Summary of Model-based EM Algorithm and its Variants

1 Introduction

1.1 Background

In practice, missing values, corruptions, and latent variables are common data problems. The missing part may cause the maximum likelihood estimate (MLE) computation very complex. The expectation-maximization (EM) algorithm is an approach for performing MLE in the presence of the above data problems. In such an algorithm, instead of performing an exact maximization, it simply returns a parameter value that does not decrease the likelihood in each iteration. Before the EM algorithm is introduced in its general form, abundant works had analyzed many EM-type algorithms. In the subsequent studies, monotonically properties and convergence results are established. Although the EM algorithm is widely used for decades, we still don't know how a suitable initialization converges to a useful estimate and mystery of mathematical and theoretical guarantees.

1.2 Related work

Tons of studies work on the EM algorithm and explore its application. Because of the importance of initialization in EM algorithm, some methods are compared, which are classification EM algorithm, stochastic EM algorithm and shorts runs of EM. It finds that the short runs of EM strategy is preferred because of simpleness, good performance in a lot of situations and less sensitivity to noisy data [1]. The property of EM algorithm estimate is closely investigated, and the results show that EM algorithm estimate does provide good asymptotic properties except for some situations in which the population means are quite close to each other and larger differences in the variances of the distributions occur [2]. For accelerating the EM algorithm, Yaming Yu (2012) shows that a trivial modification in the M-step results in an algorithm that maintains a monotonic increase in the log-likelihood, but can have an appreciably faster convergence rate [3]. In some cases, EM algorithm may still have a slow convergence issue, and the Expectation Conjugate Gradient method as an extension of EM algorithm is introduced for a faster approach [4]. Another method to solve the slow convergence issue is Quasi-Newton EM, which has a good performance in non-Gaussian noise simulation[5]. For the stochastic EM algorithm, Wolfgang Jank (2006) finds that an unlucky step size in SEM can lead to very poor algorithmic performance, and he proposes a new method for step size selection and uses EM likelihood-ascent property to monitor the progress [6]. The EM algorithm can be applied to a wide class of semiparametric mixture models, and it behaves well when applicate it to real data, showing that it is fast and easy to be implemented [7]. EM algorithm has many applications in the real world. Weng, Xiao, and Xie (2011) showed that the Stochastic approximation EM algorithm can be applied to Gaussian mixtures in a wireless sensor network and can be energy-efficient [8]. In transmitter localization, EM algorithm can also show some ability with high probability to determine direct position by replacing high dimensions with one dimension, and it outperforms the common gradient-based methods [9].

1.3 The general topic of this research

In this paper, the researchers prove rigorous guarantees on the convergence performance of the EM algorithm and gradient EM algorithm. They first analyze the population level and they apply the results to the sample level. The researchers simulate the convergence rate with three models: Gaussian mixture models, the mixture of regressions, and regression with missing covariates. These examples showed the necessity of qualified initialization and the rate of convergence of the EM and gradient EM algorithms, which confirms that various aspects of the theoretical predictions. Overall, the goal of this paper is to develop some general tools for characterizing the suitably initialized sample-based EM algorithm, and their relation to MLE.

2 Methodology

In this section, we introduce the standard EM algorithm and the gradient EM algorithm, and then both are applied to three models to illustrate the efficiency and feasibility of the algorithms.

2.1 EM algorithm and Gradient EM algorithm

For a joint density function f_{θ^*} with latent variables z and observed variables x and y, we need to obtain the MLE of θ^* by maximum the log-likelihood. In other words, $\theta \mapsto \log g_\theta(y)$, where $g_\theta(y) = \int_{\mathcal{Z}} f_\theta(y,z)\mathrm{d}z$ is the density function of the observed variable Y, is maximized.
For each $\theta \in \Omega$, define $k_\theta(z|y)$ as the conditional density of z given y. By Jensen's inequality, we have the lower bound of log-likelihood at $\theta^{'} \in \Omega$, where the equality holds when $\theta = \theta^{'}$:

$$\log g_{\theta'}(y) \geq \underbrace{\int_{\mathcal{Z}} k_\theta(z|y)\log f_{\theta'}(y,z)\mathrm{d}z}_{Q(\theta^{'}|\theta)} - \int_{\mathcal{Z}} k_\theta(z|y)\log k_\theta(z|y)\mathrm{d}z$$

Standard EM algorithm: Update $\theta^t \to \theta^{t+1}$ by two steps

- E step: Compute the bound $Q(\cdot|\theta^t)$ on the observed data.
- M step: Maximize the bound to update new parameter value $\theta^{t+1} = \mathrm{argmax}_{\theta' \in \Omega} Q(\theta^{'}|\theta^t)$.

We denote the mapping $M : \Omega \to \Omega$ by $M(\theta) := \mathrm{argmax}_{\theta'] \in \Omega} Q(\theta^{'}|\theta)$. In this case, $\theta^{t+1} = M(\theta^t)$.
Gradient EM algorithm: In this algorithm, the requirements of M-step are relaxed. We find a $\theta^{t+1} \in \Omega$ such that $Q(\theta^{t+1}|\theta^t) \geq Q(\theta^t|\theta^t)$ instead of a optimum. Therefore, given a learning rate α, it has

$$\theta^{t+1} = \theta^t + \alpha \nabla Q(\theta^t|Q(\theta^t))$$

We denote the mapping $G : \Omega \to \Omega$ by $G(\theta) = \theta + \alpha\nabla Q(\theta|\theta)$. In this case, $\theta^{t+1} = G(\theta^t)$.
For sample-based versions of EM algorithm, we have: $Q_n(\theta^{'}|\theta) = \frac{1}{n}\sum_{i=1}^n \left(\int_{\mathcal{Z}} k_\theta(z|y)\log f_{\theta'}(y_i,z)\mathrm{d}z\right)$
Therefore, $M_n(\theta) := \mathrm{argmax}_{\theta' \in \Omega} Q_n(\theta^{'}|\theta)$, $G_n(\theta) = \theta + \alpha\nabla Q_n(\theta|\theta)$.

2.2 Illustrative examples

In this section, we analyze three models: Gaussian mixture model, regression mixture model and linear regression model with missing covariates, both for EM algorithm and Gradient EM algorithm of sample-based version.

2.2.1 Gaussian Mixture models

By denoting: $w_\theta(y) = e^{-\frac{\|\theta-y\|_2^2}{2\sigma^2}}\left[e^{-\frac{\|\theta-y\|_2^2}{2\sigma^2}} + e^{-\frac{\|\theta+y\|_2^2}{2\sigma^2}}\right]^{-1}$ **and** $w_\theta(x,y) = \frac{exp\left(\frac{-(y-<x,\theta>)^2}{2\ \sigma^2}\right)}{exp\left(\frac{-(y-<x,\theta>)^2}{2\sigma^2}\right)+exp\left(\frac{-(y+<x,\theta>)^2}{2\sigma^2}\right)}$

The probability density function of this model is $f_\theta(y) = \frac{1}{2}\phi(y;\theta^*,\sigma^2 I_d) + \frac{1}{2}\phi(y;-\theta^*,\sigma^2 I_d)$, where $\phi(\cdot,\mu,\Sigma)$ denotes the density a $\mathcal{N}(\mu,\Sigma)$ and σ^2 is known. The hidden variable $Z \in \{0,1\}$ indicates that $(Y|Z=0) \sim \mathcal{N}(-\theta^*,\sigma^2 I_d)$, $(Y|Z=1) \sim \mathcal{N}(\theta^*,\sigma^2 I_d)$. The EM updates are shown in Table 1:

Target	$Q_n(\theta'\|\theta)$	$-\frac{1}{2n}\sum_{i=1}^n\left[w_\theta(y_i)\|y_i-\theta'\|_2^2 + (1-w_\theta(y_i))\|y_i+\theta'\|_2^2\right]$
Standard EM	$M_n(\theta)$	$\frac{2}{n}\sum_{i=1}^n w_\theta(y_i)y_i - \frac{1}{n}\sum_{i=1}^n y_i$
Gradient EM	$G_n(\theta)$	$\theta + \alpha\left\{\frac{1}{n}\sum_{i=1}^n(2w_\theta(y_i)-1)y_i - \theta\right\}$

Table 1: EM Updates for Gaussian Mixture Models

2.2.2 Regression Mixture models

For (y_i,x_i) drawn from linear regression model $y_i = \langle x_i,\theta^*\rangle + v_i$ with probability $\frac{1}{2}$, and otherwise from $y_i = \langle x_i,-\theta^*\rangle + v_i$, the hidden variables $\{z_i\}_{i=1}^n$ indicates that data is generated from the former model when $z_i = 1$ and otherwise when $z_i = 0$. The EM updates are shown in Table 2:

Target	$Q_n(\theta'\|\theta)$	$-\frac{1}{2n}\sum_{i=1}^n\left(w_\theta(x_i,y_i)(y_i- < x_i-\theta' >)^2 + (1-w_\theta(x_i,y_i))(y_i+ < x_i-\theta' >)^2\right)$
Standard EM	$M_n(\theta)$	$\left(\sum_{i=1}^n x_i x_i^\top\right)^{-1}\left[\sum_{i=1}^n(2w_\theta(x_i,y_i)-1)y_i x_i\right]$
Gradient EM	$G_n(\theta)$	$\theta + \frac{1}{\alpha}\left\{(2w_\theta(x_i,y_i)-1)y_i x_i - x_i x_i^\top\theta\right\}$

Table 2: EM Updates for Regression Mixture Models

2.2.3 Linear Regression Model with Missing Covariates

For a given (x,y), define x_{obs} as the observed portion of x, and θ_{obs} as the corresponding sub-vector of θ. Considering the missing rate ρ, the EM updates are shown in Table 3:

Target	$Q_n(\theta'\|\theta)$	$\frac{1}{2n}\sum_{i=1}^n < \theta', \Sigma_\theta(x_{obs,i},y_i)\theta' > + \frac{1}{n}\sum_{i=1}^n y_i < \mu_\theta(x_{obs,i},y_i),\theta' >$
Standard EM	$M_n(\theta)$	$\left[\sum_{i=1}^n \Sigma_\theta(x_{obs,i},y_i)\right]^{-1}\left[\sum_{i=1}^n y_i\mu_\theta(x_{obs,i},y_i)\right]$
Gradient EM	$G_n(\theta)$	$\theta + \alpha\left\{\frac{1}{n}\sum_{i=1}^n[y_i\mu_\theta(x_{obs,i},y_i) - \Sigma_\theta(x_{obs,i},y_i)\theta]\right\}$

Table 3: EM Updates for Linear Regression Model with Missing Covariates

where $\mu_\theta(x_{obs},y) := \begin{bmatrix}U_\theta z_{obs}\\ x_{obs}\end{bmatrix}$, $\Sigma_\theta(x_{obs},y) := \begin{bmatrix}I & U_\theta z_{obs}x_{obs}^\top\\ x_{obs}z_{obs}^\top U_\theta^\top & x_{obs}x_{obs}^\top\end{bmatrix}$, $U_\theta = \frac{1}{\|\theta_{mis}\|_2^2+\sigma^2}\begin{bmatrix}-\theta_{mis}\theta_{obs}^\top & \theta_{mis}\end{bmatrix}$, and $z_{obs} := \begin{bmatrix}x_{obs}\\ y\end{bmatrix}$.

3 Theoretical Results

In this part, we provide the statistical guarantees of model-based EM algorithm from both population based and sample based analysis. By denoting θ^* that maximizes the likelihood function, we derive the sufficient conditions where the sample-based EM algorithm for three models can converge

to an ϵ-ball around θ^*. Some basic concepts of convergence condition are shown in Table 4:

Concepts	Definition
Smallest scalar: $\epsilon_G^{unif}(n,\delta)$	$sup_{\theta\in\mathbb{B}_2(r;\theta^*)}\|\|G_n(\theta)-M(\theta)\|\|_2 \le \epsilon_G^{unif}(n,\delta)$
Smallest scalar: $\epsilon_M^{unif}(n,\delta)$	$sup_{\theta\in\mathbb{B}_2(r;\theta^*)}\|\|M_n(\theta)-M(\theta)\|\|_2 \le \epsilon_M^{unif}(n,\delta)$
First-order Stability (FOS)	$\|\|\nabla Q(M(\theta)\|\theta^*)-\nabla Q(M(\theta)\|\theta)\|\|_2 \le \gamma\|\|\theta-\theta^*\|\|_2$
Gradient Stability (GS)	$\|\|\nabla Q(\theta\|\theta^*)-\nabla Q(\theta\|\theta)\|\|_2 \le \gamma\|\|\theta-\theta^*\|\|_2$
λ-strongly concave	$Q(\theta_1\|\theta^*)-Q(\theta_2\|\theta^*)-\langle\nabla Q(\theta_2\|\theta^*),\theta_1-\theta_2\rangle \le -\frac{\lambda}{2}\|\|\theta_1-\theta_2\|\|_2^2$
μ-smooth	$Q(\theta_1\|\theta^*)-Q(\theta_2\|\theta^*)-\langle\nabla Q(\theta_2\|\theta^*),\theta_1-\theta_2\rangle \ge -\frac{\mu}{2}\|\|\theta_1-\theta_2\|\|_2^2$

Table 4: Basic Concepts of Convergence Condition

For general EM algorithm, we summarize the conditions and results of convergence in Table 5. The details of the theorem are shown in Appendix A:

Conditions	Results	Theorem
FOS(λ) and λ-strongly concave	Population-level Standard EM	Appendix A.1
Bound on ϵ_M^{unif} and population concavity	Sample-level Standard EM	Appendix A.1
GS(γ) and λ-strongly concave, μ-smooth	Population-level Gradient EM	Appendix A.2
Bound on ϵ_G^{unif} and population concavity	Sample-level Gradient EM	Appendix A.2

Table 5: Summary of General EM Convergence

We denote $\frac{\|\|\theta^*\|\|_2}{\sigma}$ as the **signal-to-noise ratio (SNR)**, which measures the difficulty of estimating. In this case, we assume $\frac{\|\|\theta^*\|\|_2}{\sigma} > \eta > 0$ for a sufficient large constant. This condition is necessary since it constrains the ML solution with lower variance and quick convergence. Under the assumption of SNR, the following part provides conditions of convergence for the specific models.

3.1 Gaussian Mixture Models

In addition the assumptions in the general EM convergence and the assumption of SNR, we add the conditions for each method in Table 6 to guarantee the convergence of Gaussian mixture models:

Method	Conditions	Result
Standard EM	C1: $r=\frac{\|\|\theta^*\|\|_2}{4}$ and $\kappa(\eta)\le e^{-c\eta^2}$ for $c>0$	$\|\|M(\theta)-\theta^*\|\|_2 \le \frac{\gamma}{\lambda}\|\|\theta-\theta^*\|\|_2$
Gradient EM	C1: $r=\frac{\|\|\theta^*\|\|_2}{4}$ and $\kappa(\eta)\le e^{-c\eta^2}$ for $c>0$	$\|\|G(\theta)-\theta^*\|\|_2 \le \left(1-\frac{2\lambda-2\gamma}{\mu+\lambda}\right)\|\|\theta-\theta^*\|\|_2$
Sample-based	$n > c_1 d\log(1/\delta)$ and $\theta^0\in\mathbb{B}_2(r;\theta^*)$ with C1	Bound with positive (c,c_1,c_2)

Table 6: Guarantee for Gaussian Mixture Models

For the sample-based EM update, the iterates have the bound with probability at least $1-\delta$:

$$\|\|\theta^t-\theta^*\|\|_2 \le \kappa^t\|\|\theta^0-\theta^*\|\|_2 + \frac{c_2}{1-\kappa}\|\|\theta^*\|\|_2\sqrt{\frac{d}{n}(\sigma^2+\|\|\theta^*\|\|_2^2)\log(\frac{1}{\delta})}$$

The standard EM bound provides a rough guideline for iteration time since the bound consists of unknown θ^* and contraction coefficient κ. It suggests that the iteration complexity should grow logarithmically in the ratio n/d. In addition, it also suggests the statistical error $\|\|\theta^t-\theta^*\|\|_2$ decrease geometrically and then level off at a plateau. On the other hand, the optimization error decreases geometrically to numerical tolerance.

3.2 Regression Mixture Models

Except for the assumptions of general EM algorithm and SNR, we add the conditions for each method in Table 7 to guarantee the convergence of regression mixture models:

Method	Conditions	Result
Standard EM	C2: $r = \frac{\|\|\theta^*\|\|_2}{32}$ and $\kappa(\eta) \leq \frac{1}{2}$	$\|\|M(\theta) - \theta^*\|\|_2 \leq \frac{\gamma}{\lambda}\|\|\theta - \theta^*\|\|_2$
Gradient EM	C2: $r = \frac{\|\|\theta^*\|\|_2}{32}$ and $\kappa(\eta) \leq \frac{1}{2}$	$\|\|G(\theta) - \theta^*\|\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)\|\|\theta - \theta^*\|\|_2$
Sample-based	$n > c_1 d \log(1/\delta)$ and $\theta^0 \in \mathbb{B}_2(r; \theta^*)$ with C2	Bound with positive (c, c_1, c_2)

Table 7: Guarantee for Regression Mixture Models

For the sample-based EM update, the iterates have the bound with probability at least $1 - \delta$:

$$||\theta^t - \theta^*||_2 \leq \kappa^t ||\theta^0 - \theta^*||_2 + c_2 ||\theta^*||_2 \sqrt{\frac{d}{n}(\sigma^2 + ||\theta^*||_2^2) \log(\frac{1}{\delta})}$$

In this case, the bound also provides guidance on the number of iterations to perform. It matches the minimax rate for estimation of a d-dimensional regression vector. Besides that, the convergence features are similar to the EM algorithm in Gaussian mixture models.

3.3 Linear Regression Model with Missing Covariates

Except for the assumptions of general EM algorithm, we assume the missing rate $\rho < \frac{1}{1+2\xi(1+\xi)}$, where $\xi = (\xi_1 + \xi_2)^2$ and the SNR satisfies $\frac{||\theta^*||_2}{\sigma} \leq \xi_1$ and $||\theta - \theta^*||_2 \leq r = \xi_2 \sigma$.

Method	Conditions	Result
Standard EM	C3: $r = \xi_2\sigma$ and $\kappa(\eta) = \frac{\xi + \rho(1+2\xi(1+\xi))}{1+\xi} < 1$	$\|\|M(\theta) - \theta^*\|\|_2 \leq \frac{\gamma}{\lambda}\|\|\theta - \theta^*\|\|_2$
Gradient EM	C3: $r = \xi_2\sigma$ and $\kappa(\eta) = \frac{\xi + \rho(1+2\xi(1+\xi))}{1+\xi} < 1$	$\|\|G(\theta) - \theta^*\|\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)\|\|\theta - \theta^*\|\|_2$
Sample-based	$n > c_1 d \log(1/\delta)$ and $\theta^0 \in \mathbb{B}_2(r; \theta^*)$ with C3	Bound with positive (c, c_1, c_2)

Table 8: Guarantee for Linear Regression Model with Missing Covariates

In this case, we require upper bound for signal-to-noise ratio. Since the norm $||\theta^*||_2$ increases, the amount of missing information increases in proportion to the amount of observed information. Intuitively, this condition is unavoidable in practice. For the sample-based EM update, the iterates have the bound with probability at least $1 - \delta$:

$$||\theta^t - \theta^*||_2 \leq \kappa^t ||\theta^0 - \theta^*||_2 + \frac{c_2}{1-\kappa}\sqrt{\frac{d}{n}(\sigma^2 + 1) \log(\frac{1}{\delta})}$$

Similar to previous models, the optimization error decays geometrically while the statistical error decays geometrically before leveling off. In conclusion, the conditions are reasonable.

4 Experimental Details

4.1 Setup

For each model, we simulate 10 different problem examples of dimension $d = 10$, sample size $n = 1000$, signal-to-noise ratio $\frac{||\theta^*||_2}{\sigma} = 2$ and missing probability $\rho = 0.2$ (only used for linear regression model with missing covariates), and then log optimization error $\log(||\theta^t - \theta^n||_2)$ and log statistical error $\log(||\theta^t - \theta^*||_2)$ are recorded across iteration count.

4.2 Results

By setting learning rate $\alpha = 0.5$ and initializing θ^0 as the way mentioned in Theoretical Results, the plots of the iteration count versus log optimization error and log statistical error for each model simulations are shown in this part.

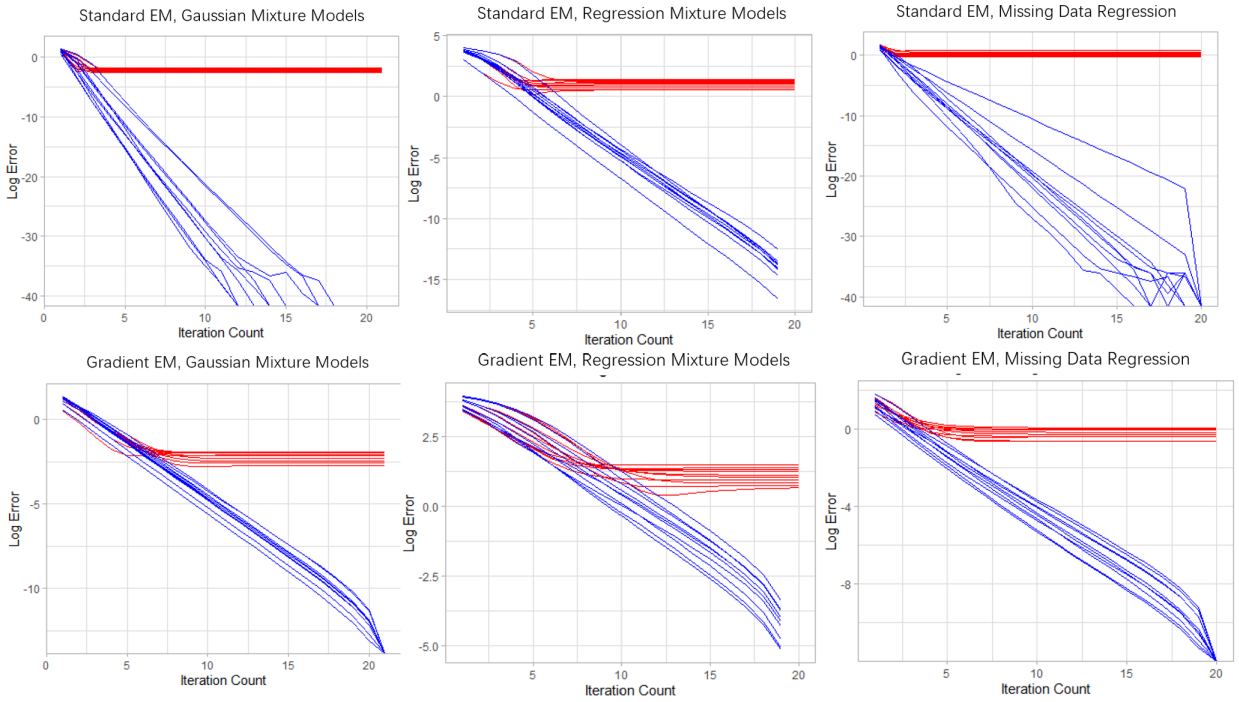

Figure 1: Log Optimization Error(Blue) and Log Statistical Error(Red)

For the simulation results of each model in Figure 1, we can see that the optimization error decreases geometrically while the statistical error decays geometrically before leveling off in both plots, which is consistent as we discussed in Theoretical Results.

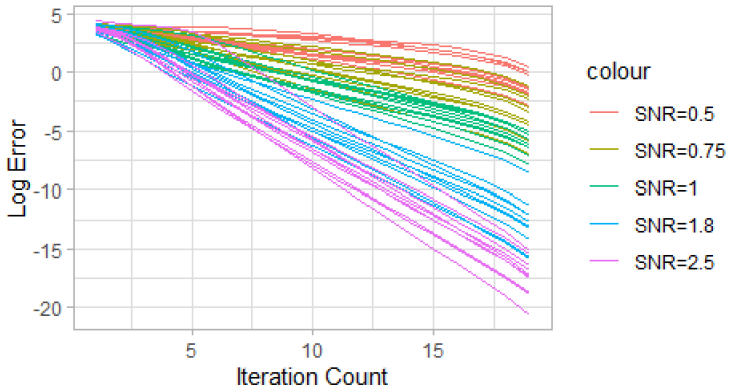

Figure 2: EM, Mixture of Gaussians

For each SNR, we performed 10 independent trials of a Gaussian mixture model with dimension d = 10 and sample size n = 1000. The Figure 2 shows that larger values of SNR lead to faster convergence rates, which is consistent with the discussion of SNR in Theoretical Results.

A Additional Proof Details

A.1 General Convergence of Standard EM Updates

Population-level Convergence: For some radius $r > 0$ and γ and λ that $0 \leq \gamma < \lambda$, suppose that the function $Q(\cdot|\theta^*)$ is λ-strongly concave and satisfy condition FOS(γ) over the ball $\mathbb{B}_2(r;\theta^*)$, then the population EM operator M is contractive for all $\theta \in \mathbb{B}_2(r;\theta^*)$ that:

$$||M(\theta) - \theta^*||_2 \leq \frac{\gamma}{\lambda}||\theta - \theta^*||_2$$

In this case, the population EM sequence $\{\theta^t\}_{t=0}^{\infty}$ is linearly convergent if $\theta^0 \in \mathbb{B}_2(r;\theta^*)$.
Sample-based Convergence: Suppose that the population EM operator $M : \Omega \to \Omega$ is contractive with parameter $\kappa \in (0,1)$ on the ball $\mathbb{B}_2(r;\theta^*)$ with the initial vector $\theta^0 \in \mathbb{B}_2(r;\theta^*)$:

1. For large sample size n which satisfies $\epsilon_M^{unif}(n,\delta) \leq (1-\kappa)r$, we have the bound:

$$||\theta^t - \theta^*||_2 \leq \kappa^t||\theta^0 - \theta^*||_2 + \frac{1}{1-\kappa}\epsilon_M^{unif}(n,\delta)$$

2. Considering sample splitting method with iteration T, we have:

$$||\theta^t - \theta^*||_2 \leq \kappa^t||\theta^0 - \theta^*||_2 + \frac{1}{1-\kappa}\epsilon_M(\frac{n}{T},\frac{\delta}{T})$$

In this case, the theorem shows the geometric convergence to the target parameter θ^*. In addition, the bound also suggests a reasonable choice of the number of iteration T with fixed sample size.

A.2 General Convergence of Gradient EM Updates

It is similar to gradient ascent algorithm that intuitively if the function $Q(\cdot|\theta)$ is close enough to $Q(\cdot|\theta^*)$ the EM operator might be converge to the target.
Population-level Convergence: For some radius $r > 0$ and $0 \leq \gamma < \lambda \leq \mu$, suppose that the function $Q(\cdot|\theta^*)$ is λ-strongly concave, μ-smooth and satisfy condition GS(γ) over the ball $\mathbb{B}_2(r;\theta^*)$, then the population EM operator G is contractive for all $\theta \in \mathbb{B}_2(r;\theta^*)$ that:

$$||G(\theta) - \theta^*||_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)||\theta - \theta^*||_2$$

In this case, the population gradient EM sequence $\{\theta^t\}_{t=0}^{\infty}$ is convergent if $\theta^0 \in \mathbb{B}_2(r;\theta^*)$.
Sample-based Convergence:Suppose that the population gradient EM operator $G : \Omega \to \Omega$ is contractive with parameter $\kappa \in (0,1)$ on the ball $\mathbb{B}_2(r;\theta^*)$ with the initial vector $\theta^0 \in \mathbb{B}_2(r;\theta^*)$:

1. For large sample size n which satisfies $\epsilon_G^{unif}(n,\delta) \leq (1-\kappa)r$, we have the bound:

$$||\theta^t - \theta^*||_2 \leq \kappa^t||\theta^0 - \theta^*||_2 + \frac{1}{1-\kappa}\epsilon_G^{unif}(n,\delta)$$

2. Considering sample splitting method with iteration T, we have:

$$||\theta^t - \theta^*||_2 \leq \kappa^t||\theta^0 - \theta^*||_2 + \frac{1}{1-\kappa}\epsilon_G(\frac{n}{T},\frac{\delta}{T})$$

In this case, it is the same as the contractive conditions in standard EM updates.
Similar to the stochastic gradient ascent algorithm, with the same assumption of sample-based convergence in gradient EM update, the **stochastic gradient EM updates** with step size $\alpha^t = \frac{3}{2\xi(t+2)}$ satisfies:

$$E[||\theta^t - \theta^*||_2^2] \leq \frac{9\sigma_G^2}{\xi^2}\frac{1}{t+2} + \left(\frac{2}{t+2}\right)^{\frac{3}{2}} ||\theta^0 - \theta^*||_2^2$$

In this case, we need to assume that $\theta^0 \in \mathbb{B}_2(\frac{r}{2}; \theta^*)$ and pick a suitable step size to achieve convergence.

B Reference

[1].Christophe Biernacki, Gilles Celeux, Gérard Govaert, Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Computational Statistics & Data Analysis, Volume 41, Issues 3–4, 2003, Pages 561-575, ISSN 0167-9473.
[2].Dechavudh Nityasuddhi, Dankmar Böhning, Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances, Computational Statistics & Data Analysis, Volume 41, Issues 3–4, 2003, Pages 591-601, ISSN 0167-9473.
[3].Yaming Yu (2012) Monotonically Overrelaxed EM Algorithms, Journal of Computational and Graphical Statistics,21:2, 518-537, DOI: 10.1080/10618600.2012.672115
[4]. Rishik Ranjan, Biao Huang, Alireza Fatehi, Robust Gaussian process modeling using EM algorithm, Journal of Process Control, Volume 42, 2016, Pages 125-136, ISSN 0959-1524.
[5].X. Jinlong, L. Jianwu and Y. Yang, "A New EM Acceleration Algorithm for Multi-user Detection," 2011 Third International Conference on Measuring Technology and Mechatronics Automation, Shangshai, 2011, pp. 150-153, doi: 10.1109/ICMTMA.2011.43.
[6].Wolfgang Jank (2006) Implementing and Diagnosing the Stochastic Approximation EM Algorithm, Journal of Computational and Graphical Statistics, 15:4, 803-829,
DOI: 10.1198/106186006X157469
[7].Laurent Bordes, Didier Chauveau, Pierre Vandekerkhove, A stochastic EM algorithm for a semiparametric mixture model, Computational Statistics & Data Analysis, Volume 51, Issue 11, 2007, Pages 5429-5443, ISSN 0167-9473.
[8]. Weng, Y.; Xiao, W.; Xie, L. Diffusion-Based EM Algorithm for Distributed Estimation of Gaussian Mixtures in Wireless Sensor Networks. Sensors 2011, 11, 6297-6316.
[9]. Elad Tzoreff, Anthony J. Weiss, Expectation-maximization algorithm for direct position determination, Signal Processing, Volume 133, 2017, Pages 32-39, ISSN 0165-1684.