EM Converges for a Mixture of Many Linear Regressions

Jeongyeol Kwon
The University of Texas at Austin

Constantine Caramanis
The University of Texas at Austin

Abstract

We study the convergence of the Expectation-Maximization (EM) algorithm for mixtures of linear regressions with an arbitrary number k of components. We show that as long as signal-to-noise ratio (SNR) is $\tilde{\Omega}(k)$, well-initialized EM converges to the true regression parameters. Previous results for $k \geq 3$ have only established local convergence for the noiseless setting, i.e., where SNR is infinitely large. Our results enlarge the scope to the environment with noises, and notably, we establish a statistical error rate that is independent of the norm (or pairwise distance) of the regression parameters. In particular, our results imply exact recovery as $\sigma \to 0$, in contrast to most previous local convergence results for EM, where the statistical error scaled with the norm of parameters. Standard moment-method approaches may be applied to guarantee we are in the region where our local convergence guarantees apply.

1 Introduction

The Expectation-Maximization (EM) algorithm is a powerful tool for statistical inference when we have samples with missing information, often modeled as latent variables. It is a general-purpose heuristic for evaluating the maximum likelihood (ML) estimator for such problems Wu et al. (1983). A canonical example is parameter estimation for the mixture of a known family of parameterized distributions such as Gaussian Mixture Models (GMM) or Mixture of Linear Regressions (MLR). In such problems, solving for maximum likelihood estimator is NP-hard due to the non-convexity of the log-likelihood function. The EM algorithm successively computes tighter lower bounds on the likelihood function; each iteration is no more complex than solving the ML problem with no missing data. Despite its simplicity and broad success in practice, a theoretical understanding of EM remains largely elusive (but see Section 1.1 for important recent results). In general, the EM algorithm may fail to converge to a global optimum of log-likelihood function. Thus, its success story is specific to problems to which the EM algorithm is applied.

In this paper, we study the convergence behavior of the EM algorithm for mixture of linear regressions with k components. We show that the EM algorithm converges to the true parameters when the signal-to-noise ratio (SNR) is larger than $\tilde{\Omega}(k)$, and the parameter is well initialized, within $\tilde{O}(1/k)$ of the true parameters (see related works and Remark 1 for some known initialization techniques). This is the first result, to the best of our knowledge, to establish the convergence of the EM algorithm in MLR with more than two components and finite SNR. Furthermore, under the same regularity conditions, we recover the results of Kwon et al. (2019) for two-component mixtures, showing that the statistical error of the sample-splitting finite sample EM algorithm is $\tilde{O}(\sigma\sqrt{k^2d/n})$ where n is the number of samples per iteration. This is significant because our analysis then implies exact recovery in the noiseless setting even with finite number of samples, in contrast to earlier work Balakrishnan et al. (2017); Klusowski et al. (2019) that only showed statistical error scales with the norm (or pairwise distance) of the regression parameters.

1.1 Related Work

Work in Balakrishnan et al. (2017) established a characterization of the local region of attraction within which EM is guaranteed to converge to a point with the statistical precision of a global optimum. This complemented work in Yi et al. (2014) that gave an analogous result for noise-less mixed regression. A key aspect in Balakrishnan et al. (2017) involves coupling an analysis of population EM to finite sample EM. Several results have followed, providing conver-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

gence results for canonical problems such as GMM or MLR. In the special case of two balanced mixtures, global convergence results have been established in Jin et al. (2016); Daskalakis et al. (2017) for GMMs, and in Kwon et al. (2019) for MLR. Beyond more than two components, a negative result for global convergence of the EM algorithm for 3-GMM has been established Jin et al. (2016), while Zhao et al. (2018); Yan et al. (2017) give a local convergence result for k-GMM with arbitrary $k \geq 3$. Attempts have been made to obtain analogous results for mixed linear regression. However, these efforts have only been successful in the setting of infinite SNR, i.e., the noiseless setting. Here, Yi et al. (2016) establishes convergence of alternating minimization, while Zhong et al. (2016) obtains a similar result by solving a non-convex formulation; work in Hand and Joshi (2018) gives a convex objective that solves the noiseless MLR problem for well-separated data.

Indeed, the problem of solving mixture of linear regressions has been extensively studied. In general, MLR is NP-hard Yi et al. (2014) due to the combinatorial nature of the problem. Therefore, it is natural to consider assumptions in the problem, and various efficient algorithms have been proposed under certain statistical assumptions Sedghi et al. (2014); Chaganty and Liang (2013); Yi et al. (2014); Chen et al. (2014); Zhong et al. (2016); Yi et al. (2016); Chen et al. (2017); Li and Liang (2018); Hand and Joshi (2018). For instance, Chen et al. (2014) proposed convex formulation which achieves the optimal minimax rate for equal-weighted 2-MLR, and later in Chen et al. (2017) extended the treatment to unequally weighted mixtures, but again focus on the mixture of only two components. As mentioned, Yi et al. (2016); Zhong et al. (2016); Hand and Joshi (2018) all propose algorithms for solving k-MLR in the noiseless setting.

A common technical tool used by many algorithms is the powerful method of moments. In the various algorithms based on method of moments Sedghi et al. (2014); Li and Liang (2018); Yi et al. (2014); Chaganty and Liang (2013); Zhong et al. (2016); Yi et al. (2016), up to third-order tensors are constructed from Gaussian regression models and tensor-decomposition algorithms are performed. The drawback of a purely moment-based method is the high sample and computational complexity. In particular, the statistical error of the resulting estimator typically scales with the norm of the regression parameters. Therefore, these methods are often used in conjunction with fast iterative algorithms, such as gradient descent Zhong et al. (2016); Li and Liang (2018) or alternating minimization Yi et al. (2014, 2016). While the work cited provides guarantees for these iterative algorithms in the noiseless setting, they are no longer consistent estimators in the presence of noise. In practice, the EM algorithm seems to obtain better results; in theory, however, the question of whether EM always converges to the global optimum for k-MLR with $k \geq 3$ is open, even when initialized in a neighborhood of true parameters. This paper provides an affirmative answer to this question.

1.2 Main Contribution

We prove local convergence of the EM algorithm for k-MLR, showing that it converges to a global optimum with high probability, when SNR is $\tilde{\Omega}(k)$, and EM is initialized in a $1/\tilde{O}(k)$-neighborhood of a global optimum. We first establish this result in the infinite sample limit, i.e., population EM. Our result generalizes the results in Balakrishnan et al. (2017) which established local convergence for a symmetrized balanced mixture of two components. We establish local convergence in the setting with arbitrary number of components and possibly unbalanced mixing weights. At a high level, our analysis proceeds by carefully constructing the event where the samples are almost correctly assigned their weights, bringing the next estimator closer to the true parameter. Given good initialization and high enough SNR, we expect most samples fall into this category. At the same time, we bound the portion of bad samples which do not fall into this event. The effect of this "leakage" is thus canceled out when the average is taken over all samples. By this construction, our convergence rate is no longer dependent on the maximum distance between regression parameters which has often appeared in the EM literature as an artifact of the analysis.

We then show the convergence of a simple variant of finite-sample EM[1] via concentration arguments. Toward this goal, we propose "event-wise" concentration of random variables as a proof strategy. Intuitively speaking, the samples that fall into the good event in population EM only induce exponentially small errors. Consequently, statistical errors from these good samples should also be exponentially small. Furthermore, they are the majority among all samples under our assumption on SNR and initialization. On the other hand, samples conditioned on bad events could incur an error as large as the norm of the parameters. However, they are in the minority, and large norms will be canceled out when divided by the total number of samples. See Section 5 and Proposition 5.3 for a detailed discussion and formal statement. Remarkably, we show that the statistical error only scales with the variance of the noise.

[1]The variant is often called *sample-splitting* since it divides entire samples into T batches and uses a new batch in every iteration.

2 Problem Setup

We consider the mixture of multiple linear regressions, where a pair of random variables $(X, y) \in \mathbb{R}^d \times \mathbb{R}$ are generated from one of k linear models:

$$\mathcal{D}_j : y = \langle X, \beta_j^* \rangle + e, \quad \text{for} \quad j = 1, ..., k$$

where e represents additive noise in the measurement with variance σ^2. Our goal is recovering regression parameters $\{\beta_j^*\}_{j=1}^k$ when the labels that indicate from which domain each pair is generated are missing. Thus, we are considering the estimation of parameters for the mixture of distributions $\{\mathcal{D}_j\}_{j=1}^k$ with mixing weights $\{\pi_j^*\}_{j=1}^k$. In the finite sample regime, we estimate $\{\beta_j^*\}_{j=1}^k$ when we have n samples $(X_i, y_i)_{i=1}^n \sim \mathcal{D}$, where $\mathcal{D} = \sum_j \pi_j^* \mathcal{D}_j$ is a mixture distribution.

In this paper, we assume that the design vector X for all linear components comes from a shared standard multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. We assume e is a zero-mean and unit-variance Gaussian random variable and independent of X. Thus, the problem is rescaled with known variance parameter σ^2.

Notation. We use d to denote the dimension of the problem and k the number of components. (X, Y) are a pair of random variables from mixture distribution $\mathcal{D}$, n is the number of samples, and (X_i, y_i) are generated samples. We define pairwise distance R_{ij}^*, and R_{min}, R_{max} as the smallest and largest distance between regression vectors of any pair of linear models:

$$R_{ij}^* = \|\beta_i^* - \beta_j^*\|, R_{min} = \min_{i \neq j} R_{ij}^*, \ R_{max} = \max_{i \neq j} R_{ij}^*.$$

We define SNR of this problem as R_{min}, which is equivalent to the ratio of minimum pairwise distance versus variance of noise. Define $\rho_\pi = \max_j(\pi_j^*)/\min_j(\pi_j^*)$ as the ratio of maximum mixing weight and minimum mixing weight, and $\pi_{min} = \min_j \pi_j^*$.

We denote the max of two scalar quantities a, b as $(a \vee b)$. When v is a vector, $\|v\|$ is l_2 norm of v. Inner product of two vectors u, v is denoted as $\langle u, v \rangle$. When A is positive semi-definite (PSD) matrix, $\|A\|_{op} = \sup_{s \in \mathbb{S}^{d-1}}(s^T A s)$ is an operator norm of A, where $\mathbb{S}^{d-1}$ represents the unit sphere in $\mathbb{R}^d$ space and s is any unit vector in $\mathbb{R}^d$.

We use $\mathbb{E}_P[X]$ to denote the expectation of random variable $X \sim P$. Thus $\mathbb{E}_{\mathcal{D}}[\cdot]$ is the expectation taken over the mixture distribution $\mathcal{D}$, and $\mathbb{E}_{\mathcal{D}_j}[\cdot]$ is the expectation taken over distribution corresponds to j^{th} linear model. We denote $\mathbb{1}_{X \in \mathcal{E}}$ an indicator function for event $\mathcal{E}$, and often use a shorthand for it $\mathbb{1}_{\mathcal{E}}$ when the context is clear. We use $\mathbb{E}[X|\mathcal{E}]$ to denote conditional expectation under event $X \in \mathcal{E}$.

For one step analysis of population EM iteration, we use β_j to denote the current estimator of j^{th} parameter, and β_j^+ to denote the next estimator resulted from EM operator. We denote $\Delta_j := \beta_j - \beta_j^*$. We denote $\tilde{\beta}_j$ and $\tilde{\beta}_j^+$ be corresponding estimators for the finite-sample EM. In the result for entire EM algorithm, $\beta_j^{(t)}$ and $\tilde{\beta}_j^{(t)}$ denote the estimator in the t^{th} step of population EM and finite-sample EM respectively. Notations for mixing weights π_i are defined in a similar manner.

3 Main Results

We state the main results for both population EM and finite-sample EM. We provide a proof sketch in the following two sections, and defer details to the Appendix.

One iteration of the population EM algorithm for this problem consists of two steps:

$$\begin{aligned}
&\text{(E-step)}: \quad w_j = \frac{\pi_j \exp(-(Y - \langle X, \beta_i \rangle)^2/2)}{\sum_{l \in [k]} \pi_l \exp(-(Y - \langle X, \beta_l \rangle)^2/2)} \\
&\text{(M-step)}: \quad \beta_j^+ = (\mathbb{E}_{\mathcal{D}}[w_j X X^\top])^{-1}(\mathbb{E}_{\mathcal{D}}[w_j X Y]), \\
&\qquad\qquad\ \ \pi_j^+ = \mathbb{E}_{\mathcal{D}}[w_j].
\end{aligned}$$

We first state our main convergence result for population EM after T iterations.

Theorem 3.1 *There exists universal constant $C, c > 0$ such that if $R_{min} \geq Ck\rho_\pi \log^2(k\rho_\pi)$, $|\pi_j^{(0)} - \pi_j^*| \leq \pi_j^*/2$, and $\|\beta_j^* - \beta_j^{(0)}\| \leq cR_{min}/(k\rho_\pi \log(k))$ for all j, then EM converges to true parameters after $T = O(\log(\max_j \|\beta_j^* - \beta_j^{(0)}\|/\epsilon))$ steps, i.e., $\max_j \|\beta_j^* - \beta_j^{(T)}\| \leq O(\epsilon)$ for all j.*

Remark 1 (Initialization with Tensor Methods) *Tensor-based methods are able to recover all parameters under the condition that the regression parameters are linearly independent*[2]*. However, tensor methods either have a poor dependence on d Chaganty and Liang (2013), or a sub-optimal sample-complexity (polynomial) dependence on R_{max}, in order to get the precision error independent of R_{max}. This is the case for a natural extension of the tensor-based method of Yi et al. (2016). Thus it is common procedure to use spectral methods to get a crude but good enough initialization, and then continue with EM when the noise is small.*

Next, we state our main results for finite-sample EM. In finite-sample EM with sample-splitting strategy, we divide n samples into T batches, and uses a fresh batch

[2]Without this structural assumption, there is no known polynomial-time algorithm for MLR, hence initialization becomes another challenging open problem.

of n/T samples per every iteration. To simplify notation, simply use n rather than n/T when it is clear from context that we are focusing on the single stage analysis. The update rule for each step is:

$$\text{(E-step)}: \; w_{i,j} = \frac{\tilde{\pi}_j \exp(-(Y - \langle X, \tilde{\beta}_j\rangle)^2/2)}{\sum_{l\in[k]} \tilde{\pi}_l \exp(-(Y - \langle X, \tilde{\beta}_l\rangle)^2/2)}$$

$$\text{(M-step)}: \; \tilde{\beta}_j^+ = \Big(\sum_{i\in[n]} w_{i,j} X_i X_i^\top\Big)^{-1}\Big(\sum_{i\in[n]} w_{i,j} X_i y_i\Big),$$

$$\tilde{\pi}_j^+ = \frac{1}{n}\sum_{i\in[n]} w_{i,j}.$$

We show similarly the convergence result for finite-sample EM after T iterations:

Theorem 3.2 *There exists universal constants $C, c > 0$ such that if $R_{min} \geq Ck\rho_\pi \log^2(k\rho_\pi)$, $|\tilde{\pi}_j^{(0)} - \pi_j^*| \leq \pi_j^*/2$, and $\|\beta_j^* - \tilde{\beta}_j^{(0)}\| \leq cR_{min}/(k\rho_\pi \log(k))$ for all j, then given n i.i.d. samples (X_i, y_i) from mixture distribution $\mathcal{D}$, where the sample complexity is $n/T = \tilde{O}((k/\pi_{min})(d/\epsilon^2))$, then with high probability, sample-splitting finite-sample EM converges to true parameters, i.e., $\|\beta_j^* - \tilde{\beta}_j^{(T)}\| \leq O(\epsilon)$ for all j after $T = O(\log(\max_j \|\beta_j^* - \tilde{\beta}_j^{(0)}\|/\epsilon))$ iterations.*

Remark 2 *The statistical error in our result is independent of R_{min} or R_{max}. This implies in the original problem where the variance of noise is σ, we have statistical precision $O(\sigma\epsilon)$. It guarantees exact recovery as $\sigma \to 0$. This is the first result showing that the statistical error rate of the EM algorithm does not depend on the distance between any two regression vectors in noisy environment, as opposed to all previous analysis on EM Balakrishnan et al. (2017); Klusowski et al. (2019); Zhao et al. (2018); Yan et al. (2017); Yi and Caramanis (2015). We provide the detailed discussion on this issue in Section 5.*

Discussion of Main Results. Several points are in order before we move on to the technical proofs. First, note that in the balanced setting (where all mixing weights are nearly equal to $1/k$), $\rho_\pi = 1$ and $\pi_{min} = 1/k$, thus the SNR condition is $\Omega(k)$, and sample complexity per each iteration is $\tilde{O}(k^2 d/\epsilon^2)$. The dependency on d and ϵ is thus optimal. We note that the $O(k^2)$ dependency appeared even in the noiseless setting Yi et al. (2016). The total number of iterations is $T = O(\log(\max_j \|\beta_j^* - \beta_j^{(0)}\|/\epsilon))$, as a result of linear convergence with constant rate in our analysis. In the original scale with noise variance σ^2, it is equivalent to $T = O(\log(\max_j \|\beta_j^* - \beta_j^{(0)}\|/\epsilon'))$, to achieve an error of $O(\epsilon')$ where $\epsilon' = \sigma\epsilon$. Note that in the extreme setting where $\sigma \to 0$, exact recovery can still be guaranteed in a finite number of steps, but it requires separate case study on the last iteration which we omit in this paper (see Lemma 3 and Corollary 1 in Yi et al. (2016)).

A natural question is whether the $\Omega(k)$ requirement for SNR $\Omega(k)$ is sharp. In a very closely related problem, the parameter estimation of GMM, Regev and Vijayaraghavan (2017) established the lower bound for minimum separation between the centers of each Gaussian component to recover all centers using a polynomial number of samples. Indeed, the bound $\Omega(\sqrt{\log k})$ established in Regev and Vijayaraghavan (2017) is a threshold above which the labels of most samples can be correctly identified (thus the majority are good samples) if the ground truth parameters are given. However, for mixed linear regression, no such lower bound result has been established. We conjecture that such lower bound might be much larger in mixtures of linear regressions, and it might be closely related to the convergence of the EM algorithm. We leave it as a main future challenge to find such lower bounds for mixed linear regression. In this paper, we focus on the local analysis of the EM algorithm under the condition where the labels of most samples can be correctly identified, if we have good estimate of ground truth parameters. We note that it might be possible to improve the logarithmic factors on the SNR condition with more refined analysis.

Doing away with sample splitting in our algorithm is also a natural and important extension, since we use it in the analysis, though it is well appreciated that in practice EM does not appear to need it. One way to avoid the sample-splitting technique is to get an uniform concentration bound over local region of interest. Indeed, some previous works on the EM algorithm does precisely this, obtaining uniform concentration of EM operators Yan et al. (2017); Zhao et al. (2018); Cai et al. (2019). However, their statistical errors have polynomial dependence on R_{max}. It is not clear how to remove this in their analysis, even if we do allow sample-splitting. We take an alternate analysis path; while we cannot seem to avoid sample splitting, we do succeed in removing this R_{max} dependence (see Section 5). We thus obtain an error rate that is free of distance between any two regression vectors. We leave it as a future work to derive the uniform concentration type result with the same statistical error rate.

4 Analysis of Population EM

We first give the sketch of the proof for population EM and provide detailed proof in Appendix A. For the ease of the presentation, we assume we know the true weights and use them in the main text, and tackle the general setting in the Appendix. We express $\beta_1^+ - \beta_1^*$

as

$$\beta_1^+ - \beta_1^* = (\mathbb{E}_{\mathcal{D}}[w_1 XX^\top])^{-1}(\mathbb{E}_{\mathcal{D}}[w_1 X(Y - \langle X, \beta_1^*\rangle)]).$$

Then, we exploit the fact that true parameters are a fixed point of the EM iteration. That is,

$$w_1^* = \frac{\pi_1 \exp(-(Y - \langle X, \beta_1^*\rangle)^2/2)}{\sum_j \pi_j \exp(-(Y - \langle X, \beta_j^*\rangle)^2/2)},$$
$$\mathbb{E}_{\mathcal{D}}[w_1^* X(Y - \langle X, \beta_1^*\rangle)] = \pi_1^* \mathbb{E}_{\mathcal{D}_1}[X(Y - \langle X, \beta_1^*\rangle)] = 0.$$

Then $\beta_1^+ - \beta_1^*$ can be re-written as

$$\beta_1^+ - \beta_1^* = (\mathbb{E}_{\mathcal{D}}[w_1 XX^\top])^{-1}(\mathbb{E}_{\mathcal{D}}[\Delta_w X(Y - \langle X, \beta_1^*\rangle)]),$$

where we defined $\Delta_w := w_1 - w_1^*$. We then bound two terms minimum singular value of $A = \mathbb{E}_{\mathcal{D}}[w_1 XX^\top]$ and norm of $B = \mathbb{E}_{\mathcal{D}}[\Delta_w X(Y - \langle X, \beta_1^*\rangle)]$.

Remark 3 *We do not exactly verify the so-called gradient smoothness (GS)-condition for population EM operator as proposed in Balakrishnan et al. (2017). The reason for that becomes clear in the finite-sample analysis: the inverse of $\mathbb{E}_{\mathcal{D}}[w_1 XX^\top]$ does not match that of finite sample EM, which has inverse of $1/n \sum_i w_{1,i} X_i X_i^\top$. If we try to control the deviation of the entire finite-sample EM operator from the population EM operator, this mismatch inevitably results in a statistical error that scales with R_{max}.*

4.1 Bounding B

We first bound B. The high-level idea of bounding B is closely related to Balakrishnan et al. (2017). Our result formalizes the proof idea in Balakrishnan et al. (2017) to make it applicable to k-mixture of regressions. Define $D_m := \max_j \|\beta_j - \beta_j^*\|$. We first treat the case when $D_m > 1$, and apply mean-value theorem for $D_m \le 1$. In either case, we construct good events with carefully chosen parameters to bound the portion of bad samples and errors induced by good samples. We start with stating our lemma on the bound of B.

Lemma 4.1 *Under the condition in Theorem 3.1, when $D_m > 1$, there exists universal constants $c_1, c_1' \in (0, 1/8)$ and $c_2, c_3, c_4 > 0$ such that:*

$$\|B\| \le \pi_1^*(c_1 + c_2 k \log(k\rho_\pi)/R_{min}) + \sum_{j\neq 1} \pi_j^* \Big(c_1'/(k\rho_\pi) + c_3 \log(R_{j1}^* k\rho_\pi)/R_{j1}^* + \left(c_4 D_m / R_{j1}^*\right) D_m \Big).$$

When $D_m \le 1$, there exists another universal constants $c_1, c_1' \in (0, 1/8)$ and $c_2, c_3 > 0$ such that:

$$\|B\| \le \pi_1^*(c_1 + c_2 k \log(k\rho_\pi)/R_{min}) D_m + \sum_{j\neq 1} \pi_j^* \Big(c_1'/(k\rho_\pi) + c_3 \log^2(R_{j1}^* k\rho_\pi)/R_{j1}^* \Big) D_m.$$

Note that when $D_m \le \tilde{O}(R_{min}/k\rho_\pi)$ and $R_{min} = \tilde{\Omega}(k\rho_\pi)$, above equations guarantee that $\|B\| \le c_B \pi_1^* D_m$ for some small constant c_B. We only present how to bound errors from other components $j \neq 1$ in the main text, but the same idea is applied to all other cases. In defining what are the good samples, we need to consider two things: (i) the noise is not abnormally large, (ii) residual error due to class mismatch is large enough to be detected. It can be formalized into the following on three events in j^{th} component for $j \neq 1$:

$$\begin{aligned}
&\mathcal{E}_{j,1} = \{|e| \le \tau_j\}, \\
&\mathcal{E}_{j,2} = \{|\langle X, \Delta_1\rangle| \vee |\langle X, \Delta_j\rangle| \le |\langle X, \beta_j^* - \beta_1^*\rangle|/4\}, \\
&\mathcal{E}_{j,3} = \{|\langle X, \beta_j^* - \beta_1^*\rangle| \ge 4\sqrt{2}\tau_j\}, \\
&\mathcal{E}_{j,good} = \mathcal{E}_{j,1} \cap \mathcal{E}_{j,2} \cap \mathcal{E}_{j,3}. \qquad (1)
\end{aligned}$$

Here, τ_j is a threshold parameter that we specify carefully in the proof. When these three events occur at the same time, it is a good sample: weights given to first component for this sample is almost 0. In fact, we can show that $|\Delta_w| \le (\pi_1^*/\pi_j^*)\exp(-\tau_j^2)$. The errors from this event can be thus bounded by

$$\|\mathbb{E}_{\mathcal{D}_j}[\Delta_w X(Y - \langle X, \beta_1^*\rangle)\mathbb{1}_{\mathcal{E}_{good}}]\| \le (\pi_1^*/\pi_j^*)\exp(-\tau_j^2) \sup_{s\in\mathbb{S}^{d-1}} \mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)|],$$

where we omitted subscript j in the notation for events. As the supremum is shown to be in order $O(R_{j1}^*)$, the choice of $\tau_j = \Theta\left(\sqrt{\log(R_{j1}^* k)}\right)$ here comes clear if we want the error less than $O\left((\pi_1^*/\pi_j^*)/k\right)$.

When one of the above events are violated, we have no control on the weights from wrong components. However, we can instead control the portion of these bad samples. For instance, consider $\mathcal{E}_1$ is violated, *i.e.*, measurement noise happens to be large. A Gaussian tail bound gives $P(\mathcal{E}_1^c) \le 2\exp(-\tau_j^2/2)$. This small probability is then used to bound the error from bad events,

$$\|\mathbb{E}_{\mathcal{D}_j}[\Delta_w X(Y - \langle X, \beta_1^*\rangle)\mathbb{1}_{\mathcal{E}_1^c}]\| \le P(\mathcal{E}_1^c) \sup_{s\in\mathbb{S}^{d-1}} \mathbb{E}_{D_j}[|\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)||\mathcal{E}_1^c].$$

We are left with bounding the expectation conditioned on $\mathcal{E}_1^c$, which turns out to be $O(R_{j1}^*)$. With the choice of $\tau_j = \Theta\left(\sqrt{\log(R_{j1}^* k\rho_\pi)}\right)$, this term is bounded by small value.

In order to handle two other cases $\mathcal{E}_2^c$ or $\mathcal{E}_3^c$, we need the following lemma:

Lemma 4.2 *Let $X \sim \mathcal{N}(0, I_d)$. Suppose any fixed vector $v \in \mathbb{R}^d$, a set of vectors $u_1, ..., u_k \in \mathbb{R}^d$ such*

that $\|u_j\| \geq \|v\|$ for all j, and constants $\alpha_1, ..., \alpha_k > 0$. Then consider two events

$$\mathcal{E} := \{|\langle X, u_j\rangle| \geq |\langle X, v\rangle|, \ \forall j = 1, ..., k\},$$
$$\mathcal{E}' := \{|\langle X, u_j\rangle| \geq \alpha_j, \ \forall j = 1, ..., k\}.$$

Then for any fixed unit vector $s \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[|\langle X, s\rangle|^2 | \mathcal{E}^c], \ \mathbb{E}[|\langle X, s\rangle|^2 | \mathcal{E}'^c] \leq C \log k, \qquad (2)$$

for some universal constant $C > 0$.

This lemma is crucial to bound the conditional expectation of errors when the sample is not in the good event set. A weaker version of this lemma appeared in Yi et al. (2016) with the (weaker) bound $O(k)$ in (2). The improvement has two consequences. First, it improves required initialization from $O(1/k^2)$-proximity to $O(1/k)$. Second, this bound is used critically in controlling the required SNR in Theorem 3.1. Using the previous results would have produced a k^2-scaling of the SNR. The proof of this lemma is given in Appendix A.1.

The rest of the proof follows similarly. The complete proof for $D_m \geq 1$ including Lemma 4.1 can be found in Appendix A.2.

When $D_m \leq 1$, we first apply the mean-value theorem to get a tighter error bound that is proportional to D_m. Then we can construct similar events that define good samples, and apply the same approach. Since $D_m \leq 1$ case involves heavy algebraic manipulation, we defer the proof for this case until Appendix D.

Remark 4 (Unknown mixing weights) *Our analysis also shows that EM succeeds when it is simultaneously estimating mixing weights and parameters. Tensor-based methods can also provide a good estimation of these mixing weights along with the estimate of regression vectors when they are not known in advance. In order to handle unknown mixing weights, the only additional requirement is the initial guess of mixing weights to be close in relative scale, i.e., $|\pi_j - \pi_j^*| \leq c\pi_j^*$ for some small $c > 0$, which we set $1/2$ as a requirement for the initial estimator.*

When $D_m \geq 1$, the only change in the proof is replacing π_j^* with π_j, i.e., using the estimator of weights instead of true weights. In this regime, it is enough to show that π_j^+ stays in the neighborhood of true mixing weights, i.e., $|\pi_j^+ - \pi_j^*| \leq \pi_j^*/2$. When $D_m \leq 1$, when we apply mean-value theorem, there is an additional term differentiated by mixing weights, which can also be bounded using the same approach. In this regime, there is also an improvement over mixing weights, i.e., $\max_j |\pi_j^+ - \pi_j^*|/\pi_j^* \leq \gamma \max(\max_j |\pi_j - \pi_j^*|/\pi_j^*, D_m)$ for some $\gamma < 1/2$ under the SNR and initialization condition we assume. Proofs for unknown mixing weights are given in Appendix A.2 for case $D_m \geq 1$, and Appendix D for case $D_m \leq 1$.

4.2 Bounding A

We give a lower bound on the minimum eigenvalue of $\mathbb{E}_{\mathcal{D}}[w_1 XX^\top]$. We first observe that

$$\mathbb{E}_{\mathcal{D}}[w_1 XX^\top] = \sum_j \pi_j^* \mathbb{E}_{\mathcal{D}_j}[w_j XX^\top] \succeq \pi_1^* \mathbb{E}_{\mathcal{D}_1}[w_1 XX^\top].$$

Then the minimum singular value of the right hand side is lower bounded by $\pi_1^*/2$. given good initialization $D_m/R_{min} = 1/\tilde{O}(k)$ and SNR $R_{min} = \tilde{\Omega}(k)$. The detailed proof including the lemma can be found in Appendix A.3.

Proof of Theorem 3.1. From Lemma 4.1, given $R_{min} = Ck\rho_\pi \log^2(k\rho_\pi) = \tilde{\Omega}(k\rho_\pi)$ and $D_m = cR_{min}/(k\rho_\pi \log k)$, we have $\|B\| \leq (\pi_1^*/4)D_m$ with proper universal constant $C, c > 0$. Similarly, from Lemma A.3. we get $\|A^{-1}\|_{op} \leq 2/\pi_1^*$. Then

$$D_m^+ := \max_j \|\beta_j^+ - \beta_j^*\| \leq \|A^{-1}\|_{op}\|B\|_2 \leq D_m/2.$$

We can conclude that after $T = O(\log(\max_j \|\beta_j^{(0)} - \beta_j^*\|/\epsilon)$ iterations, we have $\max_j \|\beta_j^{(t)} - \beta_j^*\| = O(\epsilon)$, thus we get Theorem 3.1. □

5 Finite Sample EM Analysis

In the finite sample version of EM, the estimation error at the next iteration in this problem is:

$$\tilde{\beta}_1^+ - \beta_1^* = (\sum_i w_{1,i} X_i X_i^\top)^{-1} (\sum_i w_{1,i} X_i (y_i - \langle X_i, \beta_1^*\rangle)).$$

The key quantity we will focus is the deviation of each empirical sums from their respected true means. That is,

$$A_n := 1/n \sum_i w_{1,i} X_i X_i^\top,$$
$$e_B := \big(1/n \sum_i w_{1,i} X_i (y_i - \langle X_i, \beta_1^*\rangle) - \mathbb{E}_{\mathcal{D}}[w_1 X (Y - \langle X, \beta_1^*\rangle)]\big).$$

Note that $\tilde{\beta}_1^+ - \beta_1^* = A_n^{-1} B_n$ where $B_n := B + e_B$. In the analysis of population EM, we have shown that $\|B\| \leq c_B D_m \pi_1^*$ for some universal constant c_B. Thus, we only have to bound e_B, which is the deviation of finite sample mean from true mean B. Then we analyze the minimum singular value of A_n similarly

by relating it to A. We focus on the concentration of sums in one-step iteration of EM. We assume that we use sample-splitting finite sample EM as we defined in Section 3, and we run EM for T iterations.

Before getting into our finite-sample analysis, we discuss briefly why we do not use a simpler standard concentration argument. Note that our target for giving a concentration bound is the random variable $w_1 X(Y - \langle X, \beta_1^* \rangle)$. On its own, it is a sub-exponential random variable, since $|w_1| \le 1$, X is sub-Gaussian (vector) with parameter $O(1)$, and $Y - \langle X, \beta_1^* \rangle$ is also sub-Gaussian with parameter at most $1 + R_{max}$. Thus, we can apply well-known sub-exponential tail bounds with parameter $O(R_{max})$, and a standard $1/2$ covering-net argument over the unit sphere to get a high probability guarantee. However, in this manner, we can only get a $O(R_{max}\sqrt{d/n})$ deviation of sample mean from true mean.

Most previous results established on finite-sample EM analysis have this dependency on R_{max} for statistical error Balakrishnan et al. (2017); Yi and Caramanis (2015); Klusowski et al. (2019); Yan et al. (2017); Zhao et al. (2018). In truth, however, this is an artifact of analysis and not a real phenomenon: the true statistical precision is $O(\sqrt{d/n})$ when noise is comparably less than R_{max}. For instance, in the extreme scenario, Yi et al. (2016) established exact recovery guarantee of EM in a noiseless setting, though it has not been obvious how to generalize their analysis to involve some level of noise.

Now we turn our attention to give a bound for e_B, which is given by the following lemma:

Lemma 5.1 *Suppose SNR condition $R_{min} \ge Ck\rho_\pi \log(k\rho_\pi)$ with sufficiently large $C > 0$ and initialization condition $D_m \le cR_{min}/(k\rho_\pi \log(k\rho_\pi))$ for sufficiently small $c > 0$. Given $n = \tilde{O}((k/\pi_{min})(d/\epsilon^2))$ samples, we get $\|e_B\| \le D_m \epsilon \pi_1^* + \epsilon \pi_1^*$ with high probability.*

The detailed proof is in Appendix B.1. Our proof strategy is to get a sharp concentration result is to partition random variables using indicator functions for disjoint events. Let $\mathcal{E}_j$ be the event that the the sample comes from the j^{th} component and $j \ne 1$. Then, consider events as in (1) in the population EM. We then decompose each sample using the indicator functions of these events. For simplicity of notation, let $W_i = w_{1,i} X_i (Y - \langle X_i, \beta_1^* \rangle)$. We can decompose W_i as, for instance,

$$w_{1,i} X_i (y_i - \langle X_i, \beta_1^* \rangle) = \sum_{j=1}^{k} \Big(W_i \mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,good}} + W_i \mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,1}^c} + W_i \mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,1} \cap \mathcal{E}_{j,2}^c} + W_i \mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,1} \cap \mathcal{E}_{j,2} \cap \mathcal{E}_{j,3}^c} \Big),$$

using the definition of events defined in (1). Then we can provide a finite-sample analysis with the following proposition.

Fact 5.2 *Let X be a random variable defined in some probability space, and consider a set of disjoint events $A_1, ..., A_m$ on X, such that $P(\cup_{i=1}^m A_i) = 1$. Then,*

$$P(|X - \mathbb{E}[X]| \ge t) \le \sum_{i=1}^{m} P(|X\mathbb{1}_{A_i} - \mathbb{E}[X\mathbb{1}_{A_i}]| \ge t_i),$$

for $\sum_{i=1}^m t_i = t$.

This is a restatement of the elementary union bound and the proof is immediate. It tells us that we can bound tail probabilities of decomposed random variables separately, and then collect them. If for all i, $P(|X\mathbb{1}_{A_i} - \mathbb{E}[X\mathbb{1}_{A_i}]| \ge t_i) \le \delta/m$, then $P(|X - \mathbb{E}[X]| \ge t) \le \delta$. Note that this decomposition is only for the analysis purpose and does not affect the practical implementation of the EM algorithm.

The next proposition is the key ingredient for giving a sharp concentration on each decomposed random variable.

Proposition 5.3 *Let X be a random d-dimensional vector, and A be an event defined in the same probability space with $p = P(X \in A) > 0$. Let random variable $Y = X|A$,* i.e., *X conditioned on event A, and $Z = \mathbb{1}_{X \in A}$. Let X_i, Y_i, Z_i be the i.i.d. samples from corresponding distributions. Then, equation (3) holds for any $0 \le n_e \le n$ and $t_1 + t_2 = t$.*

The proof of the Proposition 5.3 is given in Appendix C. When applied to the problem of mixed linear regression, Proposition 5.3 helps us to accurately control

$$P\left(\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \mathbb{1}_{X_i \in A} - \mathbb{E}[X\mathbb{1}_{X \in A}] \right\| \ge t \right) \le \max_{m \le n_e} P\left(\frac{1}{n} \left\| \sum_{i=1}^{m} (Y_i - \mathbb{E}[Y]) \right\| \ge t_1 \right) + P\left(\|\mathbb{E}[Y]\| \left| \frac{1}{n} \sum_{i=1}^{n} Z_i - p \right| \ge t_2 \right) + P\left(\left| \sum_{i=1}^{n} Z_i \right| \ge n_e + 1 \right) \tag{3}$$

the concentration of random vectors under different events: in a major event where samples are good, we know that $w_{i,1}$ is almost always exponentially small $O(\exp(-\tau_j^2))$ when the sample did not come from the first model. Therefore, norm of W_i conditioned on good event can be controlled with tiny $w_{i,1}$ (see Appendix B.1 for precise construction).

On bad events, such as when noise happens to be very large, the (sub-exponential) norm of W_i may be as large as R_{max}, since the weights of the wrong components could be away from zero. Fortunately, we can survive from these errors due to low chance of bad events given large SNR and good initialization. It enables us to choose n_e small enough while suppressing $P(|\sum_i^n Z_i| \geq n_e + 1)$, and n_e/n cancels out large norm of W_i conditioned on bad but rare events. This technique is critical not only for removing the dependency on R_{max}, but also obtaining as small dependency on k and π_{min} as possible.

We see in the proof that W_i conditioned on each event is another sub-exponential random vector with different sub-exponential norm. Therefore, we can give a sharp concentration bound on every decomposed random variable separately. The full details are in Appendix B.

Proof of Theorem 3.2. Given $\|e_B\| \leq D_m \epsilon \pi_1 + \epsilon \pi_1$ and results from population EM, we are left with bounding $A_n := (1/n \sum_{i\in[n]} w_{i,1} X_i X_i^\top)$. This task can be achieved via a direct application of standard concentration arguments for random matrices Vershynin (2010). When $(1/n \sum_{i\in[n]} w_{i,1} X_i X_i^\top)$ concentrates well around $\mathbb{E}_{\mathcal{D}}[w_1 X X^\top]$ in operator norm, we can conclude that the lower bound of minimum eigenvalue of sample covariance is also lower-bounded by $\pi_1^*/2$, which implies $\|A_n^{-1}\|_{op} \leq 2/\pi_1^*$ (see Appendix B.2 for the concentration result of A_n).

Then combining two results, we can conclude that

$$\begin{aligned}\|A_n^{-1}\|_{op}\|B_n\| &\leq 2{\pi_1^*}^{-1}(c_B D_m \pi_1^* + c_1 D_m \epsilon \pi_1^* + c_2 \epsilon \pi_1^*) \\ &\leq \gamma_n D_m + O(\epsilon),\end{aligned}$$

for some $\gamma_n < 1/2$ and universal constant c_B, c_1, c_2. This result for 1^{st} component is shown to hold with probability at least $1 - \delta/kT$ for a given failure probability $\delta > 0$ (see (23) in Appendix B.1 for detailed sample complexity). We can get a same result for other components, and thus we can take a union bound over k components. Thus, we have shown that with probability at least $1 - \delta/T$

$$\max_j \|\tilde{\beta}_j^+ - \beta_j^*\| \leq \gamma_n \max_j \|\tilde{\beta}_j - \beta_j^*\| + O(\epsilon).$$

Iterating over T iterations yields Theorem 3.2. □

Remark 5 (Unknown mixing weights) *Proof for mixing weights are also based on the same idea using Proposition 5.3. Mixing weights will also be well concentrated in relative scale, i.e.,* $|\tilde{\pi}_1 - \pi_1| \leq O(\epsilon)\pi_1^*$. *In the finite sample regime, mixing weights might not be exactly recovered even if noise power goes to 0. This does not conflict with the exact recovery guarantee for* β *whose statistical error is proportional to* σ *that goes to 0, since when* $\max_j \|\tilde{\beta}_j - \beta_j^*\| \geq \sigma$ *or* $D_m \geq 1$, *we do not require mixing weights to be very close (we only require* $|\pi_1 - \pi_1^*| \leq \pi_1^*/2$*) to get an improved estimator after one EM iteration. In other words, we do not require exact value of mixing weights* π *in order to get exact regression parameters* β. *Note that in the noiseless setting, we are always in* $D_m \geq 1$ *regime. Proof for concentration of mixing weights are given in Appendix B.3.*

6 Conclusion and Future Works

In this paper, we provided local convergence guarantees of both population and finite-sample EM algorithm for MLR with general k components. For our finite-sample based EM analysis, we decomposed a single random variable into multiple random variables using indicator functions, each of which corresponds to different event. With this strategy, we were able to give a near-optimal statistical error that does not depend on the distances between regression parameters. We believe our technique is applicable to other problem settings such as GMM, and other local heuristic algorithms to get an improved statistical error.

While we studied the local convergence of the EM algorithm in high SNR regime, the question whether EM converges under lower SNR condition, *i.e.* $R_{min} = o(k)$ regime, is still widely open, even when we assume we start from a very good initialization. We do not even know that under this low SNR regime, whether we can recover all parameters with polynomial number of samples (in k). It will be also interesting to consider a milder condition for the initialization requirement. Finding a polynomial-time algorithm to find a good initialization is also another challenging problem, if the linear independence assumption between regression vectors does not hold (*e.g.* $d < k$). Studying the EM algorithm in more general settings, *e.g.* with unknown and different covariance for X in each linear model, will be also an interesting future direction.

Acknowledgement

This work was partially funded by NSF grants 1609279, 1646522, and 1704778.

References

S. Balakrishnan, M. J. Wainwright, B. Yu, et al. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

T. T. Cai, J. Ma, L. Zhang, et al. Chime: Clustering of high-dimensional gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pages 1040–1048, 2013.

Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604, 2014.

Y. Chen, X. Yi, and C. Caramanis. Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory*, 64(3):1738–1766, 2017.

C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two gaussians. In *30th Annual Conference on Learning Theory*, 2017.

P. Hand and B. Joshi. A convex program for mixed linear regression with a recovery guarantee for well-separated data. *Information and Inference: A Journal of the IMA*, 7(3):563–579, 2018.

C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in neural information processing systems*, pages 4116–4124, 2016.

J. M. Klusowski, D. Yang, and W. Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 2019.

J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis. Global convergence of the EM algorithm for mixtures of two component linear regression. In *32nd Annual Conference on Learning Theory*, pages 2055–2110. PMLR, 2019.

Y. Li and Y. Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144, 2018.

O. Regev and A. Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. *Proceedings of Machine Learning Research*, 51:1223–1231, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

C. J. Wu et al. On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1):95–103, 1983.

B. Yan, M. Yin, and P. Sarkar. Convergence of gradient EM on multi-component mixture of gaussians. In *Advances in Neural Information Processing Systems*, pages 6956–6966, 2017.

X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

X. Yi, C. Caramanis, and S. Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

R. Zhao, Y. Li, and Y. Sun. Statistical convergence of the EM algorithm on gaussian mixture models. *arXiv preprint arXiv:1810.04090*, 2018.

K. Zhong, P. Jain, and I. S. Dhillon. Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pages 2190–2198, 2016.