A Review of Statistical Guarantees for the EM Algorithm with Simulation Examples

Zhi Zhang*

1 Introduction

It is very challenging to compute the maximum likelihood estimate (MLE) in most of the incomplete data problems setting. The expectation-maximization (EM) algorithm is a very useful tool to solve such problems. But there is a gap between the practical use of EM and its theoretical properties. In most models, the EM algorithm is only guaranteed to return a local optimum of the sample likelihood function without good statistical properties of MLE. The main contribution of the paper[1] is to provide a guarantee of convergence to a local optimum that matches the performance of MLE under specific conditions. The authors achieve that by two steps: first derive the statistical guarantees of the EM and first-order EM algorithms at the population level, then extend the result to the case of finite samples. In addition, the authors present three canonical problems which satisfy the conditions of the theorems and provide a tight characterizations of the region of convergence for them: symmetric mixture of two Gaussians, symmetric mixture of two regressions and linear regression with missing completely at random.

To be specific, the authors make advances over the classical results in the following three specific directions:

1. By imposing conditions of gradient smoothness, strong concavity and smoothness on the auxiliary Q-function underlying the EM algorithm, they give a quantitative characterization of the region of convergence to the population global optimum. The existing research can only guarantee convergence to a fixed point when initialized around its some neighborhood but can't quantify the size of the neighborhood.

2. The authors are the first one who analyze the EM algorithm on the population level and then quantify the minimal requirement of sample size to avoid spurious fixed points far away from the population MLE. The classical results on the EM algorithm are all based on the finite sample size and can not make sure that any fixed points of the sample likelihood are close to the population MLE.

3. A precise characterization of initialization is presented in the paper. A two-stage estimator, which involves the initialization of the moments method and the refinement of the EM algorithm, performs well empirically. The authors' theoretical results help explain this behavior and furthermore provide guideline on the refinement stage.

Apart from the contributions mentioned above, the authors' effort in the treatment of three examples is also remarkable. They show that their conditions hold in a large region around the MLE, and that the size of this region is determined by interpretable problem-dependent quantities. Extensive simulations are carried out to confirm the theorems.

*Department of Statistics, UC Davis, wwzzhan@ucdavis.edu

1.1 Related work

The EM algorithm has a long history (e.g., [2], [3], [5], [6], [11], [12], [13]) and its modern general form was introduced by Dempster, Laird and Rubin [4], who also established its well-known monotonicity properties. Wu [14] provided guarantee for the EM algorithm to converge to the unique global optimum when the likelihood is unimodal and certain regularity conditions hold. When the likelihood function is multi-modal, which is more common in reality, existing works can only guarantee convergence to some local optimum of the likelihood at an asymptotically geometric rate (see, e.g., [7], [8], [9], [10]). This type guarantee is not enough to promise convergence to a "good" local optimum. The local optimum can be far away from any global optimum of the likelihood. The paper closes the gap by guaranteeing geometric convergence to a "good" EM fixed point.

2 Methodology

The main methodological contributions of the paper is the construction of the theorems on the population EM and first-order EM algorithms, which are then used on proof the convergence theorems of their finite-sample counterparts. This innovation is very intuitive. In order to show that the algorithm can converge to a global optimal point, we must know the existence of the optimal point and have an insight on the perfect situation, e.g., when we have infinite samples. This intuition makes it natural to introduce the population EM and first-order EM algorithms. If the population EM and first-order EM algorithms perform well, people will want to know the performance of their finite-sample counterparts, which is of more interest in reality. Another intuition is to bound the difference between the sample and population function or gradient, which is called an empirical process. When the empirical process can be bounded well, which will be achieved by sufficiently large sample size, it can be expected that the behavior of the finite-sample EM and first-order EM algorithm will be very similar to their population counterparts. By this way, the authors arrived their ultimate interest by building theorems on the finite-sample EM and first-order EM algorithm.

To be more specific, at first, the author gives conditions under which the population algorithms are contractive to the global optimal point, e.g., the true parameter θ^*, when initialized in a ball around the θ^*. These conditions allow us to establish the region of attraction of θ^*.

For the first-order EM algorithms, they exploit the oracle auxiliary function $q(\theta)$, whose convergence at a geometric rate to the global optimum is guaranteed under some standard regularity conditions by classical theory on gradient methods. Then they impose the condition of gradient smoothness in the neighborhood of θ^* to make sure that the the gradient of the auxiliary function $Q(\theta|\theta)$ with respect to its first argument is close enough to $\nabla q(\theta)$. So the convergence result on $q(\theta)$ can be extended to $Q(\theta|\theta)$. Furthermore, provided that the sample size is large enough, the sample gradient $\nabla Q_n(\theta|\theta)$ will be within a small area of $\nabla Q(\theta|\theta)$ with high probability. This fact ensures the convergence properties of the empirical EM algorithm.

For the EM algorithm, the authors first define the operator $M(\theta) = \arg\max_{\theta' \in \Omega} Q(\theta|\theta')$. Then, conditioned on the first-order stability (FOS) of $Q(\cdot|\theta)$, the KKT conditions, and a similar globally strong concavity of the function $q(\theta)$, the population EM operator M is contractive over a Euclidean ball of θ^*. Then by controlling the quantity of the empirical process $M(\theta) - M_n(\theta)$ in the neighborhood of θ^* with high probability, which can be achieved by sufficiently large sample size, the authors prove the convergence of the EM iterates to the global optimal point, e.g., the true parameters.

The authors make a remark that the first order EM algorithm is identical to the gradient ascent on the marginal log-likelihood function. The intuition behind the analysis is the exploiting of the Q-function, which makes it possible to prove guarantees for a specific class of models whose log

likelihood is not concave but the Q-function is concave and satisfies some conditions.

What is of interest is the convergence property of sample-based EM algorithm. However, it is very difficult to link the estimate $\hat{\theta}$ returned by the algorithm to θ^* directly. The authors' solution is very intuitive and clear - step by step, which goes from $q(\theta)$ to $Q(\theta|\theta)$, finally to $Q_n(\theta|\theta)$. The proof of these theorems is not difficult, while the proof of the corollaries involve several techniques such as symmetrization, contraction and concentration inequalities, which are common in the field of high dimensional statistics. This is because when apply the theorems to the canonical problems, e.g., to get the corollaries, a large part of the technical effort is devoted to establishing bounds on the empirical process $\{\nabla Q(\theta|\theta) - \nabla Q_n(\theta|\theta), \theta \in \mathbb{B}_2(r;\theta^*)\}$.

Other methodological progresses are attributed to the corollaries of the three canonical problems. For instance, for the Gaussian mixture models, the authors use a simple setting with two components, balanced weights and isotropic covariances. Then the goal is to estimate the unknown mean vector θ^*. The difficulty of estimating such a mixture model can be characterized by the signal-to-noise ratio $\frac{||\theta^*||_2}{\sigma}$ (SNR). Finally, the corollary of population result for the first-order EM algorithm for Gaussian mixtures is built after verifying that the Gaussian mixture model satisfies the gradient smoothness, λ-strong concavity and μ-smoothness. The condition on SNR is intuitive and and has empirical support.

3 Theoretical Results

Let Y and Z be random variables in the sample spaces $\mathcal{Y}$ and $\mathcal{Z}$ with joint density function f_{θ^*} that belongs to some parameterized family $\{f_\theta|\theta \in \Omega\}$, where Ω is some nonempty convex set of parameters. Here Y is the observed component and Z is the latent structure in the data. For each $\theta \in \Omega$, denote $k_\theta(z|y)$ to be the conditional density of z given y. Suppose that we have n i.i.d. observations $\{y_i\}_{i=1}^n$, the standard EM algorithm maximizes the log-likelihood function $\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log\left[\int_{\mathcal{Z}} f_\theta(y_i, z_i)\, dz_i\right]$ by maximizing the finite-sample Q function $Q_n(\theta|\theta')$ at $\theta' = \theta^t$.

$$\theta^{t+1} = \arg\max_{\theta\in\Omega} Q_n\left(\theta|\theta^t\right) \quad \text{where} \quad Q_n\left(\theta|\theta'\right) = \frac{1}{n}\sum_{i=1}^n \left(\int_{\mathcal{Z}} k_{\theta'}(z|y_i) \log f_\theta(y_i, z)\, dz\right) \tag{1}$$

Definition 3.1. Denote θ^* to be the true parameter and g_{θ^*} to be the marginal density of the observed samples. Then the population counterpart of EM algorithm maximizes the log-likelihood function

$$\ell(\theta) = \int_{\mathcal{Y}} \log\left[\int_{\mathcal{Z}} f_\theta(y, z)dz\right] g_{\theta^*}(y)dy \tag{2}$$

by maximizing the population Q-function given below in each iterations at $\theta' = \theta^t$

$$Q\left(\theta|\theta'\right) = \int_{\mathcal{Y}} \left(\int_{\mathcal{Z}} k_{\theta'}(z|y) \log f_\theta(y, z)dz\right) g_{\theta^*}(y)dy \tag{3}$$

The main theoretical results of the paper can be viewed as three parts: the convergence theorems of the standard EM and the first order EM algorithm (Theorem 1 and 4 in the paper), their counterparts of the sample-based algorithms (Theorem 2 and 5(a)) and the extension of sample-splitting algorithms (Theorem 3 and 5 (b)), and the corollaries which apply these theorems to the specific three canonical problems of interest. In particular, the update rule for three versions for first order EM algorithm is given by Table 1 and there are mainly three common conditions that need to be guaranteed in these theorems.

Condition 3.1 (Gradient smoothness)**.** For an appropriately small parameter $\gamma \geq 0$, we have that

$$\|\nabla q(\theta) - \nabla Q(\theta|\theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2 \quad \text{for all } \theta \in \mathbb{B}_2(r;\theta^*) \tag{4}$$

Algorithm	Update Rules
Population-level first-order EM algorithm	$\theta^{t+1} = \theta^t + \alpha \nabla Q\left(\theta|\theta^t\right)\big\|_{\theta=\theta^t}$
Sample-based first-order EM algorithm	$\theta^{t+1} = \theta^t + \alpha \nabla Q_n\left(\theta|\theta^t\right)\big\|_{\theta=\theta^t}$
Sample-splitting first order EM algorithm	$\theta^{t+1} = \theta^t + \alpha \nabla Q_{\lfloor n/T \rfloor}\left(\theta|\theta^t\right)\big\|_{\theta=\theta^t}$

Table 1: The update rule for three first-order EM algorithm. The gradient $\nabla Q_n\left(\theta|\theta^t\right)$ is taken with respect to the first argument of the Q-function and $\nabla Q_{\lfloor n/T \rfloor}$ denotes the finite sample Q-function computed using a fresh subset of $\lfloor n/T \rfloor$ samples at each iteration.

Condition 3.2 (λ-strong concavity)**.** There is some $\lambda > 0$ such that for all pairs $\theta_1, \theta_2 \in \mathbb{B}_2\left(r; \theta^*\right)$

$$q\left(\theta_1\right) - q\left(\theta_2\right) - \left\langle \nabla q\left(\theta_2\right), \theta_1 - \theta_2 \right\rangle \leq -\frac{\lambda}{2}\left\|\theta_1 - \theta_2\right\|_2^2 \tag{5}$$

When we require this condition to hold for all pairs $\theta_1, \theta_2 \in \Omega$ we refer to this as global λ-strong concavity.

Condition 3.3 (μ-smoothness)**.** There is some $\mu > 0$ such that for all $\theta_1, \theta_2 \in \mathbb{B}_2\left(r; \theta^*\right)$

$$q\left(\theta_1\right) - q\left(\theta_2\right) - \left\langle \nabla q\left(\theta_2\right), \theta_1 - \theta_2 \right\rangle \geq -\frac{\mu}{2}\left\|\theta_1 - \theta_2\right\|_2^2 \tag{6}$$

- For theorem 1-3, there are some common parameters. They are **triplet** (γ, λ, μ) such that $0 \leq \gamma < \lambda \leq \mu$, **radius** $r > 0$, **step size** $\alpha = \frac{2}{\mu+\lambda}$ and **initialization** $\theta^0 \in \mathbb{B}_2\left(r; \theta^*\right)$.
- Theorem 1 is related to the analysis on **population level for first order EM algorithm**. This theorem guarantees a **geometric rate of convergence** toward θ^* for population level EM algorithm.
- Theorem 2 is related to the convergence rate for **sample-based first-order EM algorithm**. The extra parameter $\varepsilon_Q^{\text{unif}}(n, \delta)$ is the smallest scalar such that $\mathbb{P}\left(\sup \mathcal{E} \leq \varepsilon_Q^{\text{unif}}(n, \delta)\right) \geq 1 - \delta$ where the empirical process is defined as $\mathcal{E} = \{\nabla Q(\theta|\theta) - \nabla Q_n(\theta|\theta), \theta \in \mathbb{B}_2\left(r; \theta^*\right)\}$.
- Theorem 3 is related to the convergence rate for **mini-batch sample-based first-order EM algorithm**. The full data is divided into T subsets of size $\lfloor n/T \rfloor$. The extra parameter $\varepsilon_Q(n, \delta)$ be the smallest scalar such that for any fixed $\theta \in \mathbb{B}_2\left(r; \theta^*\right)$,

$$\mathbb{P}\left[\left\| \nabla Q_n\left(\theta|\theta^t\right)\big|_{\theta=\theta^t} - \nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t}\right\|_2 < \varepsilon_Q(n, \delta)\right] \geq 1 - \delta$$

Algorithm	Extra parameter	Convergence rate
Population	No extra parameter	$\left\|\theta^t - \theta^*\right\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu+\lambda}\right)^t \left\|\theta^0 - \theta^*\right\|_2$
Sample-based	$\varepsilon_Q^{\text{unif}}(n, \delta) \leq (\lambda - \gamma) r$	$\left\|\theta^t - \theta^*\right\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu+\lambda}\right)^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{\varepsilon_Q^{\text{unif}}(n, \delta)}{\lambda - \gamma}$
Sample-splitting	$\varepsilon_Q\left(\frac{n}{T}, \frac{\delta}{T}\right) \leq (\lambda - \gamma) r$	$\left\|\theta^t - \theta^*\right\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu+\lambda}\right)^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{\varepsilon_Q(n/T, \delta/T)}{\lambda - \gamma}$

Table 2: Summary Table for population and sample-based EM algorithm. For a specific tolerance parameter $\delta > 0$, the conditions for extra parameter in sample-based and sample splitting holds with probability at least $1 - \delta$.

The theoretical results of theorem 1-3 are summarized in Table 2. The **more detailed proof** than the paper of Theorem 1 is provided in Appendix A.1 and the proof of theorem 2 and theorem

3 are provided in Appendix A.2. Among these three theorems, conditions 3.1 through 3.3 hold over the ball $\mathbb{B}_2(r;\theta^*)$ for a reasonably large choice of r. In practice, there are some popular examples that satisfies the conditions in Theorem 1, Theorem 2 and Theorem 3. The author applies these theorems on three classes of statistical models for which the EM algorithm is frequently applied, namely, Gaussian mixture models, mixtures of regressions and regression with missing covariates. The details for theoretical illustration are provided in Appendix B. For this three specific models, the authors check the conditions we discuss above and the applies three theorems to obtain the corresponding convergence results. In these examples, the strong concavity and smoothness conditions are straightforward, whereas establishing gradient smoothness (Condition 3.1) is more challenging. Establishing that the gradient condition holds over (nearly) optimally-sized regions involves carefully leveraging properties of the generative model as well as smoothness properties of the log-likelihood function.

Besides, the author also provides with some extensions for first order EM algorithm both on population level and sample based methods. For population level, the authors establish the convergence rate for EM algorithm. For sample-based method, the authors provide some general guarantees for sample-based EM algorithm. These theorems are provided in Appendix A.3 and are quite similar to the main results above (Theorem 1-3).

4 Experimental Details

4.1 Setup

We set up the simulation experiments according to the three problems introduced in the paper (Mixture of Gaussians, Mixture of Regressions, and Missing Data Regression). We implemented the EM updates and first-order EM updates for each of problems, in a total, there are 3×2 (problem, algorithm) combinations.

For each problem-algorithm combination, we examined the logarithm statistical error $\|\theta^t - \theta^*\|_2$ and optimization error $\|\theta^t - \hat{\theta}\|_2$ over iterations. We replicated with 10 times, and fixed the dimension $d = 10$, sample size $n = 1000$, sigma $\sigma = 1$, and signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma} = 2$. In addition, while each replicate we randomly did the initialization, but we also made sure that the θ^0 was in the range of radius of convergence stated by the theorems. For missing data problem, we let $p = 0.2$ for every sample's jth dimension to be randomly missing. The code was attached in the Appendices.

4.2 Results

Fig 1 (a,c,e) shows the EM updates for three problems, (b, d, f) shows First-Order EM updates. We observe that the statistical error (red curve represented) decreases geometrically, and then level off at a plateau, while the optimization error (blue curve) keeps decreasing geometrically. The results also show that the first-order EM takes more iteration to converge. Besides, we notice that the Missing data regression has more variance compared to other two problems, this might be due to randomness introduced during the data creation makes the problem naturally harder. This randomness might also cause the statistical error fluctuation as decreasing, then increasing and eventually plateau in the first-order EM. Figure 1 (g) confirms that the convergence rate increases as the increasing of the signal-to-noise ratio.

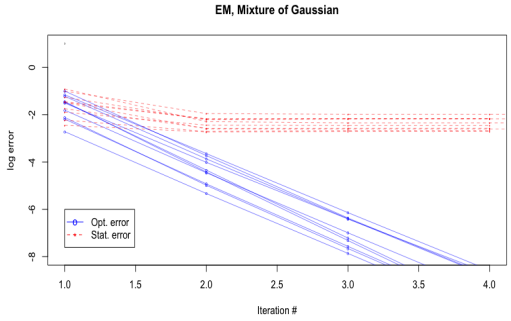
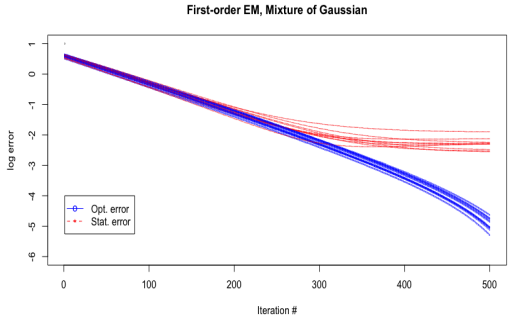
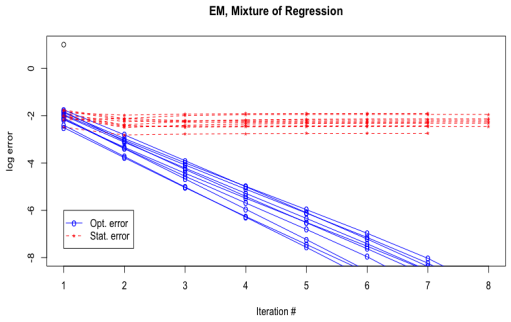
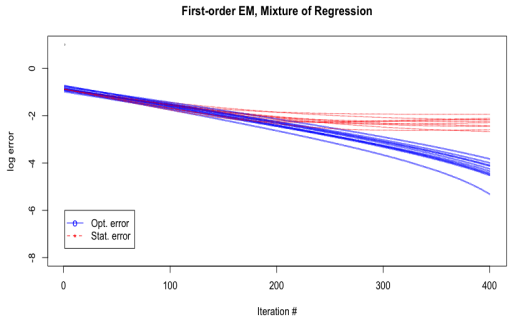
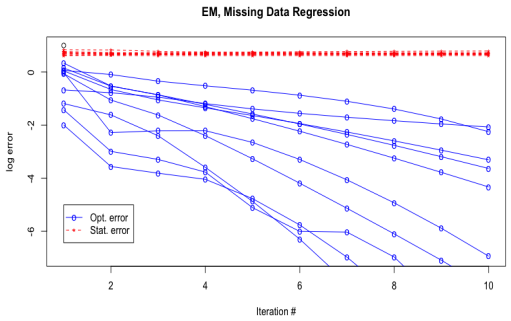
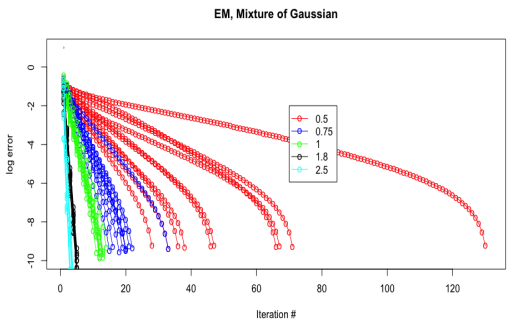
	EM	First Order EM
Mixture of Gaussians	 a	 b
Mixture of Regressions	 c	 d
Missing Data Regression	 e	 f
	Mixture of Gaussian	
EM, Various SNR	 g	

Figure 1: The full sets of replication results for figure 5,6,7,8 in the Balakrishnan's paper

References

[1] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 02 2017.

[2] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[3] Evelyn ML Beale and Roderick JA Little. Missing values in multivariate analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):129–145, 1975.

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

[5] Herman O Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194, 1958.

[6] Michael Healy and Michael Westmacott. Missing values in experiments analysed on automatic computers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 5(3):203–206, 1956.

[7] Alfred O Hero and Jeffrey A Fessler. Convergence in norm for alternating expectation-maximization (em) type algorithms. *Statistica Sinica*, pages 41–54, 1995.

[8] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[9] Xiao-Li Meng et al. On the rate of convergence of the ecm algorithm. *The Annals of Statistics*, 22(1):326–339, 1994.

[10] Xiao-Li Meng and Donald B Rubin. On the global and componentwise rates of convergence of the em algorithm. *Linear Algebra and its Applications*, 199:413–425, 1994.

[11] T Orchard and MA Woodbury. A missing information principle: theory and applications proceedings of the 6th berkeley symposium on mathematical statistics vol, 1972.

[12] Donald B Rubin. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69(346):467–474, 1974.

[13] Rolf Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, pages 49–58, 1974.

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

A Additional Proof Details

A.1 Additional Proof Details for Population-level Analysis

Lemma A.1. For a function q with the λ-strong concavity and μ-smoothness properties (Conditions 3.2 and 3.3), the oracle iterates

$$\tilde{\theta}^{t+1} = \tilde{\theta}^t + \alpha \nabla q\left(\tilde{\theta}^t\right) \quad \text{for } t = 0, 1, 2, \ldots$$

with stepsize $\alpha = \frac{2}{\mu+\lambda}$ are linearly convergent:

$$\left\|\theta^t + \alpha \nabla q(\theta)|_{\theta=\theta^t} - \theta^*\right\|_2 \leq \left(\frac{\mu-\lambda}{\mu+\lambda}\right) \left\|\theta^t - \theta^*\right\|_2 \tag{7}$$

Proof. Note that q is λ-strong concavity and μ-smooth, we see that $f = -q$ is λ-strong convex and μ-smooth. Therefore, we define

$$h(\theta) = f(\theta) - \frac{\lambda}{2}\|\theta\|_2^2$$

and we have

$$\begin{aligned}
0 &\leq (\nabla h(\theta_1) - \nabla h(\theta_2))^\top (\theta_1 - \theta_2) \\
&= (\nabla f(\theta_1) - \nabla f(\theta_2))^\top (\theta_1 - \theta_2) - \lambda\|\theta_1 - \theta_2\|_2^2 \\
&\leq (\mu - \lambda)\|\theta_1 - \theta_2\|_2^2
\end{aligned}$$

Therefore, co-coercivity of $\nabla h(\theta)$ can be written as

$$(\nabla f(\theta_1) - \nabla f(\theta_2))^\top (\theta_1 - \theta_2) \geq \frac{\lambda\mu}{\lambda+\mu}\|\theta_1 - \theta_2\|_2^2 + \frac{1}{\lambda+\mu}\|\nabla f(\theta_1) - \nabla f(\theta_2)\|_2^2$$

or equivalently

$$(\nabla q(\theta_1) - \nabla q(\theta_2))^\top (\theta_1 - \theta_2) \leq -\frac{\lambda\mu}{\lambda+\mu}\|\theta_1 - \theta_2\|_2^2 - \frac{1}{\lambda+\mu}\|\nabla q(\theta_1) - \nabla q(\theta_2)\|_2^2$$

Now, following analysis indicates the desired results

$$\begin{aligned}
\|\theta^t + \alpha\nabla q(\theta^t) - \theta^*\|_2^2 &= \|\theta^t - \theta^*\|_2^2 + 2\alpha\nabla q(\theta^t)^\top(\theta^t - \theta^*) + \alpha^2\|\nabla q(\theta^t)\|_2^2 \\
&= \|\theta^t - \theta^*\|_2^2 + 2\alpha(\nabla q(\theta^t) - \nabla q(\theta^*))^\top(\theta^t - \theta^*) + \alpha^2\|\nabla q(\theta^t) - \nabla q(\theta^*)\|_2^2 \\
&\leq \|\theta^t - \theta^*\|_2^2 - 2\alpha\left(\frac{\lambda\mu}{\lambda+\mu}\|\theta^t - \theta^*\|_2^2 + \frac{1}{\lambda+\mu}\|\nabla q(\theta^t) - \nabla q(\theta^*)\|_2^2\right) \\
&\quad + \alpha^2\|\nabla q(\theta^t) - \nabla q(\theta^*)\|_2^2 \\
&= \left(1 - 2\alpha\frac{\lambda\mu}{\lambda+\mu}\right)\|\theta^t - \theta^*\|_2^2 + \left(\alpha^2 - \frac{2\alpha}{\lambda+\mu}\right)\|\nabla q(\theta^t) - \nabla q(\theta^*)\|_2^2 \\
&= \left(1 - \frac{4\lambda\mu}{(\lambda+\mu)^2}\right)\|\theta^t - \theta^*\|_2^2 = \left(\frac{\mu-\lambda}{\mu+\lambda}\right)^2\|\theta^t - \theta^*\|_2^2
\end{aligned}$$

□

Proof of Theorem 1. By definition of the first-order EM update

$$\theta^{t+1} = \theta^t + \alpha\nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t}, \quad \text{for } t = 0, 1, 2, \ldots$$

we have

$$\begin{aligned}\left\|\theta^t + \alpha\nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t} - \theta^*\right\|_2 &= \left\|\theta^t + \alpha\nabla q(\theta)|_{\theta=\theta^t} - \alpha\nabla q(\theta)|_{\theta=\theta^t} + \alpha\nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t} - \theta^*\right\|_2 \\ &\leq \left\|\theta^t + \alpha\nabla q(\theta)|_{\theta=\theta^t} - \theta^*\right\|_2 + \alpha\left\|\nabla q(\theta)|_{\theta=\theta^t} - \nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t}\right\|_2 \\ &\leq \left(\frac{\mu-\lambda}{\mu+\lambda}\right)\left\|\theta^t - \theta^*\right\|_2 + \alpha\gamma\left\|\theta^t - \theta^*\right\|_2\end{aligned}$$

The first inequality holds by triangle inequality and the second inequality holds by Lemma A.1 and Condition 3.1. Substituting $\alpha = \frac{2}{\mu+\lambda}$ and performing some algebra yields the claim.

$$\begin{aligned}\left\|\theta^t + \alpha\nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t} - \theta^*\right\|_2 &\leq \left(\frac{\mu-\lambda}{\mu+\lambda}\right)\left\|\theta^t - \theta^*\right\|_2 + \alpha\gamma\left\|\theta^t - \theta^*\right\|_2 \\ &= \left(\frac{\mu-\lambda}{\mu+\lambda}\right)\left\|\theta^t - \theta^*\right\|_2 + \frac{2\gamma}{\lambda+\mu}\left\|\theta^t - \theta^*\right\|_2 \\ &= \frac{2\gamma+\mu-\lambda}{\lambda+\mu}\left\|\theta^t - \theta^*\right\|_2 = \left(1 - \frac{2\lambda-2\gamma}{\mu+\lambda}\right)\left\|\theta^t - \theta^*\right\|_2 \\ &\leq \left(1 - \frac{2\lambda-2\gamma}{\mu+\lambda}\right)^t\left\|\theta^0 - \theta^*\right\|_2\end{aligned}$$

□

A.2 Additional Proof Details for Sample-based Analysis

Proof of Theorem 2. With probability at least $1-\delta$ we have that for any $\theta^s \in \mathbb{B}_2\left(r;\theta^*\right)$

$$\left\|\nabla Q_n\left(\theta|\theta^s\right)|_{\theta=\theta^s} - \nabla Q\left(\theta|\theta^s\right)|_{\theta=\theta^s}\right\|_2 \leq \varepsilon_Q^{\text{unif}}\left(n,\delta\right) \tag{8}$$

We perform the remainder of our analysis under this event. Defining $\kappa = \left(1 - \frac{2\lambda-2\gamma}{\lambda+\mu}\right)$, we make following claim

Claim A.1. For each iteration $s \in \{0,1,2,\ldots\}$,

$$\left\|\theta^{s+1} - \theta^*\right\|_2 \leq \kappa\left\|\theta^s - \theta^*\right\|_2 + \alpha\varepsilon_Q^{\text{unif}}\left(n,\delta\right) \tag{9}$$

We prove this claim by induction on the iteration number. The base case is $s=0$. In this case, we have

$$\begin{aligned}\left\|\theta^1 - \theta^*\right\|_2 &= \left\|\theta^0 + \alpha\nabla Q_n\left(\theta|\theta^0\right)\big|_{\theta=\theta^0} - \theta^*\right\|_2 \\ &\leq \left\|\theta^0 + \alpha\nabla Q\left(\theta|\theta^0\right)\big|_{\theta=\theta^0} - \theta^*\right\|_2 + \alpha\left\|\nabla Q\left(\theta|\theta^0\right)\big|_{\theta=\theta^0} - \nabla Q_n\left(\theta|\theta^0\right)\big|_{\theta=\theta^0}\right\| \\ &\leq \kappa\left\|\theta^0 - \theta^*\right\|_2 + \alpha\varepsilon_Q^{\text{unif}}\left(n,\delta\right)\end{aligned}$$

The first inequality holds by triangle inequality and the second inequality holds by Theorem 1. By condition $\varepsilon_Q^{\text{unif}}\left(n,\delta\right) \leq (\lambda-\gamma)r$, we have

$$\begin{aligned}\left\|\theta^1 - \theta^*\right\|_2 &\leq \kappa\left\|\theta^0 - \theta^*\right\|_2 + \alpha\varepsilon_Q^{\text{unif}}\left(n,\delta\right) \\ &\leq \left(1 - \frac{2\lambda-2\gamma}{\lambda+\mu}\right)r + \frac{2(\lambda-\gamma)r}{\lambda+\mu} = r\end{aligned}$$

This implies $\theta^1 \in \mathbb{B}_2(r;\theta^*)$. In the induction from $s \mapsto s+1$, suppose that $\left\|\theta^s - \theta^*\right\|_2 \leq r$, and the bound (9) holds at iteration s. The same argument then implies that the bound (9) also holds

for iteration $s+1$, and that $\left\|\theta^{s+1}-\theta^*\right\|_2 \leq r$. Therefore, by induction the claim holds for each iteration $s \in \{0,1,2,3,\dots\}$.

Using this claim, we can show that

$$\begin{aligned}
\left\|\theta^t-\theta^*\right\|_2 &\leq \kappa\left\|\theta^{t-1}-\theta^*\right\|_2+\alpha\varepsilon_Q^{\mathrm{unif}}(n,\delta) \\
&\leq \kappa\left\{\kappa\left\|\theta^{t-2}-\theta^*\right\|_2+\alpha\varepsilon_Q^{\mathrm{unif}}(n,\delta)\right\}+\alpha\varepsilon_Q^{\mathrm{unif}}(n,\delta) \\
&\leq \kappa^t\left\|\theta^0-\theta^*\right\|_2+\left\{\sum_{s=0}^{t-1}\kappa^s\right\}\alpha\varepsilon_Q^{\mathrm{unif}}(n,\delta) \\
&\leq \kappa^t\left\|\theta^0-\theta^*\right\|_2+\frac{\alpha}{1-\kappa}\varepsilon_Q^{\mathrm{unif}}(n,\delta)
\end{aligned}$$

This yields the result in theorem. □

A.3 Additional Details for the Extension of Results to the EM Algorithm

In this subsection, we develop some results for the extension of first order EM algorithm.

A.3.1 Analysis of the EM Algorithm at the Population Level

We assume throughout this section that the function q is λ -strongly concave (but not necessarily smooth). In order to compactly represent the EM update, we define the operator $M : \Omega \to \Omega$

$$M(\theta) = \arg\max_{\theta' \in \Omega} Q\left(\theta'|\theta\right) \tag{10}$$

Using this notation, the EM algorithm given some initialization θ^0, produces a sequence of iterates $\left\{\theta^t\right\}_{t=0}^{\infty}$, where $\theta^{t+1} = M\left(\theta^t\right)$. By virtue of the self-consistency property and the convexity of Ω, the fixed point satisfies the first-order optimality (KKT) condition

$$\left\langle \nabla Q\left(\theta^*|\theta^*\right), \theta'-\theta^*\right\rangle \leq 0 \quad \text{for all } \theta' \in \Omega \tag{11}$$

Similarly, for any $\theta \in \Omega$, since $M(\theta)$ maximizes the function $\theta' \mapsto Q\left(\theta'|\theta\right)$ over Ω, we have

$$\left\langle \nabla Q(M(\theta)|\theta), \theta'-M(\theta)\right\rangle \leq 0 \quad \text{for all } \theta' \in \Omega \tag{12}$$

Now we introduce the following regularity condition in order to relate conditions (12) and (11): The condition involves a Euclidean ball of radius r around the fixed point θ^*, given by

$$\mathbb{B}_2\left(r;\theta^*\right) := \left\{\theta \in \Omega \mid \left\|\theta-\theta^*\right\|_2 \leq r\right\} \tag{13}$$

Definition A.1. The functions $\{Q(\cdot|\theta), \theta \in \Omega\}$ satisfy condition FOS(γ) over $\mathbb{B}_2\left(r;\theta^*\right)$ if

$$\left\|\nabla Q\left(M(\theta)|\theta^*\right)-\nabla Q(M(\theta)|\theta)\right\|_2 \leq \gamma\left\|\theta-\theta^*\right\|_2 \tag{14}$$

for all $\theta \in \mathbb{B}_2\left(r;\theta^*\right)$.

Theorem 1. *For some radius $r>0$ and pair (γ,λ) such that $0 \leq \gamma < \lambda$ suppose that the function $Q\left(\cdot|\theta^*\right)$ is globally λ-strongly concave (5), and that the FOS(γ) condition (14) holds on the ball $\mathbb{B}_2\left(r;\theta^*\right)$. Then the population EM operator M is contractive over $\mathbb{B}_2\left(r;\theta^*\right)$, in particular with*

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

for all $\theta \in \mathbb{B}_2\left(r;\theta^*\right)$.

Proof. Since both $M(\theta)$ and θ^* are in Ω, we may apply condition (11) with $\theta' = M(\theta)$ and condition (12) with $\theta' = \theta^*$ with some algebras. Then we will obtain

$$\langle \nabla Q\left(M(\theta)|\theta^*\right) - \nabla Q\left(\theta^*|\theta^*\right), \theta^* - M(\theta)\rangle \leq \langle \nabla Q\left(M(\theta)|\theta^*\right) - \nabla Q(M(\theta)|\theta), \theta^* - M(\theta)\rangle \tag{15}$$

Applying λ-strong concavity condition (5) on LHS gives us

$$\langle \nabla Q\left(M(\theta)|\theta^*\right) - \nabla Q\left(\theta^*|\theta^*\right), \theta^* - M(\theta)\rangle \geq \lambda \left\|\theta^* - M(\theta)\right\|_2^2 \tag{16}$$

Applying the FOS(γ) condition together with the Cauchy-Schwarz inequality on RHS gives us

$$\langle \nabla Q\left(M(\theta)|\theta^*\right) - \nabla Q(M(\theta)|\theta), \theta^* - M(\theta)\rangle \leq \gamma \left\|\theta^* - M(\theta)\right\|_2 \left\|\theta - \theta^*\right\|_2 \tag{17}$$

Combining the inequalities above the gives us

$$\lambda \left\|\theta^* - M(\theta)\right\|_2^2 \leq \gamma \left\|\theta^* - M(\theta)\right\|_2 \left\|\theta - \theta^*\right\|_2$$

canceling term $\left\|\theta^* - M(\theta)\right\|_2$ yields the result. □

A.3.2 Finite-sample Analysis for the EM Algorithm

We now turn to theoretical results on the sample-based version of the EM algorithm. Define the sample-based operator $M_n : \Omega \to \Omega$

$$M_n(\theta) = \arg\max_{\theta' \in \Omega} Q_n\left(\theta'|\theta\right) \tag{18}$$

where the sample-based Q-function was defined in equation (1). Given the sample size n and T iterations, the full data set is divided into T subsets of size $\lfloor n/T \rfloor$. The mini-batch EM Algorithm updates $\theta^{t+1} = M_{n/T}\left(\theta^t\right)$. For the tolerance parameter $\delta \in (0,1)$, we let $\varepsilon_M(n,\delta)$ be the smallest scalar such that for any fixed $\theta \in \mathbb{B}_2\left(r;\theta^*\right)$

$$\mathbb{P}\left(\left\|M_n(\theta) - M(\theta)\right\|_2 \leq \varepsilon_M(n,\delta)\right) \geq 1 - \delta \tag{19}$$

On the other hand, in order to analyze the standard sample-based form of EM, define $\varepsilon_M^{\text{unif}}\left(n,\delta\right)$ to be the smallest scalar for which

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \left\|M_n(\theta) - M(\theta)\right\|_2 \leq \varepsilon_M^{\text{unif}}(n,\delta) \tag{20}$$

with probability at least $1 - \delta$. With these definitions, we have the following guarantees.

Theorem 2. *Suppose that the population EM operator $M : \Omega \to \Omega$ is contractive with parameter $\kappa \in (0,1)$ on the ball $\mathbb{B}_2\left(r;\theta^*\right)$, and the initial vector θ^0 belongs to $\mathbb{B}_2\left(r;\theta^*\right)$*

(a) If the sample size n is large enough to ensure that

$$\varepsilon_M^{unif}\left(n,\delta\right) \leq (1-\kappa)r \tag{21}$$

then the EM iterates $\left\{\theta^t\right\}_{t=0}^{\infty}$ satisfy the bound

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{1}{1-\kappa}\varepsilon_M^{\text{unif}}(n,\delta) \tag{22}$$

(b) For a given iteration number T, suppose the sample size n is large enough to ensure that

$$\varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \leq (1-\kappa)r \tag{23}$$

Then the sample-splitting EM iterates $\left\{\theta^t\right\}_{t=0}^{T}$ based on $\frac{n}{T}$ samples per round satisfy the bound

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{1}{1-\kappa}\varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \tag{24}$$

Remark A.1. The proof for this Theorem is similar to Theorem 2 and Theorem 3.

B Additional Details for the Applications on Specific Statistical Models

B.1 Gaussian Mixture Models for Population Level and Sample-based Methods

Consider the two-component Gaussian mixture model with balanced weights and isotropic covariances. It can be specified by a density of the form

$$f_\theta(y) = \frac{1}{2}\phi\left(y; \theta^*, \sigma^2 I_d\right) + \frac{1}{2}\phi\left(y; -\theta^*, \sigma^2 I_d\right) \tag{25}$$

where $\phi(\cdot; \mu, \Sigma)$ denotes the density of a $\mathcal{N}(\mu, \Sigma)$ random vector in $\mathbb{R}^d$, and we have assumed that the two components are equally weighted. Suppose that the variance σ^2 is known, so that our goal is to estimate the unknown mean vector θ^* In this example, the hidden variable $Z \in \{0, 1\}$ is an indicator variable for the underlying mixture component, that is,

$$(Y|Z = 0) \sim \mathcal{N}\left(-\theta^*, \sigma^2 I_d\right) \quad \text{and} \quad (Y|Z = 1) \sim \mathcal{N}\left(\theta^*, \sigma^2 I_d\right)$$

Suppose that we are given n i.i.d. samples $\{y_i\}_{i=1}^n$ drawn from the mixture density (25). In the most general case (two components weighted by $w_\theta(y)$), the complete data $\{(y_i, z_i)\}_{i=1}^n$ corresponds to the original samples along with the component indicator variables $z_i \in \{0, 1\}$. The sample-based function Q_n takes the form

$$Q_n\left(\theta'|\theta\right) = -\frac{1}{2n}\sum_{i=1}^n \left[w_\theta\left(y_i\right)\left\|y_i - \theta'\right\|_2^2 + \left(1 - w_\theta\left(y_i\right)\right)\left\|y_i + \theta'\right\|_2^2\right] \tag{26}$$

where

$$w_\theta(y) := e^{-\frac{\|\theta - y\|_2^2}{2\sigma^2}}\left[e^{-\frac{\|\theta - y\|_2^2}{2\sigma^2}} + e^{-\frac{\|\theta + y\|_2^2}{2\sigma^2}}\right]^{-1}$$

- EM updates: This example is especially simple in that each iteration of the EM algorithm has a closed form solution, given by

$$\theta^{t+1} := \arg\max_{\theta' \in \mathbb{R}^d} Q_n\left(\theta'|\theta^t\right) = \frac{2}{n}\sum_{i=1}^n w_{\theta^t}\left(y_i\right) y_i - \frac{1}{n}\sum_{i=1}^n y_i \tag{27}$$

  Iterations of the population EM algorithm are specified analogously

$$\theta^{t+1} = 2\mathbb{E}\left[w_{\theta^t}(Y)Y\right] \tag{28}$$

  where the empirical expectation has been replaced by expectation under the mixture distribution (25).

- First-order EM updates: On the other hand, the sample-based and population first-order EM operators with step size $\alpha > 0$ are given by

$$\theta^{t+1} = \theta^t + \alpha\left\{\frac{1}{n}\sum_{i=1}^n \left(2w_{\theta^t}\left(y_i\right) - 1\right) y_i - \theta^t\right\}, \quad \text{and} \quad \theta^{t+1} = \theta^t + \alpha\left[2\mathbb{E}\left[w_{\theta^t}(Y)Y\right] - \theta^t\right] \tag{29}$$

Before we step into formal analysis on mixture of Gaussians, we provide some technical results related to the mixture of Gaussians

- For the function $f(t) = \frac{t^2}{\exp(\mu t)}$, we have

$$\sup_{t\in[0,\infty]} f(t) = \frac{4}{(e\mu)^2}, \quad \text{achieved at } t^* = \frac{2}{\mu} \tag{30}$$

and

$$\sup_{t\in[t^*,\infty]} f(t) = f(t^*), \quad \text{for } t^* \geq \frac{2}{\mu} \tag{31}$$

- For the function $g(t) = \frac{1}{(\exp(t)+\exp(-t))^2}$, we have

$$g(t) \leq \frac{1}{4} \quad \text{for all } t \in \mathbb{R} \tag{32}$$

and for any $\mu \geq 0$

$$\sup_{t\in[\mu,\infty]} g(t) \leq \frac{1}{(\exp(\mu)+\exp(-\mu))^2} \leq \exp(-2\mu) \tag{33}$$

- Similarly, for the function $g^2(t) = \frac{1}{(\exp(t)+\exp(-t))^4}$, we have

$$g^2(t) \leq \frac{1}{16} \quad \text{for all } t \in \mathbb{R} \tag{34}$$

and for any $\mu \geq 0$

$$\sup_{t\in[\mu,\infty]} g^2(t) \leq \frac{1}{(\exp(\mu)+\exp(-\mu))^4} \leq \exp(-4\mu) \tag{35}$$

B.1.1 Population Level Analysis

The difficulty of estimating such a mixture model can be characterized by the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$, and our analysis requires the SNR to be lower bounded as

$$\frac{\|\theta^*\|_2}{\sigma} > \eta \tag{36}$$

With the signal-to-noise ratio lower bound η defined above, we have the following guarantee.

Corollary B.1 (Population result for the first-order EM algorithm for Gaussian)**.** *Consider a Gaussian mixture model for which the SNR condition (36) holds for a sufficiently large η, and define the radius $r = \frac{\|\theta^*\|_2}{4}$. Then there is a contraction coefficient $\kappa(\eta) \leq e^{-c\eta^2}$ where c is a universal constant such that for any initialization $\theta^0 \in \mathbb{B}_2(r;\theta^*)$, the population first-order EM iterates with step size 1, satisfy the bound*

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 \tag{37}$$

for all $t = 1, 2, \ldots$

To prove the Corollary B.1, we need following lemma:

Lemma B.1. Under the conditions of Corollary B.1, there is a constant $\gamma \in (0,1)$ with $\gamma \leq \exp\left(-c_2\eta^2\right)$ such that

$$\|\mathbb{E}\left[2\Delta_w(Y)Y\right]\|_2 \leq \gamma \|\theta - \theta^*\|_2 \tag{38}$$

where $\Delta_w(y) := w_\theta(y) - w_{\theta^*}(y)$.

Proof. Denote $\|\cdot\|_{\text{op}}$ to be the operator norm and $\|\cdot\|_{\text{fro}}$ to be the Frobenius norm. For each $u \in [0,1]$, define $\theta_u = \theta^* + u\Delta$, where $\Delta := \theta - \theta^*$. Applying Taylor's Theorem to the function $\theta \mapsto w_\theta(Y)$ followed by expectations, gives us

$$\mathbb{E}\left[Y\left(w_\theta(Y) - w_{\theta^*}(Y)\right)\right] = 2\int_0^1 \mathbb{E}\left[\frac{YY^T}{\sigma^2\left(\exp\left(-\frac{\langle\theta_u, Y\rangle}{\sigma^2}\right) + \exp\left(\frac{\langle\theta_u, Y\rangle}{\sigma^2}\right)\right)^2}\right]\Delta du$$

and denote

$$\Gamma_u(Y) = \frac{YY^T}{\sigma^2\left(\exp\left(-\frac{\langle\theta_u, Y\rangle}{\sigma^2}\right) + \exp\left(\frac{\langle\theta_u, Y\rangle}{\sigma^2}\right)\right)^2}$$

For each choice of $u \in [0,1]$, we have $\Gamma_u(y) = \Gamma_u(-y)$. Note that the distribution of Y is symmetric around zero, we see that $\mathbb{E}\left[\Gamma_u(Y)\right] = \mathbb{E}\left[\Gamma_u(\tilde{Y})\right]$, where $\tilde{Y} \sim \mathcal{N}\left(\theta^*, \sigma^2 I\right)$, and hence that

$$\left\|\mathbb{E}\left[\left(w_\theta(Y) - w_{\theta^*}(Y)\right)Y\right]\right\|_2 \leq 2\sup_{u\in[0,1]}\left\|\mathbb{E}\left[\Gamma_u(\tilde{Y})\right]\right\|_{\text{op}}\|\Delta\|_2 \tag{39}$$

Now we need to bound the value of $\left\|\mathbb{E}\left[\Gamma_u(\tilde{Y})\right]\right\|_{\text{op}}$ uniformly over $u \in [0,1]$.

For an arbitrary fixed $u \in [0,1]$, let R be an rotating matrix such that $R\theta_u = \|\theta_u\|_2 e_1$, where $e_1 \in \mathbb{R}^d$ is the first canonical basis vector. Define the rotated random vector $V = R\tilde{Y}$ and note that $V \sim \mathcal{N}\left(R\theta^*, \sigma^2 I\right)$. Using this transformation, the operator norm of the matrix $\mathbb{E}\left[\Gamma_u(\tilde{Y})\right]$ is equal to that of

$$D = \mathbb{E}\left[\frac{VV^T}{\sigma^2\left(\exp\left(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right) + \exp\left(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right)\right)^2}\right]$$

Define

$$\alpha_1 := \mathbb{E}\left[\frac{V_1^2}{\sigma^2\left(\exp\left(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right) + \exp\left(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right)\right)^2}\right]$$

$$\alpha_2 := \mathbb{E}\left[\frac{V_1}{\sigma^2\left(\exp\left(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right) + \exp\left(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right)\right)^2}\right]$$

$$\alpha_3 := \mathbb{E}\left[\frac{1}{\sigma^2\left(\exp\left(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right) + \exp\left(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\right)\right)^2}\right]$$

and

$$\mu := R\theta^* \qquad \nu := [0, \mu_2, \mu_3, \ldots, \mu_d]^\top$$

Immediately, we have

$$
\begin{aligned}
e_1 e_1^\top &= \operatorname{diag}\{1, 0, \ldots, 0\} \\
\nu e_1^\top &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \mu_2 & 0 & 0 & \ldots & 0 \\ \mu_3 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_d & 0 & 0 & \ldots & 0 \end{bmatrix} \\
e_1 \nu^\top &= \begin{bmatrix} 0 & \mu_2 & \mu_3 & \ldots & \mu_d \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \\
e_1 \nu^\top &= \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & \mu_2^2 & \mu_2\mu_3 & \ldots & \mu_2\mu_d \\ 0 & \mu_3\mu_2 & \mu_3^2 & \ldots & \mu_3\mu_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mu_d\mu_2 & \mu_d\mu_3 & \ldots & \mu_d^2 \end{bmatrix}
\end{aligned}
$$

Therefore, we have

$$
D = \alpha_1 e_1 e_1^\top + \alpha_2 \left(\nu e_1^\top + e_1 \nu^\top \right) + \alpha_3 \nu \nu^T
$$

and

$$
\|D\|_{\text{op}} \leq \left[D \|_{\text{fro}} \leq \alpha_1 + 2\alpha_2 \| \nu \|_2 + \alpha_3 \| \nu \|_2^2 \leq \alpha_1 + 2\alpha_2 \| \theta^* \|_2 + \alpha_3 \| \theta^* \|_2^2 \right. \tag{40}
$$

Define the event, we have following claims bound the values α_1, α_2 and α_3 whenever $\frac{\|\theta^*\|_2^2}{\sigma^2} \geq \eta^2 \geq 16/3$:

Claim B.1.

$$
\alpha_1 \leq \frac{16\sigma^2}{9e^2 \|\theta^*\|_2^2} e^{-\frac{\|\theta^*\|_2^2}{32\sigma^2}} + \frac{\|\theta^*\|_2^2}{16\sigma^2} e^{-\frac{3\|\theta^*\|^2}{8\sigma^2}}
$$

Note that for α_1, we have

$$
\alpha_1 \leq \mathbb{E}\left[\frac{V_1^2/\sigma^2}{\exp\left(\frac{2\|\theta_u\|_2 V_1}{\sigma^2} \right)} \right]
$$

Conditioning on the event $\mathcal{E}$ and $\mathcal{E}^c$ yields

$$
\alpha_1 \leq \mathbb{E}\left[\frac{V_1^2/\sigma^2}{\exp\left(\frac{2\|\theta_u\|_2 V_1}{\sigma^2} \right)} \middle| \mathcal{E} \right] \mathbb{P}[\mathcal{E}] + \mathbb{E}\left[\frac{V_1^2/\sigma^2}{\exp\left(\frac{2\|\theta_u\|_2 V_1}{\sigma^2} \right)} \middle| \mathcal{E}^c \right]
$$

Applying equation (30) and (31) yields that when $\|\theta^*\|_2 \|\theta_u\|_2 \geq 4\sigma^2$

$$
\alpha_1 \leq \frac{\sigma^2}{e^2 \|\theta_u\|_2^2} \mathbb{P}[\mathcal{E}] + \frac{\|\theta^*\|_2^2}{16\sigma^2 \exp\left(\frac{\|\theta_u\|_2 \|\theta^*\|_2}{2\sigma^2} \right)}
$$

Note that in Corollary B.1, we have the condition $r = \frac{\|\theta^*\|_2}{4}$, therefore, we have

$$
\|\theta_u\|_2 = \|\theta^* + u(\theta - \theta^*)\|_2 \geq \|\theta^*\|_2 - \frac{1}{4}\|\theta^*\|_2 = \frac{3}{4}\|\theta^*\|_2 \tag{41}
$$

Thus, when $\|\theta^*\|_2^2 \geq 16\sigma^2/3$, we have

$$\alpha_1 \leq \frac{16\sigma^2}{9e^2\left\|\theta^*\right\|^2}\mathbb{P}(\mathcal{E}) + \frac{\|\theta^*\|_2^2 \exp\left(-\frac{3\|\theta^*\|_2^2}{8\sigma^2}\right)}{16\sigma^2}$$

Using equation 41, the mean of V_1 is lower bounded as

$$\begin{aligned}\mathbb{E}\left[V_1\right] &= \langle R\theta^*, e_1\rangle = \langle R\theta_u, e_1\rangle + \langle R\left(\theta^* - \theta_u\right), e_1\rangle \\ &\geq \|\theta_u\|_2 - \|\theta^* - \theta_u\|_2 \\ &\geq \frac{\|\theta^*\|_2}{2}\end{aligned}$$

Combining with by standard Gaussian tail bound

$$\mathbb{P}[\mathcal{E}] \leq \exp\left(\frac{-\|\theta^*\|_2^2}{32\sigma^2}\right) \tag{42}$$

gives us when $\|\theta^*\|_2^2 \geq 16\sigma^2/3$

$$\alpha_1 \leq \frac{16\sigma^2}{9e^2\|\theta^*\|_2^2}e^{-\frac{\|\theta^*\|_2^2}{32\sigma^2}} + \frac{\|\theta^*\|_2^2}{16\sigma^2}e^{-\frac{3\|\theta^*\|^2}{8\sigma^2}}$$

This proves our claim on α_1. The next claim bounds the value of α_2

Claim B.2.

$$\alpha_2 \leq \frac{2\|\theta^*\|_2}{\sigma^2}\exp\left(-\frac{\|\theta^*\|_2^2}{64\sigma^2}\right)$$

Similar to the case for bounding α_1, we have

$$\begin{aligned}\alpha_2 &= \mathbb{E}\left[\frac{V_1}{\sigma^2\left(\exp\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right) + \exp\left(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\right)^2}\right] \\ &\leq \sqrt{\mathbb{E}\left[\frac{V_1^2}{\sigma^2}\right]}\sqrt{\mathbb{E}\left[\frac{1}{\sigma^2\left(\exp\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right) + \exp\left(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\right)^4}\right]}\end{aligned}$$

For the first term on RHS of the inequality for α_2, we have

$$\mathbb{E}\left[\frac{V_1^2}{\sigma^2}\right] \leq \frac{\|\theta^*\|_2^2}{\sigma^2}$$

For the second term on RHS of the inequality for α_2, applying the equation (32) gives us

$$\mathbb{E}\left[\frac{1}{\sigma^2\left(\exp\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right) + \exp\left(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\right)^4}\right] = \frac{1}{\sigma^2}\mathbb{E}\left[g^2\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\right]$$

Conditioning on the event $\mathcal{E}$ and $\mathcal{E}^c$ yields

$$\begin{aligned}\frac{1}{\sigma^2}\mathbb{E}\left[g^2\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\right] &\leq \frac{1}{\sigma^2}\left[\mathbb{E}\left[g^2\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\middle|\mathcal{E}\right]\mathbb{P}[\mathcal{E}] + \mathbb{E}\left[g^2\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\middle|\mathcal{E}^c\right]\right] \\ &\leq \frac{1}{\sigma^2}\left[\frac{1}{16}\mathbb{P}[\mathcal{E}] + \exp\left(-\frac{\|\theta^*\|_2 \|\theta_u\|_2}{\sigma^2}\right)\right] \\ &\leq \frac{1}{\sigma^2}\left[\frac{1}{16}\mathbb{P}[\mathcal{E}] + \exp\left(-\frac{3\|\theta^*\|_2^2}{4\sigma^2}\right)\right]\end{aligned}$$

The second line follows by applying bound (32) to the first term, and the bound (33) with

$$\mu = \frac{\|\theta^*\|_2 [\theta_u\|_2}{4\sigma^2}$$

to the second term. The last inequality follows from the bound (41). Besides, using the bound (42) on $\mathbb{P}[\mathcal{E}]$ gives us

$$\frac{1}{\sigma^2}\mathbb{E}\left[g^2\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)\right] \leq \frac{1}{\sigma^2}\left[\frac{1}{16}\exp\left(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\right) + \exp\left(-\frac{3\|\theta^*\|_2^2}{4\sigma^2}\right)\right] \leq \frac{2}{\sigma^2}\exp\left(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\right)$$

Combining all inequalities together gives us the bound for α_2:

$$\alpha_2 \leq \frac{2\|\theta^*\|_2}{\sigma^2}\exp\left(-\frac{\|\theta^*\|_2^2}{64\sigma^2}\right)$$

Claim B.3.

$$\alpha_3 \leq \frac{2}{\sigma^2}\exp\left(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\right)$$

The derivation for α_3 is quite similar to that of α_2. Again, we first condition on the event $\mathcal{E}$ and $\mathcal{E}^c$, then apply the bound (32), (33), (41) and (42).

$$\begin{aligned}\alpha_3 &\leq \frac{1}{\sigma^2}\left[\mathbb{E}\left[g\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)|\mathcal{E}\right]\mathbb{P}[\mathcal{E}] + \mathbb{E}\left[g\left(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\right)|\mathcal{E}^c\right]\right] \\ &\leq \frac{1}{\sigma^2}\left[\frac{1}{4}\mathbb{P}[\mathcal{E}] + \exp\left(-\frac{\|\theta^*\|_2 \|\theta_u\|_2}{4\sigma^2}\right)\right] \\ &\leq \frac{1}{\sigma^2}\left[\frac{1}{4}\mathbb{P}[\mathcal{E}] + \exp\left(-\frac{3\|\theta^*\|_2^2}{16\sigma^2}\right)\right] \\ &\leq \frac{1}{\sigma^2}\left[\frac{1}{4}\exp\left(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\right) + \exp\left(-\frac{3\|\theta^*\|_2^2}{16\sigma^2}\right)\right] \\ &\leq \frac{2}{\sigma^2}\exp\left(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\right)\end{aligned}$$

Combining with three claims together, we have

$$\|2\mathbb{E}\left[\left(w_\theta(Y) - w_{\theta^*}(Y)\right)Y\right]\|_2 \leq c_1\left(1 + \frac{1}{\eta^2} + \eta^2\right)e^{-c_2\eta^2}\|\theta - \theta^*\|_2$$

if $\frac{\|\theta^*\|_2^2}{\sigma^2} \geq \eta^2 \geq 16/3$. Therefore, the upper bound (38) holds when the signal-to-noise ratio is sufficiently large. □

Using this result, we now can prove the Corollary B.1.

Proof of Corollary B.1. Note that scaling the family of Q functions by a fixed constant does not affect any of our conditions and their consequences. Thus, re-scaling Q functions by constant σ^2 does not effect the final results. In order to apply Theorem 1, we need to verify the λ-concavity (5) and μ-smoothness (6) conditions, and the GS(γ) condition (4) over the ball $\mathbb{B}_2(r;\theta^*)$. In this case, the q-function takes the form

$$q(\theta) = Q(\theta|\theta^*) = -\frac{1}{2}\mathbb{E}\left[w_{\theta^*}(Y)\|Y-\theta\|_2^2 + (1-w_{\theta^*}(Y))\|Y+\theta\|_2^2\right]$$

where the weighting function is given by

$$w_\theta(y) := \frac{\exp\left(-\|\theta-y\|_2^2/\left(2\sigma^2\right)\right)}{\exp\left(-\|\theta-y\|_2^2/\left(2\sigma^2\right)\right) + \exp\left(-\|\theta+y\|_2^2/\left(2\sigma^2\right)\right)}$$

The q-function is smooth and strongly-concave with parameters 1. Using the Lemma above, we now verify the GS condition (4). By smoothness and strong-concavity, we only need to show

$$\|\mathbb{E}\left[2\Delta_w(Y)Y\right]\|_2 < \|\theta-\theta^*\|_2$$

This claim follows immediately from the Lemma above. Thus, the GS condition holds when $\gamma < 1$. The bound on the contraction parameter follows from the fact that $\gamma \le \exp\left(-c_2\eta^2\right)$ and applying Theorem 1 yields Corollary B.1. □

B.1.2 Finite Sample Analysis

Before we step into finite sample analysis on first-order EM algorithm, we introduce some concepts that we may use during the analysis.

Definition B.1. A random variable X is said to be Rademacher Random Variable if it has following density function

$$f(k) = \frac{1}{2}(\delta(k-1)+\delta(k+1)) = \begin{cases} 1/2 & \text{if } k=-1 \\ 1/2 & \text{if } k=+1 \\ 0 & \text{otherwise} \end{cases}$$

We now analyze the convergence rate for sample-based first-order EM updates for the Gaussian mixture model. In this part, the step size is $\alpha = 1$. Consider the function

$$\varphi(\sigma;\|\theta^*\|_2) := \|\theta^*\|_2\left(1+\frac{\|\theta^*\|_2^2}{\sigma^2}\right)$$

and positive universal constants (c, c_1, c_2).

Corollary B.2 (Sample-based first-order EM guarantees for Gaussian mixture)**.** *In addition to the conditions of Corollary B.1, suppose that the sample size is lower bounded as $n \ge c_1 d\log(1/\delta)$. Then given any initialization $\theta^0 \in \mathbb{B}_2\left(\frac{\|\theta^*\|_2}{4};\theta^*\right)$, there is a contraction coefficient $\kappa(\eta) \le e^{-c\eta^2}$ such that the first order EM iterates $\left\{\theta^t\right\}_{t=0}^{\infty}$ satisfy the bound*

$$\left\|\theta^t-\theta^*\right\|_2 \le \kappa^t\left\|\theta^0-\theta^*\right\|_2 + \frac{c_2}{1-\kappa}\varphi(\sigma;\|\theta^*\|_2)\sqrt{\frac{d}{n}\log(1/\delta)} \tag{43}$$

Proof. By definition, with probability at least $1-\delta$,

$$\sup_{\theta\in\mathbb{B}_2(r;\theta^*)} \|\nabla Q_n(\theta|\theta) - \nabla Q(\theta|\theta)\|_2 \le \varepsilon_Q^{\text{unif}}(n,\delta)$$

and first-order EM algorithm updates

$$\theta^{t+1} = \theta^t + \alpha\left\{\frac{1}{n}\sum_{i=1}^n (2w_{\theta^t}(y_i)-1)\,y_i - \theta^t\right\}, \quad \text{and} \quad \theta^{t+1} = \theta^t + \alpha\left[2\mathbb{E}\left[w_{\theta^t}(Y)Y\right] - \theta^t\right]$$

Define the set

$$\mathbb{A} := \left\{\theta\in\mathbb{R}^d \middle|\, \|\theta-\theta^*\|_2 \le \|\theta^*\|_2/4\right\}$$

and random variable

$$Z := \sup_{\theta\in\mathrm{A}} \left\|\alpha\left\{\frac{1}{n}\sum_{i=1}^n (2w_\theta(y_i)-1)\,y_i - \theta\right\} - \alpha\left[2\mathbb{E}\left[w_\theta(Y)Y\right]-\theta\right]\right\|_2$$

To prove the corollary, we need to give an upper bound on $\varepsilon_Q^{\text{unif}}(n,\delta)$ that is similar to $\varepsilon_Q^{\text{unif}}(n,\delta) \le (\lambda-\gamma)r$. Now we make following claim on Z

Claim B.4. With sufficiently large constants c_1, c_2, for $n \ge c_1 d\log(1/\delta)$, we have

$$Z \le \frac{c_2\|\theta^*\|_2\left(\|\theta^*\|_2^2+\sigma^2\right)}{\sigma^2}\sqrt{\frac{d\log(1/\delta)}{n}}$$

with probability at least $1-\delta$.

We prove this claim by following steps:

- Step 1: Reduce our goal supremum on set $\mathbb{A}$ to a finite maximum over the sphere $\mathbb{S}^d$.

  For each $u\in\mathbb{R}^d$ with $\|u\|_2 = 1$, define

  $$Z_u := \sup_{\theta\in\mathbb{A}}\left\{\frac{1}{n}\sum_{i=1}^n (2w_\theta(y_i)-1)\langle y_i,u\rangle - \mathbb{E}\,(2w_\theta(Y)-1)\langle Y,u\rangle\right\}$$

  Therefore, for any pair (u,v) we have

  $$|Z_u - Z_v| \le Z\|u-v\|_2$$

  Let $\left\{u^1,\ldots,u^M\right\}$ be a $\frac{1}{2}$-covering of the sphere $S^d = \left\{v\in\mathbb{R}^d \middle| \|v\|_2 = 1\right\}$. That is, for any $v\in\mathbb{S}^d$, there exists some index $j\in[M]$ such that $\left\|v-u^j\right\|_2 \le \frac{1}{2}$. Then we can write

  $$Z_v \le Z_{u^j} + |Z_v - Z_{u^j}| \le \max_{j\in[M]} Z_{u^j} + Z\left\|v-u^j\right\|_2$$

  Combining all results above gives us

  $$Z = \sup_{v\in\mathbb{S}^d} Z_v \le 2\max_{j\in[M]} Z_{u^j} \tag{44}$$

- Step 2: Introduce a sequence of Rademacher variables.

  Let $\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, for any $\lambda > 0$, one can show that

$$\mathbb{E}\left[e^{\lambda Z_w}\right] \leq \mathbb{E}\left[\exp\left(\frac{2}{n}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^n \varepsilon_i\left(2w_\theta\left(y_i\right)-1\right)\langle y_i,u\rangle\right)\right]$$

  by using a standard symmetrization result for empirical processes given by Koltchinski et al. Besides, for any d-vectors y, θ and θ', we have the Lipschitz property

$$|2w_\theta(y) - 2w_{\theta'}(y)| \leq \frac{1}{\sigma^2}\left|\langle\theta, y\rangle - \langle\theta', y\rangle\right|$$

  Using the results on the Ledoux-Talagrand contraction for Rademacher processes given by Koltchinski et al gives us

$$\mathbb{E}\left[\exp\left(\frac{2}{n}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^n \varepsilon_i\left(2w_\theta\left(y_i\right)-1\right)\langle y_i,u\rangle\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{4}{n\sigma^2}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^n \varepsilon_i\langle\theta,y_i\rangle\langle y_i,u\rangle\right)\right]$$

  For any $\theta \in$ A, we have

$$\|\theta\|_2 \leq \frac{5}{4}\|\theta^*\|_2$$

  therefore,

$$\sup_{\theta\in\mathbb{A}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle\theta,y_i\rangle\langle y_i,u\rangle \leq \frac{5}{4}\|\theta^*\|_2\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i y_i y_i^T\right\|_{\text{op}}$$

- Step 3: Bound the random variable Z_u for a fixed $u \in \mathbb{S}^d$.

  Based on the results in previous steps, we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda Z_u}\right] &\leq \mathbb{E}\left[\exp\left(\frac{10\lambda\|\theta^*\|_2}{\sigma^2}\max_{j\in[M]}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left\langle y_i,u^j\right\rangle^2\right)\right] \\ &\leq \sum_{j=1}^M \mathbb{E}\left[\exp\left(\frac{10\lambda\|\theta^*\|_2}{\sigma^2}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left\langle y_i,u^j\right\rangle^2\right)\right]\end{aligned} \tag{45}$$

  Using Rademacher sign variable, we can represent our Gaussian mixture model as

$$y = \eta\theta^* + w$$

  where η is a Rademacher sign variable, and $w \sim \mathcal{N}\left(0, \sigma^2 I\right)$. For any $u \in \mathbb{R}^d$, one can show that

$$\mathbb{E}\left[e^{\langle u,y\rangle}\right] = \mathbb{E}\left[e^{\eta\langle u,\theta^*\rangle}\right]\mathbb{E}\left[e^{\langle u,w\rangle}\right] \leq e^{\left(\|\theta^*\|_2^2+\sigma^2\right)/2}$$

  This result shows that if $\{y_i\}_{i=1}^n$ are n i.i.d. observations, then they are sub-Gaussians with parameter at most $\gamma = \sqrt{\|\theta^*\|_2^2 + \sigma^2}$. Therefore, $\varepsilon_i\langle y_i, u\rangle^2$ is zero mean sub-exponential, and the MGF is bounded as

$$\mathbb{E}\left[e^{t\varepsilon_i(y_i,u)^2}\right] \leq e^{\gamma^4 t^2/2}$$

  for all $t > 0$ that is sufficiently small. Combined with the result (45), we will see that

$$\mathbb{E}\left[e^{\lambda Z_u}\right] \leq Me^{c\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}} \leq e^{c\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+2d}$$

Combined with the Chernoff approach, the upper bound on the MGF indicates that for a sufficiently large constant c_1, when $n \geq c_1 d \log(1/\delta)$ we will see that

$$Z \leq \frac{c_2 \left\|\theta^*\right\|_2 \gamma^2}{\sigma^2} \sqrt{\frac{d \log(1/\delta)}{n}}$$

with probability at least $1 - \delta$. This proves our claims and thus yields the Corollary B.2. □

B.2 Mixture of Regressions (MOR)

Our second example is the mixture of regressions model. In the standard linear regression model, we have i.i.d. observations $(Y, X) \in \mathbb{R} \times \mathbb{R}^d$ and the model given by

$$y_i = \langle x_i, \theta^* \rangle + v_i \tag{46}$$

with following conditions:

- $v_i \sim \mathcal{N}\left(0, \sigma^2\right)$ is the independent noise.
- The covariates x_is are independent with v_is

Besides, in this example, we also assume that

- $x_i \sim \mathcal{N}(0, I)$
- $\theta^* \in \mathbb{R}^d$ is the unknown

In the mixture of regressions problem, we observe a pair (y_i, x_i) drawn from the model (46) with probability $\frac{1}{2}$, and alternative regression model

$$y_i = \langle x_i, -\theta^* \rangle + v_i$$

with probability 1/2. Therefore, the latent variables $\{z_i\}_{i=1}^n$ represent the labels of the underlying regression model.

$$z_i = \begin{cases} 1 & \text{if the observation is from (46)} \\ 0 & \text{otherwise} \end{cases}$$

In this problem setting, we derive the EM update rule as follows:

Define the weight function

$$w_\theta(x, y) = \frac{\exp\left(\frac{-(y - \langle x, \theta \rangle)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(y - \langle x, \theta \rangle)^2}{2\sigma^2}\right) + \exp\left(\frac{-(y + \langle x, \theta \rangle)^2}{2\sigma^2}\right)} \tag{47}$$

then the sample-based Q-function is given by

$$Q_n\left(\theta' | \theta\right) = -\frac{1}{2n} \sum_{i=1}^n \left(w_\theta\left(x_i, y_i\right) \left(y_i - \left\langle x_i, \theta' \right\rangle\right)^2 + \left(1 - w_\theta\left(x_i, y_i\right)\right) \left(y_i + \left\langle x_i, \theta' \right\rangle\right)^2 \right) \tag{48}$$

The sample-based EM algorithm updates θ by maximizing above Q-function. The closed form solution to the optimization problem is given by

$$\theta^{t+1} = \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \left(\sum_{i=1}^n \left(2 w_{\theta^t}\left(x_i, y_i\right) - 1\right) y_i x_i \right) \tag{49}$$

Similarly, the update rule for population level EM algorithm is given by

$$\theta^{t+1} = 2\mathbb{E}\left[w_{\theta^t}(X,Y)YX\right] \tag{50}$$

where the expectation is taken over the joint distribution of the pair $(Y, X) \in \mathbb{R} \times \mathbb{R}^d$. On the other hand, the update rules for first order EM algorithm, for population level and sample-based, are given by

$$\theta^{t+1} = \theta^t + \alpha \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\left(2w_{\theta^t}\left(x_i, y_i\right) - 1\right) y_i x_i - x_i x_i^T \theta^t \right] \right\} \tag{51}$$

and

$$\theta^{t+1} = \theta^t + \alpha \mathbb{E}\left[2w_{\theta^t}(X,Y)YX - \theta^t\right] \tag{52}$$

where $\alpha > 0$ is a step size parameter.

B.2.1 Population Level Analysis

There are some elementary results on Gaussian random vectors that are useful in proving the main results in this section.

Lemma B.2. Given a Gaussian random vector $X \sim \mathcal{N}(0, I)$ and any fixed vectors $u, v \in \mathbb{R}^d$, then we have

$$\mathbb{E}\left[\langle X, u\rangle^2 \langle X, v\rangle^2\right] \leq 3\|u\|_2^2 \|v\|_2^2 \tag{53}$$

and

$$\mathbb{E}\left[\langle X, u\rangle^4 \langle X, v\rangle^2\right] \leq 15\|u\|_2^4 \|v\|_2^2 \tag{54}$$

Proof. For any fixed orthonormal matrix $R \in \mathbb{R}^{d\times d}$, $R^T X$ also has a $\mathcal{N}(0, I)$ distribution and

$$\mathbb{E}\left[\langle X, u\rangle^2 \langle X, v\rangle^2\right] = \mathbb{E}\left[\langle X, Ru\rangle^2 \langle X, Rv\rangle^2\right]$$

By choosing R such that $Ru = \|u\|_2 e_1$ and defining $z = Rv$, we have

$$\begin{aligned}
\mathbb{E}\left[\langle X, Ru\rangle^2 \langle X, Rv\rangle^2\right] &= \mathbb{E}\left[\|u\|_2^2 X_1^2 \sum_{i=1}^{d}\sum_{j=1}^{d} X_i X_j z_i z_j\right] \\
&= \|u\|_2^2 \left(3z_1^2 + \left(\|z\|_2^2 - z_1^2\right)\right) \\
&\leq 3\|u\|_2^2 \|z\|_2^2 \\
&= 3\|u\|_2^2 \|v\|_2^2
\end{aligned}$$

A similar argument yields the second claim. □

As in our analysis of the Gaussian mixture model, our theory applies when the signal-to-noise ratio is sufficiently large, as enforced by a condition of the form

$$\frac{\|\theta^*\|_2}{\sigma} > \eta \tag{55}$$

for a sufficiently large constant $\eta > 0$. Under a suitable lower bound on this quantity, our first result guarantees that the first-order EM algorithm is locally convergent to the global optimum θ^* and provides a quantification of the local region of convergence.

Corollary B.3 (Population result for the first-order EM algorithm for MOR)**.** *Consider any mixture of regressions model satisfying the SNR condition (55) for a sufficiently large constant η, and define the radius $r := \frac{\|\theta^*\|_2}{32}$. Then for any $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, the population first-order EM iterates with stepsize 1, satisfy the bound*

$$\left\|\theta^t - \theta^*\right\|_2 \leq \left(\frac{1}{2}\right)^t \left\|\theta^0 - \theta^*\right\|_2 \tag{56}$$

for $t = 1, 2, \ldots$

Before we prove the Corollary B.3, we need following lemma.

Lemma B.3. Under the conditions of Corollary B.3, there is a constant $\gamma < 1/4$ such that for any fixed vector $\widetilde{\Delta}$ we have

$$\left|\mathbb{E}\left[\Delta_w(X,Y)(2Z-1)\langle X, \theta^*\rangle \langle X, \tilde{\Delta}\rangle\right]\right| \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2 \tag{57}$$

and

$$\left|\mathbb{E}\left[\Delta_w(X,Y)v\langle X, \widetilde{\Delta}\rangle\right]\right| \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2 \tag{58}$$

Proof. Since the standard deviation σ is known, we can assume that $\sigma = 1$ by a simple re-scaling, then the weight function in (47) is given by

$$w_\theta(x,y) = \frac{\exp\left(\frac{-(y-\langle x,\theta\rangle)^2}{2}\right)}{\exp\left(\frac{-(y-\langle x,\theta\rangle)^2}{2}\right) + \exp\left(\frac{-(y+\langle x,\theta\rangle)^2}{2}\right)} \tag{59}$$

Note that $\Delta = \theta - \theta^*$ and $\widetilde{\Delta}$ is any fixed vector in $\mathbb{R}^d \setminus \{0\}$. We can define $\theta_u = \theta^* + u\Delta$ for a scalar $u \in [0,1]$ and our assumptions guarantee that

$$\|\Delta\|_2 \leq \frac{\|\theta^*\|_2}{32}, \quad \text{and} \quad \|\theta^*\|_2 \geq \eta \tag{60}$$

and we also have

$$\|\theta_u\|_2 \geq \|\theta^*\|_2 - \|\Delta\|_2 \geq \frac{\|\theta^*\|_2}{2} \tag{61}$$

- Proof of Inequality (57). We split the proof of this bound into two separate cases
  - Case-1: $\|\Delta\|_2 \leq 1$
    In this case, we have

$$\frac{d}{d\theta} w_\theta(X,Y) = \frac{2YX}{(\exp(Y\langle X,\theta\rangle) + \exp(-Y\langle X,\theta\rangle))^2}$$

    Taylor series with integral form remainder on the function $\theta \mapsto w_\theta(X,Y)$ gives us

$$\Delta_w(X,Y) = \int_0^1 \frac{2Y\langle X,\Delta\rangle}{\left(\exp\left(Z_u\right) + \exp\left(-Z_u\right)\right)^2} du \tag{62}$$

    where $Z_u := Y\langle X, \theta^* + u\Delta\rangle$. Substituting for $\Delta_w(X,Y)$ in inequality (57), we see that it suffices to show

$$A_u := \int_0^1 \mathbb{E}\left[\frac{2Y\langle X,\theta^*\rangle}{\left(\exp\left(Z_u\right) + \exp\left(-Z_u\right)\right)^2}(2Z-1)\langle X,\Delta\rangle\langle X,\tilde{\Delta}\rangle\right] du \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2 \tag{63}$$

    for some $\gamma \in [0, 1/4)$. To show this result, we need following auxiliary results.

Claim B.5. There is $a\gamma \in [0, 1/4)$ such that for each $u \in [0, 1]$, we have

$$\sqrt{\mathbb{E}\left[\frac{Y^2 \langle X, \theta_u \rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\right]} \leq \frac{\gamma}{14} \tag{64}$$

and

$$\sqrt{\mathbb{E}\left[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\right]} \leq \frac{\gamma}{32} \tag{65}$$

whenever $\|\Delta\|_2 \leq 1$.

Using this claim, we can bound the quantity A_u from equation (63). Note that $\theta^* = \theta_u - u\Delta$, we have $A_u = B_1 + B_2$, where

$$B_1 := \mathbb{E}\left[\frac{2Y \langle X, \theta_u \rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \tilde{\Delta}\rangle\right]$$
$$B_2 := -\mathbb{E}\left[\frac{2Yu\langle X, \Delta\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\right]$$

To show $A_u \leq \frac{\gamma}{2}\|\Delta\|_2\|\tilde{\Delta}\|_2$, we need prove that $\max\{B_1, B_2\} \leq \frac{\gamma}{4}\|\Delta\|_2\|\tilde{\Delta}\|_2$. Thus, it is natural to bound B_1 and B_2.

* Bounding B_1: By the Cauchy-Schwarz inequality, we have

$$B_1 \leq \sqrt{\mathbb{E}\left[\frac{y^2 \langle X, \theta_u \rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\right]}\sqrt{\mathbb{E}\left[4(2Z - 1)^2\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\right]}$$
$$\leq \frac{\gamma}{14}\sqrt{\mathbb{E}\left[4\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\right]}$$

  where the second step follows from the bound (64) in the claim above, and the fact that $(2Z - 1)^2 = 1$.
  Next we observe that $\mathbb{E}\left[4\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\right] \leq 12\|\Delta\|_2^2\|\widetilde{\Delta}\|_2^2$, where we have used the bound (53) in Lemma B.2. Combined with our earlier bound, we have

$$B_1 \leq \frac{\gamma}{4}\|\Delta\|_2\|\widetilde{\Delta}\|_2$$

  as claimed.

* Bounding B_2: Similarly, by Cauchy-Schwarz inequality we have

$$B_2 \leq \sqrt{\mathbb{E}\left[\frac{y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\right]}\sqrt{\mathbb{E}\left[4u^2(2Z - 1)^2\langle X, \Delta\rangle^4\langle X, \widetilde{\Delta}\rangle^2\right]}$$
$$\leq \frac{\gamma}{32}\sqrt{\mathbb{E}\left[4u^2\langle X, \Delta\rangle^4\langle X, \widetilde{\Delta}\rangle^2\right]}$$

  where the second step follows from the bound (65), and the fact that $(2Z - 1)^2 = 1$. In this case, we have

$$\mathbb{E}\left[4u^2\langle X, \Delta\rangle^4\langle X, \widetilde{\Delta}\rangle^2\right] \leq 60\|\Delta\|_2^4\|\widetilde{\Delta}\|_2^2 \leq 60\|\Delta\|_2^2\|\widetilde{\Delta}\|_2^2$$

where the first step uses the bound (54) from Lemma B.2, and the second step follows from $\|\Delta\|_2 \leq 1$. Combining these two results, we have

$$B_2 \leq \frac{\gamma}{4}\|\Delta\|_2\|\widetilde{\Delta}\|_2$$

which completes the proof of inequality (57) in the case $\|\Delta\|_2 \leq 1$

– Case-2: $\|\Delta\|_2 \geq 1$.

Our argument in this case makes use of various probability bounds on different events, which we state here for future reference. These events are related to the scalar $\tau := C_\tau\sqrt{\log\|\theta^*\|_2}$ for a constant C_τ, and are also related to the vectors

$$\Delta := \theta - \theta^*, \text{ and } \theta_u := \theta^* + u\Delta \text{ for some fixed } u \in [0,1]$$

Claim B.6 (Event bounds)**.** We have following probability bounds for the events given below

(i) For the event $\mathcal{E}_1 := \{\text{sign}(\langle X, \theta^*\rangle) = \text{sign}(\langle X, \theta_u\rangle)\}$, we have $\mathbb{P}[\mathcal{E}_1^c] \leq \|\frac{\|\Delta\|_2}{\|\theta^*\|_2}$

(ii) For the event $\mathcal{E}_2 := \{|\langle X, \theta^*\rangle| > \tau\} \cap \{|\langle X, \theta_u\rangle| > \tau\} \cap \left\{|v| \leq \frac{\tau}{2}\right\}$, we have

$$\mathbb{P}[\mathcal{E}_2^c] \leq \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2\exp\left(-\frac{\tau^2}{2}\right)$$

(iii) For the event $\mathcal{E}_3 := \{|\langle X, \theta^*\rangle| \geq \tau\} \cup \{|\langle X, \theta_u\rangle| \geq \tau\}$, we have $\mathbb{P}[\mathcal{E}_3^c] \leq \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2}$

(iv) For the event $\mathcal{E}_4 := \{|v| \leq \tau/2\}$, we have $\mathbb{P}[\mathcal{E}_4^c] \leq 2e^{-\frac{\tau^2}{2}}$

(v) For the event $\mathcal{E}_5 := \{|\langle X, \theta_u\rangle| > \tau\}$, we have $\mathbb{P}[\mathcal{E}_5^c] \leq \frac{\tau}{\|\theta_u\|_2}$

(vi) For the event $\mathcal{E}_6 := \{|\langle X, \theta^*\rangle| > \tau\}$, we have $\mathbb{P}[\mathcal{E}_6^c] \leq \frac{\tau}{\|\theta^{-1}\|_2}$

Besides, we will also use the result on controlling the second moment matrix $\mathbb{E}\left[XX^T\right]$ when conditioned on some of the events given above:

Claim B.7 (Conditional covariance bounds)**.** Conditioned on any event $\mathcal{E} \in \{\mathcal{E}_1 \cap \mathcal{E}_2, \mathcal{E}_1^c, \mathcal{E}_5^c, \mathcal{E}_6^c\}$, we have

$$\left\|\mathbb{E}\left[XX^T|\mathcal{E}\right]\right\|_{op} \leq 2$$

Here our goal is to bound the quantity

$$T = \left|\mathbb{E}\left[\Delta_w(X,Y)(2Z-1)\langle X, \theta^*\rangle\langle X, \widetilde{\Delta}\rangle\right]\right| \leq \mathbb{E}\left[\left|\Delta_w(X,Y)(2Z-1)\langle X, \theta^*\rangle\langle X, \widetilde{\Delta}\rangle\right|\right]$$

For any measurable event $\mathcal{E}$, we define

$$\Psi(\mathcal{E}) := \mathbb{E}\left[\left|\Delta_w(X,Y)(2Z-1)\langle X, \theta^*\rangle\langle X, \widetilde{\Delta}\rangle\right| |\mathcal{E}\right]\mathbb{P}[\mathcal{E}]$$

and note that by successive conditioning, we have

$$T \leq \Psi(\mathcal{E}_1 \cap \mathcal{E}_2) + \Psi(\mathcal{E}_1^c) + \Psi(\mathcal{E}_4^c) + \Psi(\mathcal{E}_5^c) + \Psi(\mathcal{E}_6^c) \tag{66}$$

We bound each of these five terms in turn and summarized the results in following claim.

Claim B.8.

$$\Psi\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq 2\|\widetilde{\Delta}\|_2 \left\|\theta^*\right\|_2 e^{-\tau^2}$$
$$\Psi\left(\mathcal{E}_1^c\right) \leq \sqrt{\mathbb{E}\left[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_1^c\right]} \sqrt{\mathbb{E}\left[\langle X, \Delta\rangle^2 | \mathcal{E}_1^c\right]} \frac{\|\Delta\|_2}{\left\|\theta^*\right\|_2} \leq \frac{2\|\widetilde{\Delta}\|_2 \|\Delta\|_2^2}{\left\|\theta^*\right\|_2}$$
$$\Psi\left(\mathcal{E}_5^c\right) \leq \frac{2\tau\|\widetilde{\Delta}\|_2 \sqrt{\tau^2 + 2\|\Delta\|_2^2}}{\left\|\theta_u\right\|_2} \leq \frac{2\tau\|\widetilde{\Delta}\|_2 \|\Delta\|_2 \sqrt{\tau^2+2}}{\left\|\theta_u\right\|_2}$$
$$\Psi\left(\mathcal{E}_6^c\right) \leq \frac{\sqrt{2}\tau^2 \|\widetilde{\Delta}\|_2}{\left\|\theta^*\right\|_2}$$

We have thus obtained bounds on all five terms in the decomposition (66). We combine these bounds with the with lower bound $\|\theta_u\|_2 \geq \frac{\|\theta^*\|_2}{2}$ from equation (61) and then perform some algebra to obtain

$$T \leq c\|\Delta\|_2 \|\widetilde{\Delta}\|_2 \left\{ \frac{\tau^2}{\left\|\theta^*\right\|_2} + \left\|\theta^*\right\|_2 e^{-\tau^2/2} \right\} + 2\|\widetilde{\Delta}\|_2 \frac{\|\Delta\|_2^2}{\left\|\theta^*\right\|_2}$$

where c is a universal constant. In particular, selecting $\tau = c_\tau \sqrt{\log \left\|\theta^*\right\|_2}$ for a sufficient large constant c_τ and the constant η in (60) sufficiently large yields the inequality (57).

- Proof of Inequality (58): Similarly, we split the proof of this bound into two separate cases
  - Case-1: $\|\Delta\|_2 \leq 1$

    As before, by using Taylor expansion of the function $\theta \mapsto \Delta_w(X, Y)$, we only need to prove

    $$\int_0^1 \mathbb{E}\left[\frac{2Yv}{\left(\exp\left(Z_u\right) + \exp\left(-Z_u\right)\right)^2} \langle X, \Delta\rangle \langle X, \tilde{\Delta}\rangle \right] du \leq \frac{\gamma}{2}\|\Delta\|_2 \|\widetilde{\Delta}\|_2$$

    For any fixed $u \in [0, 1]$, we have

    $$\begin{aligned}
    \mathbb{E}\left[\frac{2Yv\langle X, \Delta\rangle \langle X, \widetilde{\Delta}\rangle}{\left(\exp\left(Z_u\right) + \exp\left(-Z_u\right)\right)^2} \right] &\leq \sqrt{\mathbb{E}\left[\frac{4Y^2}{\left(\exp\left(Z_u\right) + \exp\left(-Z_u\right)\right)^4} \right]} \sqrt{\mathbb{E}\left[v^2 \langle X, \Delta\rangle^2 \langle X, \widetilde{\Delta}\rangle^2 \right]} \\
    &\leq \sqrt{\mathbb{E}\left[\frac{4Y^2}{\left(\exp\left(Z_u\right) + \exp\left(-Z_u\right)\right)^4} \right]} \sqrt{3\|\Delta\|_2^2 \|\widetilde{\Delta}\|_2^2} \\
    &\leq \frac{\sqrt{3}\gamma}{16}\|\Delta\|_2 \|\widetilde{\Delta}\|_2
    \end{aligned}$$

    The first step follows from Cauchy-Schwarz inequality. The second step follows from inequality (53) in Lemma B.2, the independence of v and X, and the fact that $\mathbb{E}\left[v^2\right] = 1$. The last step follows from the bound (65) in claim B.5.
  - Case-2: $\|\Delta\|_2 \geq 1$

    By Cauchy-Schwarz inequality, it suffices show that

    $$\sqrt{\mathbb{E}\left[\Delta_w^2(X, Y)\right]} \leq \frac{\gamma}{2}$$

    In claim B.6, we have introduce the scalar $\tau := C_\tau \sqrt{\log \left\|\theta^*\right\|_2}$ and the events $\mathcal{E}_1$ and $\mathcal{E}_2$. For any measurable event $\mathcal{E}$, define the function

    $$\Psi(\mathcal{E}) = \mathbb{E}\left[\Delta_w^2(X, Y) | \mathcal{E}\right] \mathbb{P}[\mathcal{E}]$$

With this notation, by successive conditioning, we have the upper bound

$$\mathbb{E}\left[\Delta_w^2(X,Y)\right] \leq \Psi\left(\mathcal{E}_1^c\right) + \Psi\left(\mathcal{E}_1 \cap \mathcal{E}_2^c\right) + \Psi\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \tag{67}$$

We control each of these terms in turn and summarize the results in following claim

Claim B.9.

$$\begin{aligned}
&\Psi\left(\mathcal{E}_1^c\right) \leq 4\mathbb{P}\left[\mathcal{E}_1^c\right] \leq 4\frac{\|\Delta\|_2}{\|\theta^*\|_2} \\
&\Psi\left(\mathcal{E}_1 \cap \mathcal{E}_2^c\right) \leq 4\mathbb{P}\left[\mathcal{E}_2^c\right] \leq 4\left\{\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2e^{-\frac{\tau^2}{2}}\right\} \\
&\Psi\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq e^{-2\tau^2}
\end{aligned}$$

Now putting them together yields

$$\sqrt{\mathbb{E}\left[\Delta_w^2(X,Y)\right]} \leq \sqrt{4\frac{\|\Delta\|_2}{\|\theta^*\|_2} + 4\left\{\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2e^{-\frac{\tau^2}{2}}\right\} + e^{-2\tau^2}}$$

By choosing C_τ sufficiently large in the definition of τ and selecting the signal-to-noise constant η in condition (60) sufficiently large, the inequality follows.

□

Remark B.1. For detailed proof of Claim B.5, Claim B.6, Claim B.2.1 and Claim B.9, you can refer to Statistical Guarantees for the EM Algorithm: From Population to Sample-based Analysis

With above Lemmas established above, we now can prove the Corollary B.3.

Proof of Corollary B.3. To prove the corollary, we need to verify the condition (3.2) (λ-strong concavity) condition (3.3) (μ-smoothness) and condition (3.1) the GS(γ) over the ball $\mathbb{B}_2(r;\theta^*)$. Under the framework of MOR, the q-function is

$$q(\theta) = Q\left(\theta|\theta^*\right) = -\frac{1}{2}\mathbb{E}\left[w_{\theta^*}(X,Y)(Y - \langle X,\theta\rangle)^2 + (1 - w_{\theta^*}(X,Y))\left(Y + \langle X,\theta\rangle\right)^2\right]$$

where

$$w_\theta(x,y) := \frac{\exp\left(-(y - \langle x,\theta\rangle)^2/\left(2\sigma^2\right)\right)}{\exp\left(-(y - \langle x,\theta\rangle)^2/\left(2\sigma^2\right)\right) + \exp\left(-(y + \langle x,\theta\rangle)^2/\left(2\sigma^2\right)\right)}$$

Note that function $Q\left(\cdot|\theta^*\right)$ is λ-strongly concave and μ-smooth with λ and μ equal to the smallest and largest (resp.) eigenvalue of the matrix $\mathbb{E}\left[XX^T\right]$, thus the strong concavity and smoothness hold with $\lambda = \mu = 1$ since $\mathbb{E}\left[XX^T\right] = I$ by assumption.

To verify condition GS, we define the difference function

$$\Delta_w(X,Y) := w_\theta(X,Y) - w_{\theta^*}(X,Y)$$

and the difference vector

$$\Delta = \theta - \theta^*$$

Using the updates given by equation (52), we need to show that

$$\|2\mathbb{E}\left[\Delta_w(X,Y)YX\right]\|_2 < \|\Delta\|_2$$

This is equivalent to

$$\left\langle 2\mathbb{E}\left[\Delta_w(X,Y)YX\right], \widetilde{\Delta}\right\rangle < \|\Delta\|_2\|\widetilde{\Delta}\|_2 \qquad \text{for } \widetilde{\Delta} \in \mathbb{R}^d \setminus \{0\}$$

Under the framework of MOR, $Y \stackrel{d}{=} (2Z-1)\langle X, \theta^*\rangle + v$ where $Z \sim \text{Bernoulli}(1/2)$ and $v \sim \mathcal{N}(0,1)$. Thus, it suffices to show that

$$\mathbb{E}\left[\Delta_w(X,Y)(2Z-1)\langle X,\theta^*\rangle\langle X,\widetilde{\Delta}\rangle\right] + \mathbb{E}\left[\Delta_w(X,Y)v\langle X,\widetilde{\Delta}\rangle\right] \leq \gamma\|\Delta\|_2\|\tilde{\Delta}\|_2 \tag{68}$$

for $\gamma \in [0, 1/2)$ to establish contractivity. In order to prove the theorem with the desired upper bound on the coefficient of contraction we need to show (68) with $\gamma \in [0, 1/4)$. In Lemma, two bounds in conjunction imply that

$$\left\langle \mathbb{E}\left[\Delta_w(X,Y)YX\right], \tilde{\Delta}\right\rangle \leq \gamma\|\Delta\|_2\|\widetilde{\Delta}\|_2$$

with $\gamma \in [0, 1/4)$ and thus the corollary B.3 holds. □

Remark B.2. Note that in Gaussian mixture model the population likelihood has global maxima at θ^* and $-\theta^*$, and a local minimum at 0. Therefore, the radius of Euclidean ball over which the iterates could converge to θ^* should be less than $\|\theta^*\|_2$.

B.2.2 Sample-based Analysis

Our result on sample based analysis for MOR is related to the quantity

$$\varphi\left(\sigma; \|\theta^*\|_2\right) = \sqrt{\sigma^2 + \|\theta^*\|_2^2}$$

with positive universal constants (c_1, c_2)

Corollary B.4. *In addition to the conditions of Corollary B.3, suppose that the sample size is lower bounded as $n \geq c_1 d\log(T/\delta)$. Then there is a contraction coefficient $\kappa \leq 1/2$ such that, for any initial vector $\theta^0 \in \mathbb{B}_2\left(\frac{\|\theta^*\|_2}{32}; \theta^*\right)$, the sample-splitting first-order EM iterates with stepsize 1, based on n/T samples per step satisfy the bound*

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 + c_2\varphi\left(\sigma; \|\theta^*\|_2\right)\sqrt{\frac{d}{n}T\log(T/\delta)} \tag{69}$$

with probability at least $1-\delta$

Proof. Similar to Corollary B.2, our goal is to bound $\varepsilon_Q(n,\delta)$ defined as

$$\mathbb{P}\left[\left\|\nabla Q_n\left(\theta|\theta^t\right)\big|_{\theta=\theta^t} - \nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta^t}\right\|_2 > \varepsilon_Q(n,\delta)\right] \leq 1-\delta$$

For the first-order EM updates for MOR, we need to control the random variable,

$$Z := \left\|\alpha\left\{\frac{1}{n}\sum_{i=1}^n \left(2w_\theta\left(y_i\right)-1\right)y_i - \theta\right\} - \alpha\left[2\mathbb{E}\left[w_\theta(Y)Y\right]-\theta\right]\right\|_2$$

Claim B.10. There are universal constants (c_1, c_2) such that given a sample size $n \geq c_1 d\log(1/\delta)$, we have

$$\mathbb{P}\left[Z > \frac{c_2\left\|\theta^*\right\|_2\left(\left\|\theta^*\right\|_2^2 + \sigma^2\right)}{\sigma^2}\sqrt{\frac{d\log(1/\delta)}{n}}\right] \leq \delta$$

When step size is $\alpha = 1$, we have

$$Z \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \left(2w_\theta\left(x_i, y_i\right) - 1\right) y_i x_i - \mathbb{E}\left(2w_\theta(X,Y) - 1\right) YX \right\|_2 + \left\| I - \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right\|_{\text{op}} \|\theta\|_2$$

Define

$$\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \quad \text{with } \Sigma = \mathbb{E}\left[XX^T\right] = I$$

$$\widehat{v} := \frac{1}{n} \sum_{i=1}^{n} \left[\mu_\theta\left(x_i, y_i\right) y_i x_i\right] \quad \text{with } v := \mathbb{E}\left[\mu_\theta(X,Y) YX\right]$$

where $\mu_\theta(x, y) := 2w_\theta(x, y) - 1$. Noting that $\mathbb{E}[YX] = 0$, we have the bound

$$Z \leq \underbrace{\|\hat{v} - v\|_2}_{T_1} + \underbrace{\|\hat{\Sigma} - \Sigma\|_{\text{op}} \|\theta\|_2}_{T_2} \tag{70}$$

We bound each of the terms T_1 and T_2 in turn.

- Bounding T_1: Let us write $\|\widehat{v} - v\|_2 = \sup_{u \in \mathbb{S}^d} Z(u)$, where

$$Z(u) := \frac{1}{n} \sum_{i=1}^{n} \mu_\theta\left(x_i, y_i\right) y_i \langle x, u\rangle - \mathbb{E}\left[\mu_\theta(X,Y) Y \langle X, u\rangle\right]$$

  The discretization over a 1/2-cover of the sphere $\mathbb{S}^d$ – say $\left\{u^1, \ldots, u^M\right\}$ gives us

$$\|\widehat{v} - v\|_2 \leq 2 \max_{j \in [M]} Z\left(u^j\right)$$

  Thus, it suffices to control the random variable $Z(u)$ for a fixed $u \in \mathbb{S}^d$. By a standard symmetrization argument we have

$$\mathbb{P}[Z(u) \geq t] \leq 2\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta\left(x_i, y_i\right) y_i \left\langle x_i, u\right\rangle \geq t/2\right]$$

  where $\{\varepsilon_i\}_{i=1}^n$ are an i.i.d. sequence of Rademacher variables. Define

$$\mathcal{E} = \left\{\frac{1}{n} \sum_{i=1}^{n} \left\langle x_i, u\right\rangle^2 \leq 2\right\}$$

  Then we have $\mathbb{P}\left[\mathcal{E}^c\right] \leq e^{-n/32}$ because each variable $\langle x_i, u\rangle$ is sub-Gaussian with parameter one. Therefore, we have

$$\mathbb{P}[Z(u) \geq t] \leq 2\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta\left(x_i, y_i\right) y_i \left\langle x_i, u\right\rangle \geq \frac{t}{2} \mid \mathcal{E}\right] + 2e^{-n/32}$$

  Using the Ledoux-Talagrand contraction for Rademacher processes and the fact that $|\mu_\theta(x, y)| \leq 1$ for all pairs (x, y) gives us

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta\left(x_i, y_i\right) y_i \left\langle x_i, u\right\rangle\right) \middle| \mathcal{E}\right] \leq \mathbb{E}\left[\exp\left(\frac{2\lambda}{n} \sum_{i=1}^{n} \varepsilon_i y_i \left\langle x_i, u\right\rangle\right) \middle| \mathcal{E}\right]$$

Note that conditioned on x_i, y_i is zero-mean and sub-Gaussian with parameter at most $\sqrt{\|\theta^*\|_2^2 + \sigma^2}$. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i y_i \langle x_i, u\rangle\right)\middle|\mathcal{E}\right] &\leq \left[\exp\left(\frac{4\lambda^2}{n^2}\left(\|\theta^*\|_2^2+\sigma^2\right)\sum_{i=1}^{n}\langle x_i,u\rangle^2\right)\middle|\mathcal{E}\right] \\ &\leq \exp\left(\frac{8\lambda^2}{n}\left(\|\theta^*\|_2^2+\sigma^2\right)\right)\end{aligned}$$

where expectations are taken over the distribution $(y_i|x_i)$ for each index i and the second inequality follows from the definition of $\mathcal{E}$. Now applying this bound on the MGF gives us

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mu_\theta\left(x_i, y_i\right) y_i \langle x_i, u\rangle \geq t/2\middle|\mathcal{E}\right] \leq \exp\left(-\frac{nt^2}{256\left(\|\theta^*\|_2^2+\sigma^2\right)}\right)$$

Note that 1/2-cover of the unit sphere $\$^d$ has at most 2^d elements. Therefore, there is a universal constant c such that

$$T_1 \leq c\sqrt{\|\theta^*\|_2^2+\sigma^2}\sqrt{\frac{d}{n}\log(1/\delta)}$$

with probability at least $1-\delta$.

- Bouding T_2:

  On the one hand, we have

  $$\|\hat{\Sigma}-\Sigma\|_{\text{op}} \leq c\sqrt{\frac{d}{n}\log(1/\delta)}$$

  with probability at least $1-\delta$ based on standard results in random matrix theory. On the other hand, note that each iteration decreases the distance to θ^* and the iterate satisfies $\|\theta\|_2 \leq 2\|\theta^*\|_2$. Therefore, we have

  $$T_2 \leq c\|\theta^*\|_2\sqrt{\frac{d}{n}\log(1/\delta)}$$

  with probability at least $1-\delta$.

Finally, applying bounds on T_1 and T_2 on the decomposition (70) yields the claim. □

Remark B.3. In Corollary B.4, the bound (69) provides guidance on the number of iterations to perform. For a given sample size n, suppose we perform $T = \left\lceil \log\left(n/d\varphi^2\left(\sigma; \|\theta^*\|_2\right)\right)\right\rceil$ iterations. Then based on the bound (69), we have

$$\left\|\theta^T-\theta^*\right\|_2 \leq c_3\varphi\left(\sigma;\|\theta^*\|_2\right)\sqrt{\frac{d}{n}\log^2\left(\frac{n}{d\varphi^2\left(\sigma;\|\theta^*\|_2\right)}\right)\log(1/\delta)} \tag{71}$$

with probability at least $1-\delta$. Besides, Corollary B.4 predicts that the statistical error $\left\|\theta^t-\theta^*\right\|_2$ should decrease geometrically and then end at a plateau.

B.3 Linear Regression with Missing Covariates

In traditional linear regression, our training/testing data $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$ are generated based on the linear model (46). Instead of observing the covariate vector $x_i \in \mathbb{R}^d$ directly, we now deal with the case that some of covariates might be missing. Consider $\tilde{x}_i \in \mathbb{R}^d$ with components

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, & \text{with probability } 1-\rho \\ *, & \text{with probability } \rho \end{cases} \tag{72}$$

where $\rho \in [0, 1)$ is the probability of missingness.

For a given sample (x, y), let x_{obs} denote the observed portion of x and θ_{obs} denote the corresponding sub-vector of θ. Define the missing portions x_{mis} to be the missing portion of x and θ_{mis} to be the corresponding sub-vector of θ.

W.L.O.G. we can assume that the coordinates are permuted and the missing values are in the first block. Under the framework of the Linear Regression with Missing Covariates, the EM algorithm imputes the conditional mean and conditional covariance using the current parameter estimate θ. Immediately, the conditional mean of X given (x_{obs}, y) is

$$\mu_\theta(x_{\text{obs}}, y) := \begin{bmatrix} \mathbb{E}(x_{\text{mis}}|x_{\text{obs}}, y, \theta) \\ x_{\text{obs}} \end{bmatrix} = \begin{bmatrix} U_\theta z_{\text{obs}} \\ x_{\text{obs}} \end{bmatrix} \tag{73}$$

where

$$U_\theta = \frac{1}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2} \begin{bmatrix} -\theta_{\text{mis}}\theta_{\text{obs}}^T & \theta_{\text{mis}} \end{bmatrix} \quad \text{and} \quad z_{\text{obs}} := \begin{bmatrix} x_{\text{obs}} \\ y \end{bmatrix} \in \mathbb{R}^{|x_{\text{obs}}|+1} \tag{74}$$

Similarly, for the second moment matrix, we have

$$\Sigma_\theta(x_{\text{obs}}, y) := \mathbb{E}\left[XX^T|x_{\text{obs}}, y, \theta\right] = \begin{bmatrix} I & U_\theta z_{\text{obs}} x_{\text{obs}}^T \\ x_{\text{obs}} z_{\text{obs}}^T U_\theta^T & x_{\text{obs}} x_{\text{obs}}^T \end{bmatrix} \tag{75}$$

For a given parameter θ, the EM update is to maximize

$$Q_n(\theta'|\theta) := -\frac{1}{2n}\sum_{i=1}^n \left\langle \theta', \Sigma_\theta(x_{\text{obs},i}, y_i)\,\theta' \right\rangle + \frac{1}{n}\sum_{i=1}^n y_i \left\langle \mu_\theta(x_{\text{obs},i}, y_i), \theta' \right\rangle \tag{76}$$

The sample-based EM iterations are given as

$$\theta^{t+1} := \left[\sum_{i=1}^n \Sigma_{\theta^t}(x_{\text{obs},i}, y_i)\right]^{-1} \left[\sum_{i=1}^n y_i \mu_{\theta^t}(x_{\text{obs},i}, y_i)\right] \tag{77}$$

and the population EM iterations are given as

$$\theta^{t+1} := \left\{\mathbb{E}\left[\Sigma_{\theta^t}(X_{\text{obs}}, Y)\right]\right\}^{-1} \mathbb{E}\left[Y\mu_{\theta^t}(X_{\text{obs}}, Y)\right] \tag{78}$$

On the other hand, the sample-based first-order EM algorithm with step size α performs

$$\theta^{t+1} = \theta^t + \alpha\left\{\frac{1}{n}\sum_{i=1}^n \left[y_i\mu_{\theta^t}(x_{\text{obs},i}, y_i) - \Sigma_{\theta^t}(x_{\text{obs},i}, y_i)\,\theta^t\right]\right\} \tag{79}$$

and the population-based first order EM algorithm with step size α performs

$$\theta^{t+1} = \theta^t + \alpha\mathbb{E}\left[Y\mu_{\theta^t}(X_{\text{obs}}, Y) - \Sigma_{\theta^t}(X_{\text{obs}}, Y)\,\theta^t\right] \tag{80}$$

B.3.1 Population Analysis

In this example, we also give some conditions on the signal-to-noise ratio and the radius of contractivity r (the radius of the region around θ^* within which the population EM algorithm is convergent to a global optimum). Define

$$\xi_1 := \frac{\|\theta^*\|_2}{\sigma} \quad \text{and} \quad \xi_2 := \frac{r}{\sigma} \tag{81}$$

The condition for ρ in our first corollary is given by

$$\rho < \frac{1}{1 + 2\xi(1+\xi)} \quad \text{where } \xi := (\xi_1 + \xi_2)^2 \tag{82}$$

Corollary B.5 (Population contractivity for missing covariates)**.** *Given any missing covariate regression model with missing probability ρ satisfying the bound (82) the first-order EM iterates with stepsize* 1*, satisfy the bound*

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 \quad \textit{for } t = 1, 2, \ldots \tag{83}$$

where $\kappa \equiv \kappa(\xi, \rho) := \left(\frac{\xi + \rho(1+2\xi(1+\xi))}{1+\xi}\right)$.

Proof. As before, we need to verify three conditions (μ-smooth, λ-strongly concave and that the GS condition) are satisfied for the function q which takes the form

$$q(\theta) = \frac{1}{2}\left\langle \theta, \mathbb{E}\left[\Sigma_{\theta^*}\left(X_{\text{obs}}, Y\right)\right]\theta\right\rangle - \left\langle \mathbb{E}\left[Y\mu_{\theta^*}\left(X_{\text{obs}}, Y\right)\right], \theta\right\rangle$$

where the vector $\mu_{\theta^*} \in \mathbb{R}^d$.

- Smoothness and strong concavity.

  Note that

  $$\nabla^2 q(\theta) = \mathbb{E}\left[\Sigma_{\theta^*}\left(X_{\text{obs}}, Y\right)\right]$$

  By fixing a pattern of missingness and then averaging over (X_{obs}, Y), it is easy to verify that

  $$\mathbb{E}\left[\Sigma_{\theta^*}\left(X_{\text{obs}}, Y\right)\right] = \begin{bmatrix} I & U_{\theta^*}\begin{bmatrix} I \\ \theta^{*T}_{\text{obs}} \end{bmatrix} \\ \begin{bmatrix} I & \theta^*_{\text{obs}} \end{bmatrix} U^T_{\theta^*} & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

  Therefore, smoothness and strong concavity hold with $\mu = \lambda = 1$.

- GS Condition.

  To verify this condition, we need to show that there is a scalar $\gamma \in [0, 1)$ such that

  $$\|\mathbb{E}[V]\|_2 \leq \gamma \left\|\theta - \theta^*\right\|_2$$

  where the vector

  $$V = V\left(\theta, \theta^*\right) = \Sigma_{\theta^*}\left(X_{\text{obs}}, Y\right)\theta - Y\mu_{\theta^*}\left(X_{\text{obs}}, Y\right) - \Sigma_{\theta}\left(X_{\text{obs}}, Y\right)\theta + Y\mu_{\theta}\left(X_{\text{obs}}, Y\right) \tag{84}$$

  We can compute the expectation over (X_{obs}, Y) assuming that the first block is missing as below

  $$\mathbb{E}_{X_{\text{obs}}, Y}[V] = \begin{bmatrix} (\theta_{\text{mis}} - \theta^*_{\text{mis}}) + \pi_1\theta_{\text{mis}} \\ \pi_2\left(\theta_{\text{obs}} - \theta^*_{\text{obs}}\right) \end{bmatrix} \tag{85}$$

where

$$\pi_1 := \frac{\|\theta^*_{\text{mis}}\|_2^2 - \|\theta_{\text{mis}}\|_2^2 + \|\theta_{\text{obs}} - \theta^*_{\text{obs}}\|_2^2}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2}$$
$$\pi_2 := \frac{\|\theta_{\text{mis}}\|_2^2}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2}$$

Claim B.11. π_1 and π_2 can be bounded independently of the missingness pattern by

$$\pi_1 \leq 2\left(\xi_1 + \xi_2\right)\frac{\|\theta - \theta^*\|_2}{\sigma} \quad \text{and} \quad \pi_2 \leq \delta := \frac{1}{1 + \left(1/\left(\xi_1 + \xi_2\right)\right)^2} < 1 \tag{86}$$

To prove this claim, note that $\|\theta_{\text{mis}}\|_2 - \|\theta^*_{\text{mis}}\|_2 \leq \|\theta_{\text{mis}} - \theta^*_{\text{mis}}\|_2$ thus, by assumption, we have

$$\|\theta_{\text{mis}}\|_2 \leq \|\theta^*_{\text{mis}}\|_2 + \xi_2\sigma \leq \left(\xi_1 + \xi_2\right)\sigma \tag{87}$$

and thus

$$\begin{aligned}\|\theta^*_{\text{mis}}\|_2^2 - \|\theta_{\text{mis}}\|_2^2 &= \left(\|\theta_{\text{mis}}\|_2 - \|\theta^*_{\text{mis}}\|_2\right)\left(\|\theta_{\text{mis}}\|_2 + \|\theta^*_{\text{mis}}\|_2\right) \\ &\leq \left(2\xi_1 + \xi_2\right)\sigma\,\|\theta_{\text{mis}} - \theta^*_{\text{mis}}\|_2\end{aligned}$$

Therefore, the bound for π_1 follows from $\|\theta_{\text{obs}} - \theta^*_{\text{obs}}\|_2^2 \leq \xi_2\sigma\,\|\theta_{\text{obs}} - \theta^*_{\text{obs}}\|_2$. Besides,

$$\pi_2 = \frac{\|\theta_{\text{mis}}\|_2^2}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2} = \frac{1}{1 + \sigma^2/\|\theta_{\text{mis}}\|_2^2} \overset{\text{(i)}}{=} \frac{1}{1 + \left(1/\left(\xi_1 + \xi_2\right)\right)^2} < 1$$

where the first inequality follows from equation (87)

Now using the results in the claim above, we can then average over the missing pattern. Note that each coordinate is missing independently with probability ρ, we have

$$|\mathbb{E}[V]|_i \leq |\rho|\theta_i - \theta_i^*| + \rho\pi_1\,|\theta_i| + (1-\rho)\pi_2|\theta_i - \theta_i^*\|$$

Let $\eta := (1-\rho)\delta + \rho < 1$, we have

$$\begin{aligned}\|\mathbb{E}[V]\|_2^2 &\leq \eta^2\,\|\theta - \theta^*\|_2^2 + \rho^2\pi_1^2\|\theta\|_2^2 + 2\pi_1\eta\rho\,\|\theta, \theta - \theta^*\| \\ &\leq \left\{\eta^2 + \rho^2\|\theta\|_2^2\frac{4\left(\xi_1 + \xi_2\right)^2}{\sigma^2} + \frac{4\eta\rho\|\theta\|_2\left(\xi_1 + \xi_2\right)}{\sigma}\right\}\|\theta - \theta^*\|_2^2\end{aligned}$$

Define

$$\gamma^2 := \eta^2 + \rho^2\|\theta\|_2^2\frac{4\left(\xi_1 + \xi_2\right)^2}{\sigma^2} + \frac{4\eta\rho\|\theta\|_2\left(\xi_1 + \xi_2\right)}{\sigma}$$

By assumption, we have $\|\theta^*\|_2 \leq \xi_1\sigma$ and $\|\theta - \theta^*\|_2 \leq \xi_2\sigma$, and hence $\|\theta\|_2 \leq \left(\xi_1 + \xi_2\right)\sigma$. Thus, the coefficient γ^2 can be bounded as

$$\gamma^2 \leq \eta^2 + 4\rho^2\left(\xi_1 + \xi_2\right)^4 + 4\eta\rho\left(\xi_1 + \xi_2\right)^2$$

Under the conditions of the corollary, we have $\gamma < 1$ which completes the proof.

□

B.3.2 Sample-based Analysis

Corollary B.6 (Sample-splitting first-order EM guarantees for missing covariates)**.** *In addition to the conditions of Corollary B.5 suppose that the sample size is lower bounded as $n \geq c_1 d \log(1/\delta)$. Then there is a contraction coefficient $\kappa < 1$ such that, for any initial vector $\theta^0 \in \mathbb{B}_2(\xi_2\sigma; \theta^*)$, the sample-splitting first-order EM iterates with step size 1, based on n/T samples per iteration satisfy the bound*

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{c_2\sqrt{1+\sigma^2}}{1-\kappa}\sqrt{\frac{d}{n}T\log(T/\delta)} \tag{88}$$

with probability at least $1-\delta$

Proof. Similar to Corollary B.4, our goal is to bound $\varepsilon_Q(n,\delta)$ defined as

$$\mathbb{P}\left[\left\|\nabla Q_n\left(\theta|\theta^t\right)\big|_{\theta=\theta'} - \nabla Q\left(\theta|\theta^t\right)\big|_{\theta=\theta'}\right\|_2 > \varepsilon_Q(n,\delta)\right] \leq 1-\delta$$

For any fixed $\theta \in \mathbb{B}_2(r;\theta^*) = \left\{\theta \in \mathbb{R}^d \| \theta - \theta^*\|_2 \leq \xi_2\sigma\right\}$, we need to upper bound the random variable,

$$Z = \left\|\frac{1}{n}\sum_{i=1}^n \left[y_i\mu_\theta\left(x_{\text{obs},i}, y_i\right) - \Sigma_\theta\left(x_{\text{obs},i}, y_i\right)\theta\right] - \mathbb{E}\left[Y\mu_\theta\left(X_{\text{obs}}, Y\right) - \Sigma_\theta\left(X_{\text{obs}}, Y\right)\theta\right]\right\|_2$$

with high probability. We define:

$$T_1 := \left\|\left[\mathbb{E}\Sigma_\theta\left(x_{\text{obs}}, y\right)\theta - \frac{1}{n}\sum_{i=1}^n \Sigma_\theta\left(x_{\text{obs},i} + y_i\right)\theta\right]\right\|_2$$

$$T_2 := \left\| \left[\mathbb{E}\left(y\mu_\theta\left(x_{\text{obs}}, y\right)\right) - \frac{1}{n}\sum_{i=1}^n y_i\mu_\theta\left(x_{\text{obs},i}, y_i\right)\right]\right\|_2$$

Let $z_i \in \mathbb{R}^d$ be i-th canonical vector with ones in the positions of observed covariates and the notation $\odot$ be the element-wise product. Now we bound T_1 and T_2 respectively.

- Controlling T_1: Define

$$\bar{\Sigma} = \mathbb{E}\left[\Sigma_\theta\left(x_{\text{obs}}, y\right)\right]$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \Sigma_\theta\left(x_{\text{obs},i}, y_i\right)$$

  Then we have

$$T_1 \leq \|\bar{\Sigma} - \widehat{\Sigma}\|_{\text{op}}\|\theta\|_2 \leq \|\bar{\Sigma} - \widehat{\Sigma}\|_{\text{op}}\left(\xi_1 + \xi_2\right)\sigma$$

  where the second step follows since any vector $\theta \in \mathbb{B}_2(r;\theta^*)$ has ℓ_2-norm bounded as $\|\theta\|_2 \leq (\xi_1+\xi_2)\sigma$.

  Claim B.12. For any fixed vector $u \in S^d$, the random variable $\langle u, (\bar{\Sigma} - \widehat{\Sigma})u\rangle$ is zero-mean and sub-exponential.

  The key idea is to use the expression

$$\Sigma_\theta\left(x_{\text{obs}}, y\right) = I_{\text{mis}} + \mu_\theta\mu_\theta^T - \left((1-z)\odot\mu_\theta\right)\left((1-z)\odot\mu_\theta\right)^T$$

  where I_{mis} is the identity matrix on the diagonal sub-block corresponding to the missing entries. Since the square of any sub-Gaussian random variable has sub-exponential tails. Thus, we only need to show that each of the random variables $\langle\mu_\theta, u\rangle$, and $\langle(1-z)\odot\mu_\theta, u\rangle$ are sub-Gaussian. To show this, we need to verify that μ_θ is sub-Gaussian and here we state this argument as following Lemma

Lemma B.4. Under the conditions of Corollary B.5, the random vector $\mu_\theta(x_{\text{obs}}, y)$ is sub-Gaussian with a constant parameter.

For the detailed proof of this lemma and the claim above, you can refer to Page 115 in Statistical Guarantees for the EM Algorithm: From Population to Sample-based Analysis.

By above claim and referring to some standard arguments in random matrix theory, we have

$$\|\bar{\Sigma} - \widehat{\Sigma}\|_{\text{op}} \leq c\sqrt{\frac{d}{n}\log(1/\delta)}$$

with probability at least $1 - \delta$ when $n > d$.

- Controlling T_2: Note that

$$T_2 = \sup_{\|u\|_2=1} \left| \mathbb{E}\left[y \left\langle \mu_\theta(x_{\text{obs}}, y), u \right\rangle\right] - \frac{1}{n}\sum_{i=1}^{n} y_i \left\langle \mu_\theta(x_{\text{obs},i}, y_i), u \right\rangle \right|$$

Using the discretization argument with a 1/2-cover $\left\{u^1, \ldots, u^M\right\}$ of the sphere with $M \leq 2^d$ elements gives us

$$T_2 \leq 2 \max_{j\in[M]} \left| \mathbb{E}\left[y \left\langle \mu_\theta(x_{\text{obs}}, y), u^j \right\rangle\right] - \frac{1}{n}\sum_{i=1}^{n} y_i \left\langle \mu_\theta(x_{\text{obs},i}, y_i), u^j \right\rangle \right|$$

Each term in this maximum is the product of two zero-mean variables. To bound T_2, we have following observations

(i) y is sub-Gaussian with parameter at most $\sqrt{\|\theta^*\|_2^2 + \sigma^2} \leq c\sigma$;

(ii) $\langle \mu_\theta, u \rangle$ is sub-Gaussian with constant parameter by Lemma B.4.

(iii) The product of any two sub-Gaussian variables is sub-exponential.

Therefore, by standard sub-exponential tail bounds, we have

$$\mathbb{P}[T_2 \geq t] \leq 2M \exp\left(-c \min\left\{\frac{nt}{\sqrt{1+\sigma^2}}, \frac{nt^2}{1+\sigma^2}\right\}\right)$$

Since $M \leq 2^d$ and $n > c_1 d$, we have

$$T_2 \leq c\sqrt{1+\sigma^2}\sqrt{\frac{d}{n}\log(1/\delta)}$$

with probability at least $1 - \delta$.

Combining our bounds on T_1 and T_2, we have

$$\varepsilon_Q(n, \delta) \leq c\sqrt{1+\sigma^2} \times \sqrt{\frac{d}{n}\log(1/\delta)}$$

with probability at least $1 - \delta$ and the Corollary B.6 follows from Theorem 2. □

Remark B.4. If we set $T = c \log n$ for a sufficiently large constant c, then the bound (88) implies that

$$\left\|\theta^T - \theta^*\right\|_2 \leq c'\sqrt{1+\sigma^2}\sqrt{\frac{d}{n}\log^2(n/\delta)}$$

with probability at least $1 - \delta$.

C Additional Experiment

Besides looking at the optimization error for different SNR for Mixture of Gaussians, using EM, we did the following addtional experiments: First-Order-EM, for Mixture of Gaussians (Fig 2); EM, for Mixture of Regressions (Fig 3); First-Order-EM, for Mixture of Regressions (Fig 4).

All of them confirmed that the optimization error deceased faster with larger values of signal-to-noise ratio.

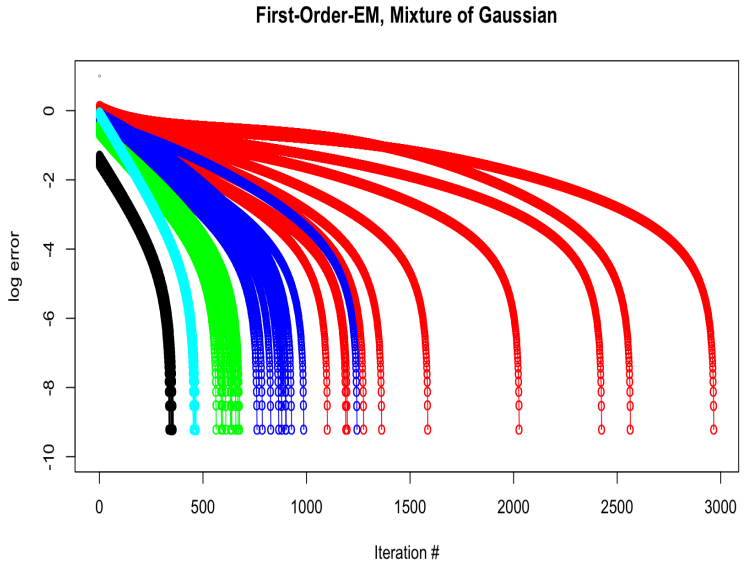


Figure 2: First-Order-EM: Mixture of Gaussians for Different SNR

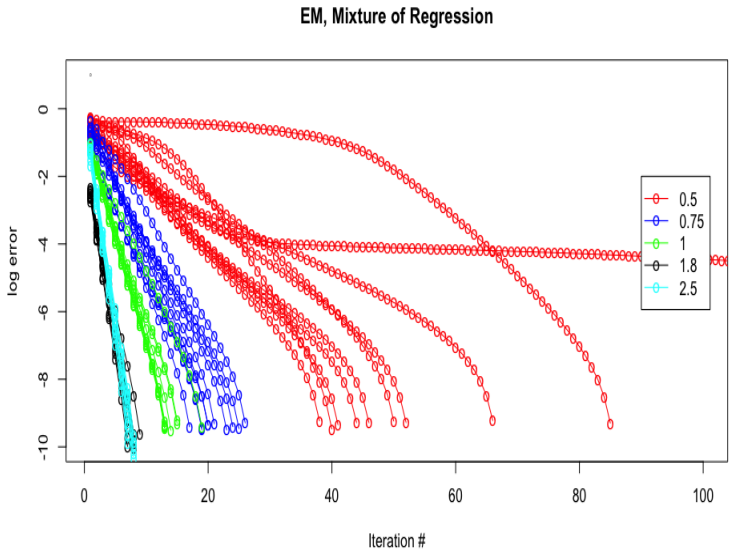


Figure 3: EM: Mixture of Regressions for Different SNR

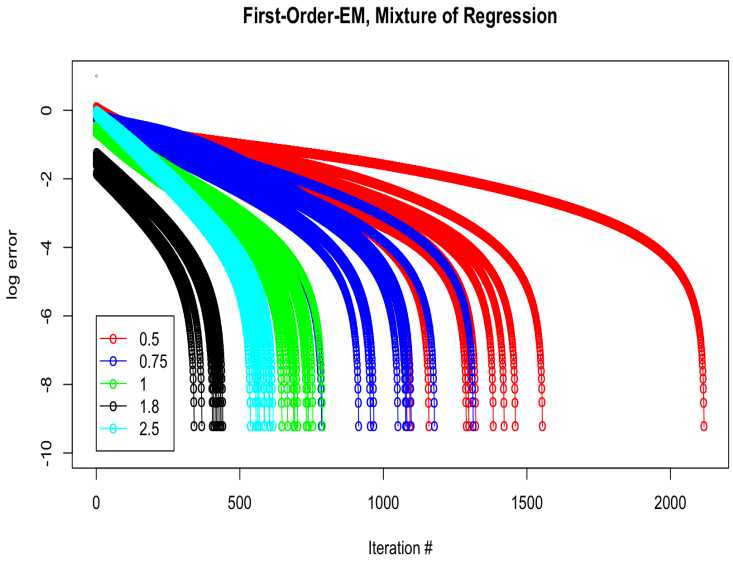


Figure 4: First-Order-EM: Mixture of Regressions for Different SNR

D R Code For Experiments

D.1 Mixture of Gaussians

```
## EM-Algorithm for GMM
#X: the data matrix, n X d. n is the number of observations, d is the dimension of
    features.
#eta: the tolerance of difference of the theta values.
EMGMM<-function(data,theta0,sigma0,alpha=0.01,eta=0.00001){
  X=data$X
  sigma20=sigma0^2
  n = nrow(X); d = ncol(X)
  theta.t = list()
  repeat({
    first.term=0
    second.term=0
    for(i in 1:n){
      xi=X[i,]
      w.n=exp(-norm(theta0-xi,"2")^2/(2*sigma20))
      w.p=exp(-norm(theta0+xi,"2")^2/(2*sigma20))
      w=w.n*(w.n+w.p)^(-1)
      first.term=first.term+w*xi
      second.term=second.term+xi
    }
    theta=2/n*first.term-1/n*second.term
    err = norm(theta - theta0,"2")
    theta.t[[length(theta.t)+1]] = theta
    theta0=theta
    #print(err)
    if(err<eta)
      break
  })
  return(list(theta.hat=theta,theta.t=theta.t))

 ## First-Order EM-Algorithm for GMM
FEMGMM<-function(data,theta0,sigma0,alpha=0.01,eta=0.0001){
  X=data$X
  sigma20=sigma0^2
  n = nrow(X); d = ncol(X)
  theta.t = list()
  repeat({
    gradient=0
    for(i in 1:n){
      xi=X[i,]
      w.n=exp(-norm(theta0-xi,"2")^2/(2*sigma20))
      w.p=exp(-norm(theta0+xi,"2")^2/(2*sigma20))
      w=w.n*(w.n+w.p)^(-1)
      gradient=gradient+(2*w-1)*xi
    }
    update=1/n*gradient-theta0
    theta=theta0+alpha*update
    err = norm(theta - theta0,"2")
    theta.t[[length(theta.t)+1]] = theta
    theta0=theta
    #print(err)
```

```
    if(err<eta)
      break
  })
  return(list(theta.hat=theta,theta.t=theta.t))
}

## GMM Data Generation
GMMdata<-function(c,sigma){
  # c decide the norm of the theta.opt
  n = 1000;k=2;d=10;sigma2 = sigma^2;
  gauss = rnorm(d)
  length = norm(gauss,'2')
  c=c
  ## construct the theta star that length is c
  theta.opt=c*gauss/length

  theta = matrix(1:(k*d),nrow=k,byrow = T)
  theta[1,] = theta.opt
  theta[2,] = -theta.opt
  pi = c(1,1)/k
  Z = sample(1:k,n,prob=pi,replace = T)
  X = matrix(0,nrow=n,ncol=d)
  for(i in 1:n){
    X[i,] = rnorm(d,mean=theta[Z[i],],sd=sqrt(sigma2))
  }
  return(list(X=X,theta=theta.opt))
}
```

D.2 Mixture of Regressions

```
##  EM for Mixture of Regressions
EMMR<-function(data,theta0,sigma0,alpha=0.01,eta=0.0001){
  X=data$X
  y=data$y
  sigma20=sigma0^2
  n = nrow(X); d = ncol(X)
  theta.t=list()
  repeat({
    first.term=0
    second.term=0
    for(i in 1:n){
      xi=X[i,]
      yi=y[i,]
      w.n=exp(-(yi-xi%*%theta0)^2/(2*sigma20))
      w.p=exp(-(yi+xi%*%theta0)^2/(2*sigma20))
      w=w.n*(w.n+w.p)^(-1)
      first.term=first.term+xi%*%t(xi)
      second.term=second.term+(2*w-1)*yi*xi
    }
    theta=solve(first.term)%*%second.term
    err = norm(theta - theta0,"2")
    theta.t[[length(theta.t)+1]] = theta
    theta0=theta
```

```
    #print(err)
    if(err<eta)
      break
  })
  return(list(theta.hat=theta,theta.t=theta.t))
}


## first-order EM for Mixture of Regressions
FEMMR<-function(data,theta0,sigma0,alpha=0.01,eta=0.0001){
  X=data$X
  y=data$y
  sigma20=sigma0^2
  n = nrow(X); d = ncol(X)
  theta.t=list()
  repeat({
    gradient=0
    for(i in 1:n){
      xi=X[i,]
      yi=y[i,]
      w.n=exp(-(yi-xi%*%theta0)^2/(2*sigma20))
      w.p=exp(-(yi+xi%*%theta0)^2/(2*sigma20))
      w=w.n/(w.n+w.p)
      gradient=gradient+(2*w-1)*yi*xi-xi%*%t(xi)%*%theta0
    }
    update=1/n*gradient
    theta=theta0+alpha*update
    err = norm(theta - theta0,"2")
    theta.t[[length(theta.t)+1]] = theta
    theta0=theta
    #print(err)
    if(err<eta)
      break
  })
  return(list(theta.hat=theta,theta.t=theta.t))
}

## Mixture of regression data
MRdata<-function(c,sigma){
  sigma2 = sigma^2
  n = 1000;k=2;d=10;sigma2 = 1;
  gauss = rnorm(d)
  length = norm(gauss,'2')
  c=c
  theta.opt=c*gauss/length
  theta = matrix(1:(k*d),nrow=k,byrow = T)
  theta[1,] = theta.opt
  theta[2,] = -theta.opt
  pi = c(1,1)/k
  Z = sample(1:k,n,prob=pi,replace = T)
  X = matrix(0,nrow=n,ncol=d)
  y = matrix(0,nrow=n, ncol=1)
  for(i in 1:n){
    xi = rnorm(d,mean=0,sd=1)
    vi = rnorm(1, mean=0,sd=1)
```

```
    X[i,] = xi
    y[i,] = xi%*%theta[Z[i],]+vi
  }
  return(list(X=X,y=y,theta=theta.opt))
}
```

D.3 Linear Regression with Missing Covariates

```
## EM for LRMC
EMLRMC<-function(data,theta0,sigma0,alpha=0.01,eta=0.0001){
  X=data$X
  y=data$y
  sigma20=sigma0^2
  n = nrow(X); d = ncol(X)
  theta.t = list()
  repeat({
    first.term=0
    second.term=0
    for(i in 1:n){
      xi=X[i,]
      yi=y[i,]
      NAindex <-which(is.na(xi))
      numNA = length(NAindex)

      if(numNA>0)
      {
        NonNAindex <- which(!is.na(xi))
        xi.obs=xi[NonNAindex]
        xi.mis=xi[NAindex]
        theta0.obs=theta0[NonNAindex]
        theta0.mis=theta0[NAindex]
        z.obs=c(xi.obs,yi)
        U.theta=cbind(-theta0.mis%*%t(theta0.obs),theta0.mis)*1
            /(norm(theta0.mis,type="2")^2+sigma20)
        u=c(U.theta%*%z.obs,xi.obs)
        p1=matrix(1, numNA, numNA)
        p2=U.theta%*%z.obs%*%t(xi.obs)
        p3=xi.obs%*%t(z.obs)%*%t(U.theta)
        p4=xi.obs%*%t(xi.obs)
        #conditional second moment matrix
        csmm=rbind(cbind(p1,p2),cbind(p3,p4))
        second.term=second.term+yi*u
        first.term=first.term+csmm
      }
      else
      {
        second.term=second.term+yi*xi #d*1 vector
        first.term=first.term+xi%*%t(xi) #d*d matrix
      }

    }
    theta=solve(first.term)%*%second.term
    err = norm(theta - theta0,"2")
```

```
    theta.t[[length(theta.t)+1]] = theta
    theta0=theta
    #print(err)
    if(err<eta)
      break
  })
  return(list(theta.hat=theta,theta.t=theta.t))
}



## First-Order-EM for LRMC

FEMLRMC<-function(data,theta0,sigma0,alpha=0.01,eta=0.0001){
  X=data$X
  y=data$y
  sigma20=sigma0^2
  n = nrow(X); d = ncol(X)
  theta.t = list()
  repeat({
    gradient=0
    for(i in 1:n){
      xi=X[i,]
      yi=y[i,]
      NAindex <-which(is.na(xi))
      numNA = length(NAindex)
      if(numNA>0)
      {
        NonNAindex <- which(!is.na(xi))
        xi.obs=xi[NonNAindex]
        xi.mis=xi[NAindex]
        theta0.obs=theta0[NonNAindex]
        theta0.mis=theta0[NAindex]
        z.obs=c(xi.obs,yi)
        U.theta=cbind(-theta0.mis%*%t(theta0.obs),theta0.mis)*1
              /(norm(theta0.mis,type="2")^2+sigma20)
        u=c(U.theta%*%z.obs,xi.obs)
        p1=matrix(1, numNA, numNA)
        p2=U.theta%*%z.obs%*%t(xi.obs)
        p3=xi.obs%*%t(z.obs)%*%t(U.theta)
        p4=xi.obs%*%t(xi.obs)
        #conditional second moment matrix
        csmm=rbind(cbind(p1,p2),cbind(p3,p4))
        gradient=gradient+yi*u-csmm%*%theta0
      }
      else
      {
        gradient=gradient+yi*xi-xi%*%t(xi)%*%theta0
      }

    }
    update=1/n*gradient
    theta=theta0+alpha*update
    err = norm(theta - theta0,"2")
    theta.t[[length(theta.t)+1]] = theta
```

```
    theta0=theta
    #print(err)
    if(err<eta)
      break
  })
  return(list(theta.hat=theta,theta.t=theta.t))
}

## Linear Regression with Missing Covariates data generation

LRMCdata<-function(c,sigma){
  n = 1000;k=2;d=10;sigma2 =sigma^2;p = 0.2
  y = matrix(0,nrow=n, ncol=1)
  gauss = rnorm(d)
  length = norm(gauss,'2')
  c=c
  theta.opt=c*gauss/length
  #print(norm(theta.opt,'2'))
  theta = matrix(1:(k*d),nrow=k,byrow = T)
  theta[1,] = theta.opt
  theta[2,] = -theta.opt
  pi = c(1,1)/k
  Z = sample(1:k,n,prob=pi,replace = T)
  X = matrix(0,nrow=n,ncol=d)
  y = matrix(0,nrow=n, ncol=1)
  for(i in 1:n){
    xi = rnorm(d,mean=0,sd=1)
    mask = sample(c(NA,1), prob=c(p, 1-p), replace=TRUE, size=d)
    X[i,] = xi*mask
    vi = rnorm(1, mean=0,sd=1)
    y[i,] = sum(X[i,]*theta[Z[i],], na.rm = TRUE) +vi
  }
  return(list(X=X,y=y,theta=theta.opt))
}
```

D.4 Experiments

```
## function to run 10 times for a given theta.opt and sigma and given algorithm
r1 = function(c,sigma,datafunc,algofunc){
  data = datafunc(c,sigma)
  theta.opt = data$theta
  theta0=c()
  ## based on Fig3, the theta.opt_i - sqrt(c/10)<=theta0_i<=theta.opt_i + sqrt(c/10)
  dif = runif(1,c*sqrt(1/10)-0.5,c*sqrt(1/10)-0.5)
  for(i in 1:length(theta.opt)){
    p = sample(c(1,-1),1)
    value = theta.opt[i] + p*dif
    theta0 = append(theta0, value)
  }
  norm(theta.opt,'2')
  norm(theta0,'2')
  norm((theta.opt-theta0),'2')
  # call the function
```

```
  mod=algofunc(data,theta0,sigma)
  theta.t = mod$theta.t
  theta.hat = mod$theta.hat
  end = length(theta.t)

  print(end)
  err.opt = c()
  err.stat = c()
  for(i in 1:end){
    theta = theta.t[[i]]
    e1=log(norm(theta.hat-theta,'2'))
    e2=log(norm(theta.opt-theta,'2'))
    #print(e1)
    #print(e2)
    err.opt=c(err.opt, e1)
    err.stat=c(err.stat, e2)
  }
  return(list(err.opt=err.opt,err.stat=err.stat))
}

## Fig5(a)
rs = lapply(rep(2,10), function(c){
  sigma = 1
  out=r1(c,sigma,GMMdata,EMGMM)
})

end = 0
for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  end_ = length(err.opt)
  end = max(end,end_)
}

x = seq(1:end-1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-8,1),
     main='EM, Mixture of Gaussian',cex=1,lwd=1)

for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  err.stat=rs[[i]]$err.stat
  end = length(err.opt)-1
  x = seq(1:end)
  points(x, err.opt[1:end], col="blue", pch="o",cex=1,lwd=1)
  lines(x, err.opt[1:end], col="blue", lty=1, cex=1,lwd=1)
  points(x, err.stat[1:end], col="red", pch="*",cex=1,lwd=1)
  lines(x, err.stat[1:end], col="red", lty=2,cex=1,lwd=1)
}
legend(1,-6,legend=c("Opt. error","Stat. error"), col=c("blue","red"),
       pch=c("o","*"),lty=c(1,2), ncol=1)


## Fig 5(b)
rs = lapply(rep(2,10), function(c){
  sigma = 1
  out=r1(c,sigma,GMMdata,FEMGMM)
```

```
})

end = 0
for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  end_ = length(err.opt)
  end = max(end,end_)
}

end=500
x = seq(1:end-1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-6,1),
     main='First-order EM, Mixture of Gaussian',cex=0.3,lwd=0.3)

for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  err.stat=rs[[i]]$err.stat
  end = length(err.opt)-1
  end=500
  x = seq(1:end)
  points(x, err.opt[1:end], col="blue", pch="o",cex=0.3,lwd=0.3)
  lines(x, err.opt[1:end], col="blue", lty=1, cex=0.3,lwd=0.3)
  points(x, err.stat[1:end], col="red", pch="*",cex=0.3,lwd=0.3)
  lines(x, err.stat[1:end], col="red", lty=2,cex=0.3,lwd=0.3)
}
legend(1,-4,legend=c("Opt. error","Stat. error"), col=c("blue","red"),
       pch=c("o","*"),lty=c(1,2), ncol=1)


## Fig6
end = 0
rss = list()
snrs = c(0.5,0.75,1,1.8,2.5)
repnum=10
for(i in snrs){
  rs = lapply(rep(i,repnum), function(i){
    sigma = 1
    out=r1(i,sigma,GMMdata,EMGMM)
  })
  for(j in 1:repnum){
    err.opt=rs[[j]]$err.opt
    end_ = length(err.opt)
    end = max(end,end_)
  }
  rss[[length(rss)+1]]=rs
}


x = seq(1:end+1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-10,1),
     main='EM, Mixture of Gaussian',cex=0.3,lwd=0.3)

colors=c('red','blue','green','black','cyan')
for(j in 1:length(snrs)){
  rs=rss[[j]]
```

```
  color=colors[j]
  for(i in 1:repnum){
    err.opt=rs[[i]]$err.opt
    end = length(err.opt)-1
    x = seq(1:end)
    points(x, err.opt[1:end], col=color, pch="o",cex=1,lwd=1)
    lines(x, err.opt[1:end], col=color, lty=1, cex=1,lwd=1)
  }
}
legend(70,-2,legend=snrs, col=colors,
       pch=rep("o",5),lty=rep(1,5), ncol=1)


## Fig7(a)
## based on the theorem, the theta.opt_i - sqrt(c/10)<=theta0_i<=theta.opt_i + sqrt(c/10)
rs = lapply(rep(2,10), function(c){
  sigma = 1
  out=r1(c,sigma,MRdata,EMMR)
})

end = 0
for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  end_ = length(err.opt)
  end = max(end,end_)
}


x = seq(1:end-1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-8,1),
     main='EM, Mixture of Regression',cex=1,lwd=1)

for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  err.stat=rs[[i]]$err.stat
  end = length(err.opt)-1
  x = seq(1:end)
  points(x, err.opt[1:end], col="blue", pch="o",cex=1,lwd=1)
  lines(x, err.opt[1:end], col="blue", lty=1, cex=1,lwd=1)
  points(x, err.stat[1:end], col="red", pch="*",cex=1,lwd=1)
  lines(x, err.stat[1:end], col="red", lty=2,cex=1,lwd=1)
}
legend(1,-6,legend=c("Opt. error","Stat. error"), col=c("blue","red"),
       pch=c("o","*"),lty=c(1,2), ncol=1)


## Fig 7(b)
rs = lapply(rep(2,10), function(c){
  sigma = 1
  out=r1(c,sigma,MRdata,FEMMR)
})

end = 0
for(i in 1:10){
  err.opt=rs[[i]]$err.opt
```

```
  end_ = length(err.opt)
  end = max(end,end_)
}

end=400
x = seq(1:end-1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-8,1),
     main='First-order EM, Mixture of Regression',cex=0.3,lwd=0.3)

for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  err.stat=rs[[i]]$err.stat
  end = length(err.opt)-1
  end=400
  x = seq(1:end)
  points(x, err.opt[1:end], col="blue", pch="o",cex=0.3,lwd=0.3)
  lines(x, err.opt[1:end], col="blue", lty=1, cex=0.3,lwd=0.3)
  points(x, err.stat[1:end], col="red", pch="*",cex=0.3,lwd=0.3)
  lines(x, err.stat[1:end], col="red", lty=2,cex=0.3,lwd=0.3)
}
legend(1,-6,legend=c("Opt. error","Stat. error"), col=c("blue","red"),
       pch=c("o","*"),lty=c(1,2), ncol=1)



## fig6 suplement for regression
end = 0
rss = list()
snrs = c(0.5,0.75,1,1.8,2.5)
repnum=10
for(i in snrs){
  rs = lapply(rep(i,repnum), function(i){
    sigma = 1
    out=r1(i,sigma,MRdata,EMMR)
  })
  for(j in 1:repnum){
    err.opt=rs[[j]]$err.opt
    end_ = length(err.opt)
    end = max(end,end_)
  }
  rss[[length(rss)+1]]=rs
}

end=100
x = seq(1:end+1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-10,1),
     main='EM, Mixture of Regression',cex=0.3,lwd=0.3)

colors=c('red','blue','green','black','cyan')
for(j in 1:length(snrs)){
  rs=rss[[j]]
  color=colors[j]
  for(i in 1:repnum){
    err.opt=rs[[i]]$err.opt
```

```
    end=100
    end = length(err.opt)-1
    x = seq(1:end)
    points(x, err.opt[1:end], col=color, pch="o",cex=1,lwd=1)
    lines(x, err.opt[1:end], col=color, lty=1, cex=1,lwd=1)
  }
}
legend(90,-2,legend=snrs, col=colors,
       pch=rep("o",5),lty=rep(1,5), ncol=1)


## Fig8(a)
rs = lapply(rep(2,10), function(c){
  sigma = 1
  out=r1(c,sigma,LRMCdata,EMLRMC)
})

end = 0
for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  end_ = length(err.opt)
  end = max(end,end_)
}

end=10
x = seq(1:end-1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-7,1),
     main='EM, Missing Data Regression')

for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  err.stat=rs[[i]]$err.stat
  end = length(err.opt)-1
  end=10
  x = seq(1:end)
  points(x, err.opt[1:end], col="blue", pch="o")
  lines(x, err.opt[1:end], col="blue", lty=1)
  points(x, err.stat[1:end], col="red", pch="*")
  lines(x, err.stat[1:end], col="red", lty=2)
}
legend(1,-5,legend=c("Opt. error","Stat. error"), col=c("blue","red"),
       pch=c("o","*"),lty=c(1,2), ncol=1)


## Fig8(b)
rs = lapply(rep(2,10), function(c){
  sigma = 1
  out=r1(c,sigma,LRMCdata,FEMLRMC)
})

end = 0
for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  end_ = length(err.opt)
  end = max(end,end_)
```

```
}

end=500
x = seq(1:end+1)
plot(x,xlab="Iteration #", ylab="log error",ylim=c(-7.5,1),
     main='First-order EM, Missing Data Regression',cex=0.3,lwd=0.3)

for(i in 1:10){
  err.opt=rs[[i]]$err.opt
  err.stat=rs[[i]]$err.stat
  end = length(err.opt)-1
  end =500
  x = seq(1:end)
  points(x, err.opt[1:end], col="blue", pch="o",cex=0.3,lwd=0.3)
  lines(x, err.opt[1:end], col="blue", lty=1, cex=0.3,lwd=0.3)
  points(x, err.stat[1:end], col="red", pch="*",cex=0.3,lwd=0.3)
  lines(x, err.stat[1:end], col="red", lty=2,cex=0.3,lwd=0.3)
}
legend(1,-5,legend=c("Opt. error","Stat. error"), col=c("blue","red"),
       pch=c("o","*"),lty=c(1,2), ncol=1)
```