ey are bounded too. Hence, there exists a positive number M such that $Eyy^T \le M$, $Ey \le M$, $||f_l^T\Sigma_l^{*-1} - f_j^T\Sigma_j^{*-1}||_\infty \le M$, and $||f_l^T\Sigma_l^{*-1}f_l - f_j^T\Sigma_j^{*-1}f_j||_\infty \le M$. Let $\sigma_1 = \max_{l\ne j}|\Sigma_l^*|^{1/2}/|\Sigma_j^*|^{1/2}$, $\sigma_2 = \max_{l\ne j}\mathrm{tr}((\Sigma_l^{*-1} - \Sigma_j^{*-1})E(yy^T))$, and $\sigma_3 = \max_{l\ne j}||(f_l^T(\Sigma_l^*)^{-1} - f_j^T(\Sigma_j^*)^{-1})Ey||_\infty$ where the notation $\mathrm{tr}(\cdot)$ stands for the trace of a matrix. Hence we get the following inequality:

$$\begin{aligned}
&\int \frac{g_l(x,\theta_0^*)}{g_j(x,\theta_0^*)}\frac{p_l(y|x,\Theta^*)}{p_j(y|x,\Theta^*)}p(y,x|\Theta^*)\,dydx \\
&\le \left|\left|\frac{g_l}{g_j}\right|\right|_\infty \sigma_1 \int \exp\left\{\frac{1}{2}\mathrm{tr}((\Sigma_j^{*-1} - \Sigma_l^{*-1})yy^T) + (f_l^T\Sigma_l^{*-1}\right. \\
&\quad \left. - f_j^T\Sigma_j^{*-1})y + \frac{1}{2}(f_j^T\Sigma_j^{*-1}f_j - f_l^T\Sigma_l^{*-1}f_l)\right\}p(y,x|\Theta^*)\,dydx \\
&\le \frac{(1-\omega K+\omega)\sigma_1}{\omega}\exp\left(\frac{\sigma_2}{2} + \sigma_3 + \frac{M}{2}\right) \triangleq \beta.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} h_j(t) \\
&= 1\Bigg/\left(1+\sum_{l\neq j}\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\left(\frac{g_l(x^{(t)},\theta_0^*)p_l(y^{(t)}|x^{(t)},\Theta^*)}{g_j(x^{(t)},\theta_0^*)p_j(y^{(t)}|x^{(t)},\Theta^*)}\right)\right) \\
&\geq \frac{1}{1+(K-1)\beta}.
\end{aligned}$$

Let $\delta = 1/(1+(K-1)\beta)$, and we reach the conclusion of the lemma. Meanwhile, because $\sum_{j=1}^{K} h_j(t) = 1$, we further get $\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N} h_j(t) \leq 1-(K-1)\delta$.

Before we give the last lemma, we need to discuss the covariance matrix of the random vector X subject to $P(x)$, that is, $\mathrm{cov}(X) = E[(X-EX)(X-EX)^T]$. In general, it is positive definite. Obviously it is always nonnegative definite. It is not positive definite only if X is distributed on a subspace of R^n (i.e., the space of x). That is, $P(x)$ is degenerated and distributed on a subspace of R^n. In the following analysis, we always assume that $\mathrm{cov}(X)$ is positive definite, without regard to those degenerated cases.

Lemma 4. *Let $a(t) = a(x^{(t)}, y^{(t)})$ be a $K(n+1)\times d_1$–dimensional matrix function, $b(t) = b(x^{(t)}, y^{(t)})$ be an $m(n+1)\times d_2$–dimensional matrix function, and $c(t) = c(x^{(t)}, y^{(t)})$ be a $d_3\times d_4$–dimensional matrix function, where d_1, d_2, d_3, and d_4 can be any positive integer.*

i. If $lim_{N\to\infty}\frac{1}{N}a(t) = o(e^{0.5-\varepsilon}(\Theta^*))$, then $lim_{N\to\infty}R_g^{-1}a(t) = o(e^{0.5-\varepsilon}(\Theta^*))$. If $lim_{N\to\infty}R_g^{-1}a(t) = 0$, then $lim_{N\to\infty}\frac{1}{N}a(t) = 0$.

ii. If $lim_{N\to\infty}\frac{1}{N}b(t) = o(e^{0.5-\varepsilon}(\Theta^*))$, then $lim_{N\to\infty}R_j^{-1}b(t) = o(e^{0.5-\varepsilon}(\Theta^*))$.

iii. If $lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}c(t) = o(e^{0.5-\varepsilon}(\Theta^*))$, then $lim_{N\to\infty}\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}c(t) = o(e^{0.5-\varepsilon}(\Theta^*))$.

Proof. (i). Let EX, EXX^T be the expectations of X and XX^T. Since $\mathrm{cov}(X)$ is positive definite, there exists an orthogonal matrix Q such that $Q^T EX = \beta_1 e_1$, where $\beta_1^2 = ||EX||_2$ and $e_1\in\mathbb{R}^n$ denotes $(1,0,\ldots,0)^T$. There exists an orthogonal matrix P such that $\mathrm{cov}(X) = P\Lambda P^T$, $\Lambda = \mathrm{diag}\{\lambda_1,\ldots,\lambda_n\}$, where $\lambda_1\geq\cdots\geq\lambda_n > 0$ are the eigenvalues of $\mathrm{cov}(X)$. Noticing that $EXX^T = \mathrm{cov}(X) + EXEX^T$, we therefore have $EX^T(EXX^T)^{-1}EX = \beta_1 e_1^T\cdot Q^T P(\Lambda+\beta_1^2 e_1 e_1^T)^{-1}P^T Q\cdot\beta_1 e_1 = \frac{\beta_1^2}{\lambda_1+\beta_1^2} < 1$. The determinant of matrix $\begin{pmatrix} EXX^T & EX \\ EX^T & 1\end{pmatrix} \triangleq A$ is $|EXX^T|(1-EX^T(EXX^T)^{-1}EX) > 0$. Therefore, the

symmetric matrix $\begin{pmatrix} EXX^T & EX \\ EX^T & 1 \end{pmatrix}$ is positive definite with its maximum and minimum eigenvalues denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. Because $\alpha(1-\alpha) \le g_j(t)(1-g_j(t)) \le \frac{1}{4}$, we get $\lambda_{\max}(\lim_{N\to\infty} R_g/N) \le \frac{1}{4}\lambda_{\max}(A)$ and $\lambda_{\min}(\lim_{N\to\infty} R_g/N) \ge \alpha(1-\alpha)\lambda_{\min}(A)$. Hence, if $\lim_{N\to\infty} \frac{1}{N} a(t) = o(e^{0.5-\varepsilon}(\Theta^*))$, we get

$$\lim_{N\to\infty} R_g^{-1} \sum_{t=1}^{N} a(t) = \lim_{N\to\infty} \left(\frac{1}{N} R_g\right)^{-1} \left(\frac{1}{N} \sum_{t=1}^{N} a(t)) = o(e^{0.5-\varepsilon}(\Theta^*)\right).$$

Analogously, if $\lim_{N\to\infty} R_g^{-1} a(t) = 0$, then $\lim_{N\to\infty} \frac{1}{N} a(t) = 0$.

(ii). $\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} X_t \Sigma_j^{*-1} X_t^T = \begin{pmatrix} \Sigma_j^{*-1} \otimes EXX^T & \Sigma_j^{*-1} \otimes EX \\ \Sigma_j^{*-1} \otimes EX^T & \Sigma_j^{*-1} \otimes 1 \end{pmatrix} \triangleq B.$

Because A and Σ_j^{*-1} are positive definite, it can be proved in a similar way as the proof of A that B is positive definite. By lemma 3, there is a positive number δ such that $\delta \le Eh_j(X)$. Let λ be any eigenvalue of $\lim_{N\to\infty} \frac{1}{N} R_j$; then we get $\delta\lambda_{\min}(B) \le \lambda \le \lambda_{\max}(B)$. Further, $\lambda_{\max}^{-1}(B) \le \lambda^{-1} \le \lambda_{\min}^{-1}(B)/\delta$. Therefore, if $\lim_{N\to\infty} \frac{1}{N} b(t) = o(e^{0.5-\varepsilon}(\Theta^*))$, we have

$$\lim_{N\to\infty} R_j^{-1} \sum_{t=1}^{N} b(t) = \lim_{N\to\infty} \left(\frac{1}{N} R_j\right)^{-1} \left(\frac{1}{N} \sum_{t=1}^{N} b(t)\right) = o(e^{0.5-\varepsilon}(\Theta^*)).$$

(iii). By lemma 3, there is a positive number δ such that $\lim_{N\to\infty} 1/(\frac{1}{N}\sum_{t=1}^{N} h_j(t)) \le 1/\delta$. Therefore, when $\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} c(t) = o(e^{0.5-\varepsilon}(\Theta^*))$, we have

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{\sum_{t=1}^{N} h_j(t)} \sum_{t=1}^{N} c(t) &= \lim_{N\to\infty} \frac{1}{\frac{1}{N}\sum_{t=1}^{N} h_j(t)} \left(\frac{1}{N} \sum_{t=1}^{N} c(t)\right) \\ &= o(e^{0.5-\varepsilon}(\Theta^*)). \end{aligned}$$

4 Main Theorems

With the explicit expressions of $P(\Theta^*)$ and $H(\Theta^*)$ with the training data set $\mathcal{S} = \{x^{(t)}, y^{(t)}\}_{t=1}^{N}$, we can obtain formulas for the blocks of $P(\Theta^*)H(\Theta^*)$, which can be written as follows. For notational convenience, we denote $[\theta_0, \theta_j, \Sigma_j] = [\theta_0^*, \theta_j^*, \Sigma_j^*]$ throughout this section.

$$P(\Theta^*)H(\Theta^*) = \operatorname{diag}[P_g, P_1, \ldots, P_K, P_{\Sigma_1}, \ldots, P_{\Sigma_K}]$$

$$\times \begin{pmatrix} H_{\theta_0,\theta_0^T} & H_{\theta_0,\theta_1^T} & \cdots & H_{\theta_0,\theta_K^T} & H_{\theta_0,\Sigma_1^T} & \cdots & H_{\theta_0,\Sigma_K^T} \\ H_{\theta_1,\theta_0^T} & H_{\theta_1,\theta_1^T} & \cdots & H_{\theta_1,\theta_K^T} & H_{\theta_1,\Sigma_1^T} & \cdots & H_{\theta_1,\Sigma_K^T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ H_{\theta_K,\theta_0^T} & H_{\theta_K,\theta_1^T} & \cdots & H_{\theta_K,\theta_K^T} & H_{\theta_K,\Sigma_1^T} & \cdots & H_{\theta_K,\Sigma_K^T} \\ H_{\Sigma_1,\theta_0^T} & H_{\Sigma_1,\theta_1^T} & \cdots & H_{\Sigma_1,\theta_K^T} & H_{\Sigma_1,\Sigma_1^T} & \cdots & H_{\Sigma_1,\Sigma_K^T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ H_{\Sigma_K,\theta_0^T} & H_{\Sigma_K,\theta_1^T} & \cdots & H_{\Sigma_K,\theta_K^T} & H_{\Sigma_K,\Sigma_1^T} & \cdots & H_{\Sigma_K,\Sigma_K^T} \end{pmatrix}$$

$$= \begin{pmatrix} P_g H_{\theta_0,\theta_0^T} & P_g H_{\theta_0,\theta_1^T} & \cdots & P_g H_{\theta_0,\theta_K^T} & P_g H_{\theta_0,\Sigma_1^T} & \cdots & P_g H_{\theta_0,\Sigma_K^T} \\ P_1 H_{\theta_1,\theta_0^T} & P_1 H_{\theta_1,\theta_1^T} & \cdots & P_1 H_{\theta_1,\theta_K^T} & P_1 H_{\theta_1,\Sigma_1^T} & \cdots & P_1 H_{\theta_1,\Sigma_K^T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ P_K H_{\theta_K,\theta_0^T} & P_K H_{\theta_K,\theta_1^T} & \cdots & P_K H_{\theta_K,\theta_K^T} & P_K H_{\theta_K,\Sigma_1^T} & \cdots & P_K H_{\theta_K,\Sigma_K^T} \\ P_{\Sigma_1} H_{\Sigma_1,\theta_0^T} & P_{\Sigma_1} H_{\Sigma_1,\theta_1^T} & \cdots & P_{\Sigma_1} H_{\Sigma_1,\theta_K^T} & P_{\Sigma_1} H_{\Sigma_1,\Sigma_1^T} & \cdots & P_{\Sigma_1} H_{\Sigma_1,\Sigma_K^T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ P_{\Sigma_K} H_{\Sigma_K,\theta_0^T} & P_{\Sigma_K} H_{\Sigma_K,\theta_1^T} & \cdots & P_{\Sigma_K} H_{\Sigma_K,\theta_K^T} & P_{\Sigma_K} H_{\Sigma_K,\Sigma_1^T} & \cdots & P_{\Sigma_K} H_{\Sigma_K,\Sigma_K^T} \end{pmatrix}.$$

Based on the expressions of the Hessian blocks (Gerald, 1980; Horn & Johnson, 1986) and P matrix through the IRLS approach, letting $g_j(t)$, $s_j(t)$, $f_j(t)$ denote $g_j(x^{(t)}, \theta_0)$, $s_j(x^{(t)}, \theta_0)$, $f_j(x^{(t)}, \theta_j)$ respectively, we have:

$$P_g H_{\theta_0,\theta_0^T} = \gamma_g (R_g)^{-1} \sum_{t=1}^{N} \sum_{j=1}^{K} \Big(h_j(t) \frac{\partial s_j(t)}{\partial \theta_0} - h_j(t) \sum_{l=1}^{K} h_l(t) \frac{\partial s_l(t)}{\partial \theta_0} - g_j(t) \frac{\partial s_j(t)}{\partial \theta_0} + g_j(t) \sum_{l=1}^{K} g_l(t) \frac{\partial s_l(t)}{\partial \theta_0} \Big) \frac{\partial s_j(t)}{\partial \theta_0^T},$$

$$P_g H_{\theta_0,\theta_j^T} = (R_g)^{-1} \sum_{t=1}^{N} \gamma_g h_j(t) \big[(y^{(t)} - f_j(t))^T \Sigma_j^{-1} X_t^T \big] \otimes \left[\frac{\partial s_j(t)}{\partial \theta_0} - \sum_{l=1}^{K} h_l(t) \frac{\partial s_l(t)}{\partial \theta_0} \right],$$

$$P_g H_{\theta_0,\Sigma_j^T} = -\frac{1}{2} (R_g)^{-1} \sum_{t=1}^{N} h_j(t) \gamma_g \mathrm{vec}^T \big[\Sigma_j^{-1} - U_j(t) \big]$$

$$\otimes\left[\frac{\partial s_j(t)}{\partial\theta_0}-\sum_{l=1}^{K}h_l(t)\frac{\partial s_l(t)}{\partial\theta_0}\right],$$

$$P_jH_{\theta_j,\theta_0^T}=(R_j)^{-1}\sum_{t=1}^{N}h_j(t)\left[X_t\Sigma_j^{-1}\left(y^{(t)}-f_j(t)\right)\right]$$

$$\otimes\left[\frac{\partial s_j(t)}{\partial\theta_0^T}-\sum_{l=1}^{K}h_l(t)\frac{\partial s_l(t)}{\partial\theta_0^T}\right],$$

$$P_jH_{\theta_j,\theta_i^T}=(R_j)^{-1}\sum_{t=1}^{N}\gamma_{ij}(t)[(y^{(t)}-f_i(t))^T\Sigma_i^{-1}X_t^T]$$

$$\otimes[X_t\Sigma_j^{-1}(y^{(t)}-f_j(t))]-\delta_{ij}I,$$

$$P_jH_{\theta_j,\Sigma_i^T}=-\frac{1}{2}(R_j)^{-1}\sum_{t=1}^{N}\gamma_{ij}(t)\mathrm{vec}[\Sigma_i^{-1}-U_i(t)]^T$$

$$\otimes\left[X_t\Sigma_j^{-1}(y^{(t)}-f_j(t)\right]-\frac{1}{2}(R_j)^{-1}\sum_{t=1}^{N}\delta_{ij}h_j(t)$$

$$\times\{[(y^{(t)}-f_j(t))^T\Sigma_j^{-1}]\otimes(X_t\Sigma_j^{-1})$$

$$+(X_t\Sigma_j^{-1})\otimes\left[\left(y^{(t)}-f_j(t)\right)^T\Sigma_j^{-1}\right]\},$$

$$P_{\Sigma_j}H_{\Sigma_j,\theta_0^T}=-\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}(\Sigma_j\otimes\Sigma_j)\mathrm{vec}[\Sigma_j^{-1}-U_j(t)]$$

$$\otimes\left[h_j(t)\frac{\partial s_j(t)}{\partial\theta_0}-h_j(t)\sum_{l=1}^{K}h_l\frac{\partial s_l(t)}{\partial\theta_0}\right]^T,$$

$$P_{\Sigma_j}H_{\Sigma_j,\theta_i^T}=-\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}\gamma_{ij}(t)(\Sigma_j\otimes\Sigma_j)\{\mathrm{vec}\left[\Sigma_j^{-1}-U_j(t)\right]$$

$$\otimes\left[\left(y^{(t)}-f_i(t)\right)^T\Sigma_i^{-1}X_t^T\right]\}-\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}\delta_{ij}h_j(t)$$

$$(\Sigma_j\otimes\Sigma_j)\{\left[\Sigma_j^{-1}\left(y^{(t)}-f_j(t)\right)\right]\otimes\left[\Sigma_j^{-1}X_t^T\right]$$

$$+[\Sigma_j^{-1}X_t^T]\otimes[\Sigma_j^{-1}(y^{(t)}-f_j(t))]\},$$

$$P_{\Sigma_j}H_{\Sigma_j,\Sigma_i^T}=-\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}\delta_{ij}h_j(t)(\Sigma_j\otimes\Sigma_j)(\mathrm{vec}^T[I_m]\otimes I_m\otimes I_m)$$

$$\times (I_m \otimes M(t) \otimes I_m)(I_m \otimes I_m \otimes \mathrm{vec}[I_m])$$
$$+ \frac{1}{2} \frac{1}{\sum_{t=1}^{N} h_j(t)} \sum_{t=1}^{N} \gamma_{ij}(t)(\Sigma_j \otimes \Sigma_j)\mathrm{vec}^T[\Sigma_i^{-1} - U_i(t)]$$
$$\otimes \mathrm{vec}[\Sigma_j^{-1} - U_j(t)],$$

where I_d is the dth-order identity matrix and

$$U_j(t) = \Sigma_j^{-1}\big(y^{(t)} - f_j(t)\big)\big(y^{(t)} - f_j(t)\big)^T \Sigma_j^{-1},$$
$$M(t) = \frac{\partial \Sigma_j^{-1}}{\partial \Sigma_j} - \frac{\partial \Sigma_j^{-1}}{\partial \Sigma_j}\big[\big(y^{(t)} - f_j(t)\big)\big(y^{(t)} - f_j(t)\big)^T \Sigma_j^{-1}\big] \otimes I_m$$
$$- I_m \otimes [\Sigma_j^{-1}(y^{(t)} - f_j(t))(y^{(t)} - f_j(t))^T] \frac{\partial \Sigma_j^{-1}}{\partial \Sigma_j}. \tag{4.1}$$

We now have our first theorem on the EM algorithm for ME through the IRLS approach as follows:

Theorem 1. *Given i.i.d. sample data $\{x^{(t)}, y^{(t)}\}_1^N$ from a mixture of K expert networks of parameters Θ^* with the help of $P(x)$ constrained on a bounded, closed set $\mathcal{D}$, that is, $\int_{\mathcal{D}} P(x)dx = 1$, for any $x \in \mathcal{D}$, $P(x) > 0$ and Θ^* as well as x satisfy conditions 1 to 3. When $e(\Theta^*)$ is considered as an infinitesimal, as it tends to zero, for the EM algorithm for ME through the IRLS approach, we have*

$$\lim_{N\to\infty} P(\Theta^*)H(\Theta^*)$$
$$= \begin{pmatrix} -\gamma_g \widetilde{P}_g G + o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) - I_{K^2(n+1)} & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) - I_{Km^2} \end{pmatrix},$$

where ε is an arbitrarily small positive number, and $\widetilde{P}_g \triangleq \lim_{N\to\infty}(R_g/N)^{-1}$,

$$G \triangleq \lim_{N\to\infty} \frac{1}{N}$$
$$\times \begin{pmatrix} \sum_{t=1}^{N}(g_1(t) - g_1^2(t)) \frac{\partial s_1(t)}{\partial \theta_{01}} \frac{\partial s_1(t)}{\partial \theta_{01}^T} & \cdots & -\sum_{t=1}^{N} g_1(t) g_{K-1}(t) \frac{\partial s_1(t)}{\partial \theta_{01}} \frac{\partial s_{K-1}(t)}{\partial \theta_{0(K-1)}^T} \\ \vdots & \vdots & \vdots \\ -\sum_{t=1}^{N} g_{K-1}(t) g_1(t) \frac{\partial s_{K-1}(t)}{\partial \theta_{0(K-1)}} \frac{\partial s_1(t)}{\partial \theta_{01}^T} & \cdots & \sum_{t=1}^{N}(g_{K-1}(t) - g_{K-1}^2(t)) \frac{\partial s_{K-1}(t)}{\partial \theta_{0(K-1)}} \frac{\partial s_{K-1}(t)}{\partial \theta_{0(K-1)}^T} \end{pmatrix}.$$

Accordingly, we have an upper bound for the asymptotic convergence rate of the EM algorithm:

$$r \leq ||I - \gamma_g \widetilde{P}_g G|| + o(e^{0.5-\varepsilon}(\Theta^*)). \tag{4.2}$$

Proof. Letting $e_1 = (1, 0, \ldots, 0)^T, \ldots, e_{K-1} = (0, 0, \ldots, 1)^T$ denote the canonical basis vectors of $\mathbb{R}^{K-1}$, by equation 2.4, we then have

$$\frac{\partial s_j(t)}{\partial \theta_0} = e_j \otimes [(x^{(t)})^T, 1]^T.$$

For convenience of notation, we let x_t denote $[(x^{(t)})^T; 1]^T$; we thus have $\partial s_j(t)/\partial \theta_0 = e_j \otimes x_t$. Since we always set $s_K(t) = 0$, we have $\partial s_K(t)/\partial \theta_0 = 0$. We begin to consider the block $P_g H_{\theta_0,\theta_0^T}$. By lemmas 2 and 4, we have:

$$\begin{aligned}
&\lim_{N\to\infty} \gamma_g (R_g)^{-1} \sum_{t=1}^{N} \sum_{j=1}^{K} (h_j(t) \frac{\partial s_j(t)}{\partial \theta_0} - h_j(t) \sum_{l=1}^{K} h_l(t) \frac{\partial s_l(t)}{\partial \theta_0}) \frac{\partial s_j(t)}{\partial \theta_0^T} \\
&= \lim_{N\to\infty} \gamma_g (R_g)^{-1} \sum_{t=1}^{N} \begin{pmatrix} (h_1(t) - h_1^2(t)) x_t x_t^T & \cdots & -h_1(t) h_{K-1}(t) x_t x_t^T \\ \vdots & \ddots & \vdots \\ -h_{K-1}(t) h_1(t) x_t x_t^T & \cdots & (h_{K-1}(t) - h_{K-1}^2(t)) x_t x_t^T \end{pmatrix} \\
&= \lim_{N\to\infty} \gamma_g (R_g)^{-1} \sum_{t=1}^{N} \begin{pmatrix} \sum_{l=2}^{K-1} \gamma_{1l}(t) x_t x_t^T & \cdots & -\gamma_{1(K-1)}(t) x_t x_t^T \\ \vdots & \ddots & \vdots \\ -\gamma_{1(K-1)}(t) x_t x_t^T & \cdots & \sum_{l=1}^{K-2} \gamma_{(K-1)l}(t) x_t x_t^T \end{pmatrix} \\
&= o(e^{0.5-\varepsilon}(\Theta^*)).
\end{aligned}$$

Hence, $P_g H_{\theta_0,\theta_0^T} = -\gamma_g \widetilde{P}_g G + o(e^{0.5-\varepsilon}(\Theta^*))$.

We then consider the block $P_g H_{\theta_0,\theta_j^T}$. The elements of $\partial s_j(t)/\partial \theta_0 - \sum_{l=1}^{K} h_l(t) \partial s_l(t)/\partial \theta_0$ are $-h_l(t) x_t$ $(l \neq j)$, or $(1 - h_j(t)) x_t$. By lemma 2, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \gamma_g h_j(t) \left[(y^{(t)} - f_j(t))^T \Sigma_j^{-1} X_t^T \right] \\
&\quad \otimes \left[\frac{\partial s_j(t)}{\partial \theta_0} - \sum_{l=1}^{K} h_l(t) \frac{\partial s_l(t)}{\partial \theta_0} \right] = o(e^{0.5-\varepsilon}(\Theta^*)),
\end{aligned}$$

and by lemma 4, we further get

$$\lim_{N\to\infty} P_g H_{\theta_0,\theta_j^T} = o(e^{0.5-\varepsilon}(\Theta^*)).$$

Analogously, we can get

$$\lim_{N\to\infty} P_g H_{\theta_0,\Sigma_j^T} = o(e^{0.5-\varepsilon}(\Theta^*)),$$

$$\lim_{N\to\infty} P_j H_{\theta_j,\theta_0^T} = o(e^{0.5-\varepsilon}(\Theta^*)),$$

$$\lim_{N\to\infty} P_j H_{\theta_j,\theta_i^T} = o(e^{0.5-\varepsilon}(\Theta^*)) - \delta_{ij} I,$$

$$\lim_{N\to\infty} -\frac{1}{2}\sum_{t=1}^{N} \gamma_{ij}(t)(R_j)^{-1}\mathrm{vec}[\Sigma_i^{-1} - U_i(t)]^T \otimes [X_t \Sigma_j^{-1}(y^{(t)} - f_j(t)] = o(e^{0.5-\varepsilon}(\Theta^*)),$$

$$\lim_{N\to\infty} P_{\Sigma_j} H_{\Sigma_j,\theta_0^T} = o(e^{0.5-\varepsilon}(\Theta^*)),$$

$$\lim_{N\to\infty} -\frac{1}{\sum_{t=1}^{N} h_j(t)}\sum_{t=1}^{N} \gamma_{ij}(t)(\Sigma_j \otimes \Sigma_j)\{\mathrm{vec}\left[\Sigma_j^{-1} - U_j(t)\right] \otimes [(y^{(t)} - f_i(t))^T \Sigma_i^{-1} X_t^T]\} = o(e^{0.5-\varepsilon}(\Theta^*)),$$

$$\lim_{N\to\infty} \frac{1}{2}\frac{1}{\sum_{t=1}^{N} h_j(t)}\sum_{t=1}^{N} \gamma_{ij}(t)(\Sigma_j \otimes \Sigma_j)\mathrm{vec}^T\left[\Sigma_i^{-1} - U_i(t)\right] \otimes \mathrm{vec}[\Sigma_j^{-1} - U_j(t)] = o(e^{0.5-\varepsilon}(\Theta^*)).$$

We further consider the block $\frac{1}{2}(R_j)^{-1}\sum_{t=1}^{N}\delta_{ij}h_j(t)[(y^{(t)} - f_j(t))^T\Sigma_j^{-1}] \otimes (X_t\Sigma_j^{-1})$. According to the EM iteration, equation 2.7, we have

$$\lim_{N\to\infty} R_j^{-1}\sum_{t=1}^{N} h_j(t) X_t \Sigma_j^{-1}(y^{(t)} - f_j(t)) = 0.$$

Suppose that $\Sigma_j^{-1} = (\sigma_{kl}), k, l = 1, \ldots, m$, and $y^{(t)} - f_j(t) \triangleq z(t)$. We then have

$$h_j(t) X_t \Sigma_j^{-1}(y^{(t)} - f_j(t)) = h_j(t)\Bigg[\sum_{l=1}^{m} x_1^{(t)}\sigma_{1l}z_l(t), \ldots, \sum_{l=1}^{m} x_n^{(t)}\sigma_{1l}z_l(t), \sum_{l=1}^{m} x_1^{(t)}\sigma_{2l}z_l(t), \ldots, \sum_{l=1}^{m} x_n^{(t)}\sigma_{2l}z_l(t), \ldots, \sum_{l=1}^{m} x_n^{(t)}\sigma_{ml}z_l(t), \sum_{l=1}^{m}\sigma_{1l}z_l(t), \ldots, \sum_{l=1}^{m}\sigma_{ml}z_l(t)\Bigg]^T.$$

By lemma 4, we also have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}h_j(t)\sum_{l=1}^{m}x_d^{(t)}\sigma_{kl}z_l(t)=0,\ \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}h_j(t)\sum_{l=1}^{m}\sigma_{kl}z_l(t)=0 \tag{4.3}$$

for all $d\in\{1,\dots,n+1\}, k\in\{1,\dots,m\}$. Since

$$\begin{aligned}&h_j(t)\left[(y^{(t)}-f_j(t))^T\Sigma_j^{-1}\right]\otimes(X_t\Sigma_j^{-1})\\&\quad=h_j(t)\left[\sum_{i=1}^{m}z_i(t)\sigma_{1i},\dots,\sum_{i=1}^{m}z_i(t)\sigma_{mi}\right]^T\otimes(X_t\Sigma_j^{-1}),\end{aligned}$$

$$X_t\Sigma_j^{-1}=\left(\begin{array}{cccccccc|ccc}\sigma_{11}x_1^{(t)}&\cdots&\sigma_{11}x_n^{(t)}&\cdots&\sigma_{m1}x_1^{(t)}&\cdots&\sigma_{m1}x_n^{(t)}&\sigma_{11}&\cdots&\sigma_{m1}\\\sigma_{12}x_1^{(t)}&\cdots&\sigma_{12}x_n^{(t)}&\cdots&\sigma_{m2}x_1^{(t)}&\cdots&\sigma_{m2}x_n^{(t)}&\sigma_{12}&\cdots&\sigma_{m2}\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\\sigma_{1m}x_1^{(t)}&\cdots&\sigma_{1m}x_n^{(t)}&\cdots&\sigma_{mm}x_1^{(t)}&\cdots&\sigma_{mm}x_n^{(t)}&\sigma_{1m}&\cdots&\sigma_{mm}\end{array}\right)^T,$$

the elements of $h_j(t)[(y^{(t)}-f_j(t))^T\Sigma_j^{-1}]\otimes(X_t\Sigma_j^{-1})$ take the forms of $h_j(t)\sum_{i=1}^{m}z_i(t)\sigma_{pi}\sigma_{kl}x_d^{(t)}$ and $h_j(t)\sum_{i=1}^{m}z_i(t)\sigma_{pi}\sigma_{kl}$, where $d\in\{1,\dots,n\}$, $p,k,l\in\{1,\dots,m\}$. According to equation 4.3, we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}h_j(t)\sum_{i=1}^{m}z_i(t)\sigma_{pi}\sigma_{kl}x_d^{(t)}=0$$

and

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}h_j(t)\sum_{i=1}^{m}z_i(t)\sigma_{pi}\sigma_{kl}=0.$$

Therefore, we have

$$\lim_{N\to\infty}\frac{1}{2}(R_j)^{-1}\sum_{t=1}^{N}\delta_{ij}h_j(t)\left[(y^{(t)}-f_j(t))^T\Sigma_j^{-1}\right]\otimes\left(X_t\Sigma_j^{-1}\right)=0.$$

Similarly, we also have

$$\lim_{N\to\infty}\frac{1}{2}\sum_{t=1}^{N}\delta_{ij}h_j(t)(R_j)^{-1}\left(X_t\Sigma_j^{-1}\right)\otimes\left[(y^{(t)}-f_j(t))^T\Sigma_j^{-1}\right]=0,$$

$$\begin{aligned}\lim_{N\to\infty}&\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}\delta_{ij}h_j(t)(\Sigma_j\otimes\Sigma_j)\left\{\left[\Sigma_j^{-1}(y^{(t)}-f_j(t))\right]\right.\\&\otimes\left[\Sigma_j^{-1}X_t^T\right]+\left[\Sigma_j^{-1}X_t^T\right]\\&\left.\otimes\left[\Sigma_j^{-1}(y^{(t)}-f_j(t))\right]\right\}=0.\end{aligned}$$

By the EM iteration, equation 2.8, we have

$$\lim_{N\to\infty}\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}h_j(t)\left[\Sigma_j-(y^{(t)}-f_j(t))(y^{(t)}-f_j(t))^T\right]=0,$$

which helps us get the following limitation about equation 4.1:

$$\lim_{N\to\infty}\frac{1}{\sum_{t=1}^{N}h_j(t)}\sum_{t=1}^{N}h_j(t)M(t)=-\frac{\partial\Sigma_j^{-1}}{\partial\Sigma_j}.$$

In this way, we get

$$\begin{aligned}&\lim_{N\to\infty}-\frac{1}{\sum_{t=1}^{N}h_j(t)}(\Sigma_j\otimes\Sigma_j)\\&\quad\times\sum_{t=1}^{N}h_j(t)(\mathrm{vec}^T[I_m]\otimes I_m\otimes I_m)(I_m\otimes M(t)\otimes I_m)(I_m\otimes I_m\otimes\mathrm{vec}[I_m])\\&\quad=-(\Sigma_j\otimes\Sigma_j)(\Sigma_j^{-1}\otimes\Sigma_j^{-1})=-I.\end{aligned}$$

Summing up all the results, we obtain:

$$\begin{aligned}&\lim_{N\to\infty}P(\Theta^*)H(\Theta^*)\\&=\begin{pmatrix}-\gamma_g\widetilde{P}_gG+o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*))\\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*))-I_{K^2(n+1)} & o(e^{0.5-\varepsilon}(\Theta^*))\\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*))-I_{Km^2}\end{pmatrix},\end{aligned}$$

According to equation 2.14 and the norm inequality, we finally have

$$\begin{aligned} r &\le \lim_{N\to\infty} ||I + P(\Theta^*)H(\Theta^*)|| = ||I + \lim_{N\to\infty} P(\Theta^*)H(\Theta^*)|| \\ &= \left|\left| \begin{pmatrix} I - \gamma_g \widetilde{P}_g G + o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \end{pmatrix} \right|\right| \\ &= ||I - \gamma_g P_g G|| + o(e^{0.5-\varepsilon}(\Theta^*)). \end{aligned}$$

Remark 1. According to equation 4.2, the asymptotic convergence rate of the EM algorithm for ME through the IRLS approach is generally bounded by a positive number $||I - \gamma_g P_g G||$ since $\gamma_g P_g G \neq I$ in general, even if the average overlap measure of the ME model tends to zero. So we can consider that the EM algorithm through the IRLS approach maintains a linear convergence rate around the true solution with a large sample.

We further consider the EM algorithm for ME through the Newton-Raphson approach and have our second theorem as follows:

Theorem 2. *Under the same assumptions as stated in theorem 1, for the EM algorithm for ME through the Newton-Raphson approach, we have*

$$\begin{aligned} &\lim_{N\to\infty} P(\Theta^*)H(\Theta^*) \\ &= \begin{pmatrix} -\alpha I_{(K-1)(n+1)} + o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) - I_{K^2(n+1)} & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) - I_{Km^2} \end{pmatrix}, \end{aligned}$$

where ε is an arbitrarily small positive number. Accordingly, we have an upper bound of the asymptotic convergence rate of the EM algorithm:

$$r \le |1 - \alpha| + o(e^{0.5-\varepsilon}(\Theta^*)). \tag{4.4}$$

Proof. As compared to the proof of theorem1, we need only to compute $P_{\theta_0} H$. For the Newton-Raphson method, $P_{\theta_0} = -\alpha H_g$, where H_g is the Hessian matrix of equation 2.10. Actually, H_g has the following expression:

$$H_g = \sum_{t=1}^{N} \sum_{j=1}^{K} (-g_j(t) \frac{\partial s_j(t)}{\partial \theta_0} + g_j(t) \sum_{l=1}^{K} g_l(t) \frac{\partial s_l(t)}{\partial \theta_0}) \frac{\partial s_j(t)}{\partial \theta_0^T}.$$

In comparison with the expression of H_{θ_0,θ_0^T}, we find that H_g is just the sum of the last two terms of H_{θ_0,θ_0^T}. Then we have

$$\begin{aligned}
&\lim_{N\to\infty} -\alpha H_g^{-1} H_{\theta_0,\theta_0^T} \\
&= \lim_{N\to\infty} -\alpha H_g^{-1} \sum_{t=1}^{N}\sum_{j=1}^{K}(h_j(t)\frac{\partial s_j(t)}{\partial\theta_0} - h_j(t)\sum_{l=1}^{K} h_l(t)\frac{\partial s_l(t)}{\partial\theta_0})\frac{\partial s_j(t)}{\partial\theta_0^T} - \alpha \\
&= \lim_{N\to\infty} -\alpha \left(\frac{1}{N}H_g\right)^{-1} \\
&\quad \times \frac{1}{N}\sum_{t=1}^{N}\begin{pmatrix} \sum_{l=2}^{K-1}\gamma_{1l}(t)x_t x_t^T & \cdots & -\gamma_{1(K-1)}(t)x_t x_t^T \\ \vdots & \ddots & \vdots \\ -\gamma_{1(K-1)}(t)x_t x_t^T & \cdots & \sum_{l=1}^{K-2}\gamma_{(K-1)l}(t)x_t x_t^T \end{pmatrix} - \alpha.
\end{aligned}$$

By lemma 2, we further have

$$\lim_{N\to\infty} -\alpha H_g^{-1} H_{\theta_0,\theta_0^T} = -\alpha + o(e^{0.5-\varepsilon}(\Theta^*)).$$

Thus, we have

$$\lim_{N\to\infty} ||I + P_{\theta_0} H_{\theta_0,\theta_0^T}|| = 1 - \alpha + o(e^{0.5-\varepsilon}(\Theta^*)).$$

Since the rest of the proof is identical to that of theorem 1, we therefore have

$$\begin{aligned}
&\lim_{N\to\infty} P(\Theta^*)H(\Theta^*) \\
&= \begin{pmatrix} -\alpha I_{(K-1)(n+1)} + o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) - I_{K^2(n+1)} & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) - I_{Km^2} \end{pmatrix}.
\end{aligned}$$

According to equation 2.14 and the norm inequality, we finally have

$$\begin{aligned}
r &\le \lim_{N\to\infty} ||I + P(\Theta^*)H(\Theta^*)|| = ||I + \lim_{N\to\infty} P(\Theta^*)H(\Theta^*)|| \\
&= \left|\left|\begin{pmatrix} (1-\alpha)I_{(K-1)(n+1)} + o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \\ o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) & o(e^{0.5-\varepsilon}(\Theta^*)) \end{pmatrix}\right|\right| \\
&= |1-\alpha| + o(e^{0.5-\varepsilon}(\Theta^*)).
\end{aligned}$$

Specifically for the Newton approach with $\alpha = 1$, by theorem 2, we have the following corollary.

Corollary 1. *Under the same assumptions as stated in theorem 2, for the EM algorithm for ME through the Newton approach, we have an upper bound of its asymptotic convergence rate:*

$$r = o(e^{0.5-\varepsilon}(\Theta^*)). \tag{4.5}$$

Remark 2. Corollary 1 has proved that the asymptotic convergence rate of the EM algorithm for ME through the Newton approach locally around the true solution Θ^* tends to zero as the average overlap measure $e(\Theta^*)$ tends to zero. In other words, the large sample local convergence rate for the EM algorithm tends to be asymptotically superlinear when $e(\Theta^*)$ tends to zero.

5 Experimental Results

To substantiate our theoretical results on the asymptotic convergence of the EM algorithm for ME through the Newton or IRLS approach, we implement the EM algorithm on two groups of synthetic data sets with attenuating measures of overlap among the expert distributions. We first consider the data set from a mixture of two experts: $K = 2$. The experts are two line segments with noises—$y = a_1 x + a_2 + n_t$ for $x \in [x_L, x_U]$ and $y = b_1 x + b_2 + n_t$ for $x \in [x_L^{'}, x_U^{'}]$, where $n_t \sim \mathcal{N}(0, \sigma^2)$ denotes a gaussian distribution with zero mean and variance σ^2. In order to make the average overlap measure between two experts attenuate to zero, we push the two intervals $[x_L, x_U]$ and $[x_L^{'}, x_U^{'}]$ away along the x-axis. In our experiments, we let $a_1 = 1$, $a_2 = -1$, $b_1 = -1$, $b_2 = 1$, $\sigma_1^2 = \sigma_2^2 = 0.4$, $[x_L, x_U] = [-2, 1] - m_x$ and $[x_L^{'}, x_U^{'}] = [1, 4] + m_x$, where the variable m_x increases from -0.5 to 1.25. Along each noisy line segment or expert in the mixture, we generate 5000 i.i.d. samples. Typically we select $m_x = -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25$, and establish eight data sets denoted by $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_8$, respectively. For illustration, four of them are sketched in Figure 1. Obviously as m_x increases gradually, the average overlap measure (AOM) of the two experts attenuates to zero (see Figure 2). We run the EM algorithm through the Newton approach on the eight data sets 50 times with different randomly initialized parameters, and the algorithm is terminated when the change of the log-likelihood function between two epochs is less than 10^{-5}. We compute the absolute errors between the average estimated parameters and the corresponding true parameters: $\Delta_{ij} = |\theta_j^i - \theta^{*i}_j|$, $\Delta^j = |\sigma_j^2 - \sigma_j^{*2}|$. The experimental results of the EM algorithm through the Newton approach on those eight data sets are listed in Table 1.

It can be seen from Table 1 that as the AOM of a data set falls from a considerable value (i.e., 0.09), the parameter estimation becomes more accurate. Specifically, the accuracy rate of parameter estimation on $\mathcal{S}_2$ is higher than that on $\mathcal{S}_1$, and the accuracy rate of parameter estimation on $\mathcal{S}_4$ is higher than that on $\mathcal{S}_3$. However, as the AOM moves closer to zero,

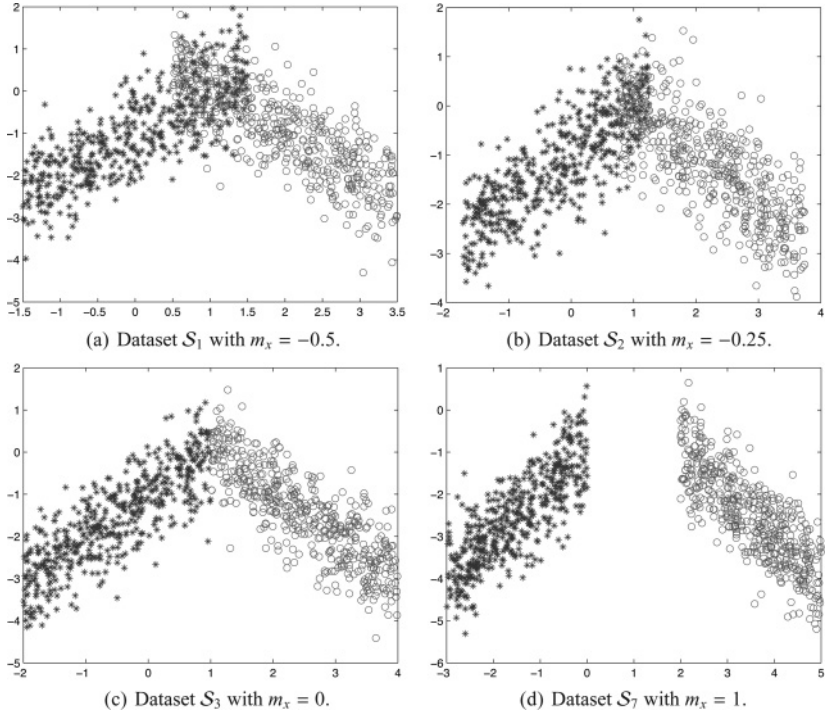


Figure 1: Sketches of four typical data sets of the first group with attenuating AOMs. The notations $*$ and o represent samples from the two classes or experts, respectively. (a, b, c, d). Sketches of the data sets $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, and $\mathcal{S}_7$, respectively, where the overlap measure of $\mathcal{S}_1$ is the largest and that of $\mathcal{S}_7$ is the smallest, being close to zero.

the accuracy rate of parameter estimation remains almost the same, but the number of epochs for the convergence of the EM algorithm decreases considerably. That is, the EM algorithm converges at a higher speed as the AOM decreases. This result is consistent with our theoretical result that the large sample local convergence rate for the EM algorithm tends to be asymptotically superlinear as $e(\Theta^*)$ is close to zero.

We further implement the EM algorithms through both the IRLS and Newton approaches on the second group of five synthetic three-category data sets with attenuating AOMs (shown in Figure 3 and denoted by $\mathcal{S}_a, \ldots, \mathcal{S}_e$, respectively), where data points in each data set are generated from a mixture of three gaussian distributions centered at $[-\gamma, 0]$, $[0, \gamma]$, $[\gamma, 0]$, respectively, with $\gamma > 0$ dominating its AOM. In each data set, we generate 1000 i.i.d. samples from a gaussian distribution. In the first four data sets, the gaussian distributions have a common covariance matrix [0.5,

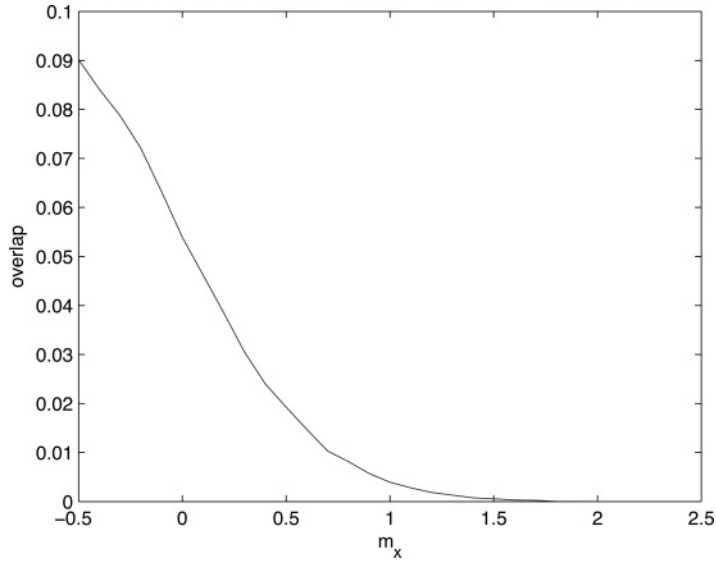


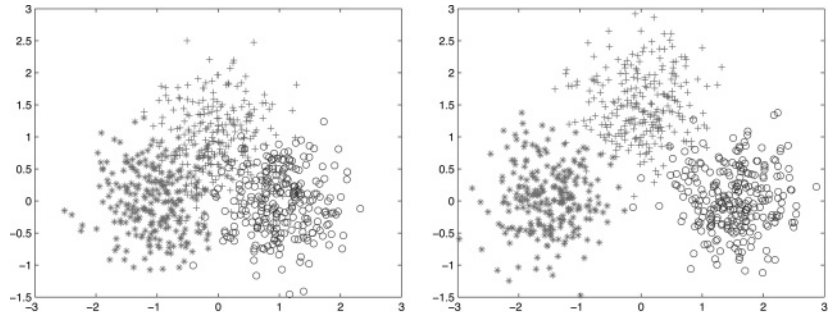
Figure 2: Sketch of the AOM of the two experts with respect to m_x.

Table 1: Experimental Results of the EM Algorithm for ME Through the Newton Approach on Eight Synthetic Data Sets with Attenuating AOMs.

Data Set	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	$\mathcal{S}_6$	$\mathcal{S}_7$	$\mathcal{S}_8$
m_x	−0.5	−0.25	0	0.25	0.5	0.75	1	1.25
AOM	0.090	0.075	0.055	0.035	0.019	0.009	0.004	0.001
Δ_{11}	0.029	0.022	0.036	0.008	0.017	0.004	0.009	0.009
Δ_{12}	0.089	0.011	0.012	0.019	0.007	0.001	0.002	0.06
Δ_{21}	0.042	0.013	0.014	0.001	0.038	0.010	0.001	0.020
Δ_{22}	0.242	0.015	0.034	0.034	0.006	0.003	0.001	0.029
Δ^1	0.002	0.011	0.011	0.014	0.012	0.012	0.003	0.005
Δ^2	0.001	0.001	0.018	0.008	0.018	0.001	0.003	0.004
LLF	−0.9946	−0.9758	−0.9601	−0.9756	−0.957	−0.9537	−0.9494	−0.9549
CR	0.924	0.973	0.956	0.908	0.860	0.605	0.324	0.033
Epochs	24.6	31.6	32.6	35.9	27.1	23.2	20.84	20.2

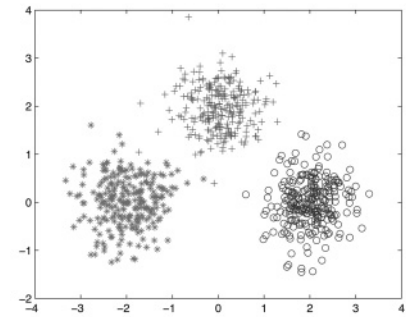
Note: Epochs denotes the number of epochs the EM algorithm has taken before stop; CR denotes the convergence rate; which is the maximum eigenvalue of the matrix $I + P(\Theta)H(\Theta)$; and LLF denotes the obtained log-likelihood function on a given data set.

0; 0, 0.5], where $\gamma = 1, 1.5, 2, 2.5$, respectively. As to the fifth data set $\mathcal{S}_e$, the gaussian distributions keep the same centers as those of $\mathcal{S}_d$ but have a different covariance matrix [0.3, 0; 0, 0.3]. This three-category problem is used to evaluate the convergence performance of the EM algorithms through

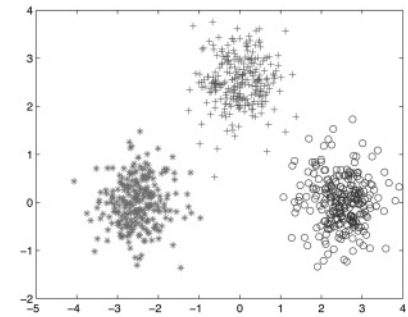


(a) Dataset $\mathcal{S}_a$ with the AOM being 0.0668.

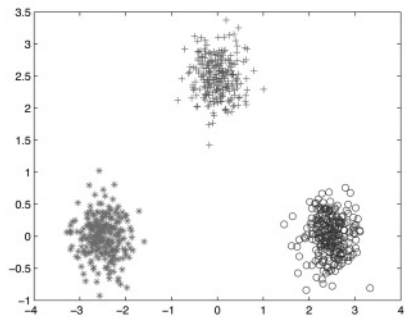
(b) Dataset $\mathcal{S}_b$ with the AOM being 0.0261.



(c) Dataset $\mathcal{S}_c$ with the AOM being 0.0023.



(d) Dataset $\mathcal{S}_d$ with the AOM being 0.001.



(e) Dataset $\mathcal{S}_e$ with the AOM being 10^{-13}.

Figure 3: Sketches of five typical data sets of the second group with attenuating AOMs. Data points of each data set are generated from a mixture of three gaussian distributions and denoted by $*$, $+$, and o, respectively.

both the IRLS and Newton approaches on the data sets with attenuating AOMs.

In the experiments, an ME architecture consisting of three experts is adopted, and the learning rates for the two approaches are both set to 1.

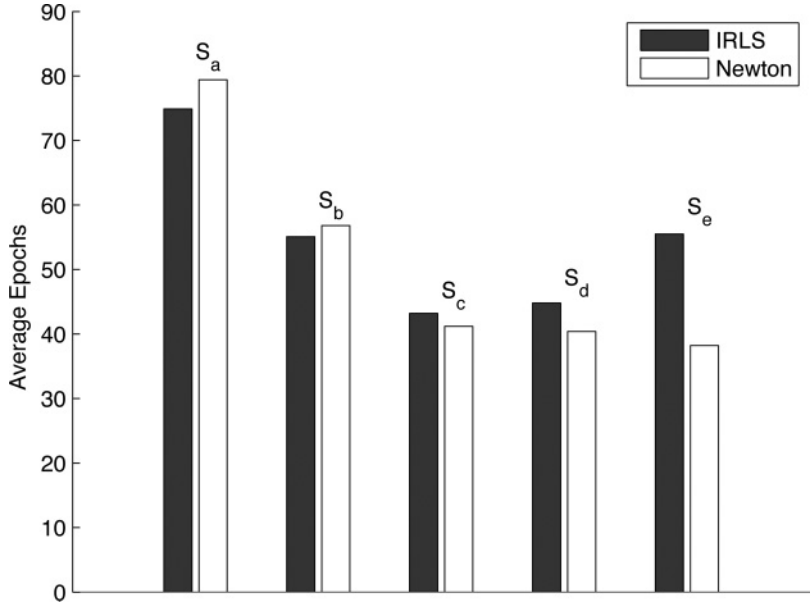


Figure 4: Average number of epochs taken for the convergence of the EM algorithm through either the IRLS or Newton approach on $\mathcal{S}_a, \cdots, \mathcal{S}_e$, respectively.

In the same way, we run the EM algorithms through both the IRLS and Newton approaches 50 times with different randomly initialized parameters, and the algorithms stop when the change of the log-likelihood function between two epochs is less than 10^{-5}. The average numbers of epochs taken for the convergence of the EM algorithm through either the IRLS or Newton approach on these data sets is illustrated in Figure 4. It can be observed that on the data sets with a relatively large overlap (e.g., $\mathcal{S}_a$, $\mathcal{S}_b$), the EM algorithm through the IRLS approach may converge a bit faster than the EM algorithm through the Newton approach. However, the EM algorithm through the Newton approach converges much faster than the EM algorithm through the IRLS approach on the lower AOM data sets (e.g., $\mathcal{S}_c, \mathcal{S}_d, \mathcal{S}_e$). In fact, on $\mathcal{S}_e$ whose AOM is very close to zero, the EM algorithm through the Newton approach is terminated after about 38 epochs, while the EM algorithm through the IRLS approach is terminated after about 55 epochs (see Figure 5). The experimental results are consistent with our theoretical results on the asymptotic convergence rates of the EM algorithms through the IRLS and Newton approaches. As the AOM tends to zero, the asymptotic convergence rate of the Newton approach also tends to zero. However, the asymptotic convergence rate of the IRLS approach does not tend to zero because the projection matrix for the IRLS approach is different

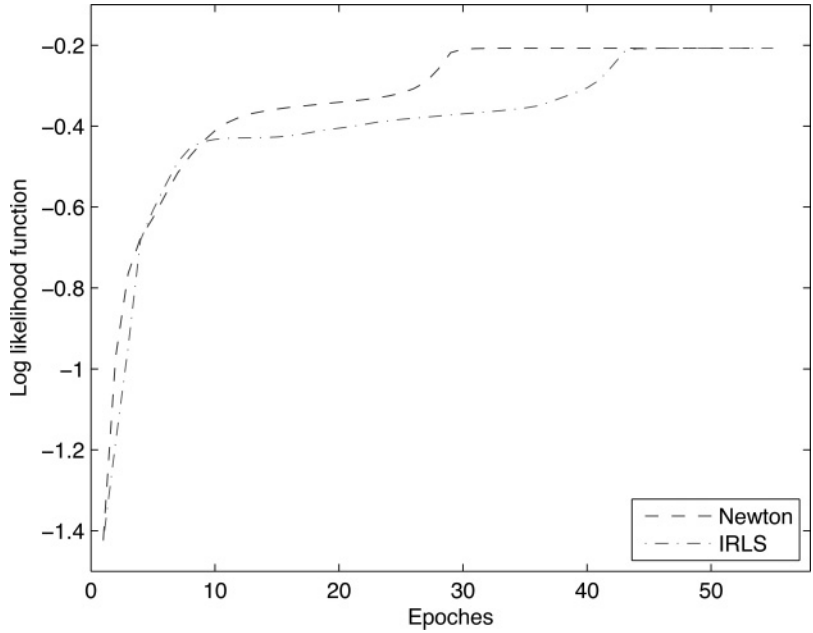


Figure 5: Evolution sketch of the log-likelihood function with respect to the number of epochs during the EM iterations on data set $\mathcal{S}_e$, where the dashed and dot-dash lines represent values of the log-likelihood functions of the Newton and IRLS approaches, respectively.

from that for the Newton approach. This is why the IRLS approach could not converge as fast as the Newton approach when the AOM is very small.

6 Conclusion

We have presented an analysis on the asymptotic convergence rate of the EM algorithm for the mixture-of-experts architecture through the IRLS and Newton-Raphson approaches. By introducing the average overlap measure of the ME architecture, we obtain an upper bound of the asymptotic convergence rate of the EM algorithm for both approaches. Specifically, for the Newton approach with the large sample, when the average overlap tends to zero, the EM algorithm tends to converge superlinearly. Moreover, these theoretical results are demonstrated by simulation experiments.

Acknowledgments

This work was supported by the Natural Science Foundation of China, grant 60771061.

References

Chen, K., Xu, L., & Chi, H. (1999). Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks, 12*, 1229–1252.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39*, 1–38.

Gerald, S. R. (1980). *Matrix derivatives*. New York: Dekker.

Horn, R. A., & Johnson, C. R. (1986). *Matrix analysis*. Cambridge: Cambridge University Press.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.

Jordan, M. I., & Jacobs, R. A. (1992). Hierarchies of adaptive experts. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing system, 4* (pp. 985–992). San Francisco: Morgan Kaufmann.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation, 6*, 181–214.

Jordan, M. I., & Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Computation, 8*, 1409–1431.

Ma, J., & Fu, S. (2005). On the correct convergence of the EM algorithm for gaussian mixtures. *Pattern Recognition, 38*, 2602–2611.

Ma, J., & Xu, L. (2005). Asymptotic convergence properties of the EM algorithm with respect to the overlap in the mixture. *Neurocomputing, 68*, 105–129.

Ma, J., Xu, L., & Jordan, M. I. (2000). Asymptotic convergence rate of the EM algorithm for gaussian mixtures. *Neural Computation, 12*, 2881–2907.

Meng, X. L. (1994). On the rate of convergence of the ECM algorithm. *Annals of Statistics, 22*, 326–339.

Ng, S-K., & McLachlan, G. J. (2004). Using the EM algorithm to train neural networks: Misconceptions and a new algorithm for multiclass classification. *IEEE Transactions on Neural Networks, 15*, 738–749.

Xu, L. (1997). Comparative analysis on convergence rates of the EM algorithms and its two modifications for gaussian mixtures. *Neural Processing Letters, 6*, 69–76.

Xu, L., & Jordan, M. I. (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural Computation, 8*, 129–151.

Xu, L., Jordan, M. I., & Hinton, G. E. (1994). An alternative model for mixtures of experts. In G. Tasauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems, 7* (pp. 633–640). Cambridge, MA: MIT Press.

Yang, Y., & Ma, J. (2009). A single loop EM algorithm for the mixture of experts architecture. *Lecture Notes in Computer Science, 5552*, 956–968.

Received July 14, 2010; accepted January 15, 2011.