On the Minimax Optimality of the EM Algorithm for Learning Two-Component Mixed Linear Regression

Jeongyeol Kwon **Nhat Ho** **Constantine Caramanis**
University of Texas, Austin

Abstract

We study the convergence rates of the EM algorithm for learning two-component mixed linear regression under all regimes of signal-to-noise ratio (SNR). We resolve a long-standing question that many recent results have attempted to tackle: we completely characterize the convergence behavior of EM, and show that the EM algorithm achieves minimax optimal sample complexity under all SNR regimes. In particular, when the SNR is sufficiently large, the EM updates converge to the true parameter θ^* at the standard parametric convergence rate $\mathcal{O}((d/n)^{1/2})$ after $\mathcal{O}(\log(n/d))$ iterations. In the regime where the SNR is above $\mathcal{O}((d/n)^{1/4})$ and below some constant, the EM iterates converge to a $\mathcal{O}(\mathrm{SNR}^{-1}(d/n)^{1/2})$ neighborhood of the true parameter, when the number of iterations is of the order $\mathcal{O}(\mathrm{SNR}^{-2}\log(n/d))$. In the low SNR regime where the SNR is below $\mathcal{O}((d/n)^{1/4})$, we show that EM converges to a $\mathcal{O}((d/n)^{1/4})$ neighborhood of the true parameters, after $\mathcal{O}((n/d)^{1/2})$ iterations. Notably, these results are achieved under mild conditions of either random initialization or an efficiently computable local initialization. By providing tight convergence guarantees of the EM algorithm in middle-to-low SNR regimes, we fill the remaining gap in the literature, and significantly, reveal that in low SNR, EM changes rate, matching the $n^{-1/4}$ rate of the MLE, a behavior that previous work had been unable to show.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 INTRODUCTION

The expectation-maximization (EM) algorithm is a general-purpose heuristic to compute a maximum-likelihood estimator (MLE) for problems with missing information (Dempster et al., 1997; Wu, 1983; Redner and Walker, 1984). In general, computing the MLE is intractable due to the non-concave nature of log-likelihood functions in the presence of missing data. The EM algorithm iteratively computes a tighter lower bound on log-likelihood functions, with each iteration no more complex than solving a maximum-likelihood (ML) problem without missing data. Due to its simplicity and broad success in practice, EM is one of the most popular methods-of-choice in a variety of applications (Jordan and Xu, 1995; Ma et al., 2000; Li et al., 2009; Chen and Li, 2009).

Recent years have witnessed remarkable progress in establishing theory describing the non-asymptotic convergence of EM to the true parameters on canonical examples such as a mixture of Gaussian distributions and mixed linear regression (see Prior Art below). In such models, a key factor in the analysis is the separation between components, or the "signal strength". Most prior work has studied strongly separated instances (high SNR) and established linear convergence of the EM algorithm with the standard parametric statistical rate $n^{-1/2}$. In contrast, the understanding of the EM algorithm in the weakly separated settings (low SNR), especially mixed linear regression, remains incomplete.

Our contributions: In this paper, we aim to fill the remaining gap in the literature with the minimax optimal sample complexity of the EM algorithm for learning two-component mixed linear regression in the weakly separated regime. In so doing, we provide a complete picture of the EM algorithm under all signal-to-noise ratio (SNR) regimes for symmetric two-component mixed linear regression, namely, $\frac{1}{2}\mathcal{N}(-X^\top\theta^*, (\sigma^*)^2) + \frac{1}{2}\mathcal{N}(X^\top\theta^*, (\sigma^*)^2)$ where $\sigma^* = 1$ is given and X follows the standard multivariate normal distribution in d dimensions. We define SNR as

$\eta := \|\theta^*\|$ since $\sigma^* = 1$. Notably, our results are obtained under mild conditions of either random initialization or an efficiently computable local initialization. While simplified, the model is complex enough to capture the most interesting behaviors of the EM algorithm for learning a mixed linear regression with two components, and reveals statistical behaviors in the low-to-middle SNR regimes that previous analysis had missed. In summary, our contributions are as follows.

1. **High-to-middle SNR regimes**: when $(d/n)^{1/4} \lesssim \|\theta^*\|$ (up to some logarithmic factor), the EM updates converges to θ^* within a neighborhood of $\mathcal{O}(\max\{1, \|\theta^*\|^{-1}\}(d/n)^{1/2})$ after $\mathcal{O}(\max\{1, \|\theta^*\|^{-2}\}\log(n/d))$ number of iterations.

2. **Low SNR regime**: when $\|\theta^*\| \lesssim (d/n)^{1/4}$ (up to some logarithmic factor), the EM algorithm converge to θ^* within a neighborhood of $\mathcal{O}((d/n)^{1/4})$ when the number of iterations is of the order of $\mathcal{O}((n/d)^{1/2})$.

3. **Global Convergence**: We demonstrate that EM converges from *any* randomly initialized point with high probability. Furthermore, we do not require sample-splitting in our analysis.

While we discuss the tightness of our result in a great detail in Section 2.3, we briefly explain the significance of our results. We focus primarily on two aspects of the EM algorithm: (i) statistical rate, and (ii) computational complexity. In the high SNR regime, we have linear convergence to true parameters within $\sqrt{d/n}$ rate as noted previously in the literature. In contrast, in the low SNR regime when $\|\theta^*\| \lesssim (d/n)^{1/4}$, the statistical rate is $(d/n)^{1/4}$. We explain this transition in statistical rate with a convergence property of the population EM in the middle-to-low SNR regimes. The upper bound given by EM matches the known lower bound for this problem in all SNR regimes (Chen et al., 2014). For the computational complexity, the number of iterations increases quadratically in the inverse of SNR until SNR reaches $(d/n)^{1/4}$. Interestingly, the number of iterations is naturally interpolated at SNR $= (d/n)^{1/4}$ from $\|\theta^*\|^{-2}\log(n/d)$ to $\sqrt{n/d}$. More in-depth discussions on the results (e.g., detailed comparison to previous works, proof techniques we use, etc.) are provided in Section 2.3.

1.1 Prior Art

While the classical results on the EM algorithm only guaranteed asymptotic convergence to *stationary points* (Wu, 1983), the seminal work (Balakrishnan et al., 2017) proposed a general framework to study a non-asymptotic convergence of the EM algorithm to *true parameters*. Motivated by this work, there has been a flurry of work studying the convergence of the EM algorithm to the true parameters for various kinds of regular mixture models (see e.g., (Yi et al., 2014, 2016; Xu et al., 2016; Yan et al., 2017; Daskalakis et al., 2017; Kwon and Caramanis, 2020b; Dwivedi et al., 2020b; Kwon and Caramanis, 2020a)). Most of the work in this line require strong separation compared to the noise level, i.e., considers the high SNR regime. Using this condition, it establishes linear convergence of EM to parameter estimates that lie within $(d/n)^{1/2}$-radius around the true location parameters. In contrast, relatively little understanding is available when different components in a mixture model are weakly separated (i.e., middle-to-low SNR). In particular, even for simple settings of two-component mixed linear regression that we consider in this work, our understanding on the EM algorithm still remains incomplete, for as we show, not only the techniques, but also the conclusions of past analysis no longer hold in the weakly separated regime.

The first convergence guarantees for EM under mixed linear regression was established in a noise-free setting (Yi et al., 2014, 2016). Subsequent results succeeded in treating the noisy setting (see (Balakrishnan et al., 2017)) for a mixture of two linear regressions, when the the signal strength $\|\theta^*\|$ is significantly larger than the noise variance σ^* (high SNR). Work in Kwon and Caramanis (2020b) extended the results in Balakrishnan et al. (2017) and Yi et al. (2016) to a more general setting of learning a mixture of k-component linear regressions when the SNR is $\Omega(k)$. However, it has not been obvious how to extend any of these results to the weakly separated regimes.

Recently, Kwon et al. (2019) has established the global convergence of the EM algorithm for learning a mixture of two linear regressions in all SNR regimes. While their result guarantees convergence of EM in all SNR regimes, the characterization of this convergence falls short in two aspects: (i) their analysis relies on the sample-splitting, (ii) their result is sub-optimal in terms of SNR in low SNR regime. In order to elaborate more on the second aspect, the statistical rate in Kwon et al. (2019) is given as $O(\eta^{-6}n^{-1/2})$ given that the sample size $n \gtrsim \eta^{-6}$ is sufficiently large. However, it is known that in the limit setting of $\eta \to 0$, the rate of MLE slows down to $n^{-1/4}$ (Chen, 1995; Ho and Nguyen, 2016; Ho et al., 2019). The result in Kwon et al. (2019) fails to capture this important property in relation to EM, and gives little insight on what happens when there is a large overlap between components. Our results tighten the sub-optimal analysis for middle SNR regime in Kwon et al. (2019) and fill in the remaining gap in the literature by providing a tight convergence guarantee of the EM algorithm in low SNR regime.

In a closely related problem of learning mixtures of two Gaussians, Dwivedi et al. (2020a, 2018, 2020b) recently studied an extreme case of the over-specified mixture models, *i.e.,* there is no separation between two components. However, their analysis is restricted to strictly over-specified settings, and it has not been obvious to extend their result to weakly-separated models. In another recent work, Wu and Zhou (2019) has studied the EM algorithm for learning a mixture of two weakly-separated location Gaussians, establishing a minimax rate of the EM algorithm after $O(\sqrt{n/d})$ iterations in middle-to-low SNR regimes. However, their result requires the initialization to be already within a small Euclidean ball of $(d/n)^{1/4}$-radius, which is very restrictive. Our result does not suffer from small initialization issue as in Wu and Zhou (2019). Furthermore, our proof strategy can be applied to resolve the open issue with small initialization in Wu and Zhou (2019).

We note in passing that the problem of solving mixed linear regressions is an interesting problem by itself. It arises in a number of applications (De Veaux, 1989; Grün et al., 2007), and has been extensively studied with various algorithms proposed (see e.g., (Chaganty and Liang, 2013; Chen et al., 2014; Sedghi et al., 2016; Yi et al., 2016; Li and Liang, 2018; Chen et al., 2019; Karmalkar et al., 2019; Raghavendra and Yau, 2020)). The special case of a mixture of two-component linear regressions is by now well understood (Yi et al., 2014; Chen et al., 2014; Kwon et al., 2019; Ghosh and Ramchandran, 2020). In this work, rather than solving a mixed linear regression itself, we focus on the rigorous study of the EM algorithm.

2 CONVERGENCE RATES OF EM

In this section, we first formulate symmetric mixed linear regression with two components and EM updates for this model in Section 2.1. Then, we state our main results with the convergence behaviors of EM algorithm under all regimes of SNR in Section 2.2. Then, we provide a detailed discussion with the tightness of the results in Section 2.3, and possible extensions to more unknowns in Section 2.4.

2.1 Problem setup

We assume that the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated from a symmetric two-component mixed linear regression, whose density function has the following form:

$$g_{\text{true}}(x, y) := \big(\frac{1}{2} f(y| - (\theta^*)^\top x, \sigma^*) + \frac{1}{2} f(y|(\theta^*)^\top x, \sigma^*)\big) \bar{f}(x), \quad (1)$$

where $\sigma^* = 1$ is given and θ^* is an unknown parameter. Furthermore, we assume that $\bar{f}(x)$ is the density of standard multivariate Gaussian distribution, i.e., $X \sim \mathcal{N}(0, I_d)$. In order to estimate θ^*, we fit the data by using symmetric two-component mixed linear regression, which is given by:

$$g_{\text{fit}}(x, y; \theta) := \big(\frac{1}{2} f(y| - \theta^\top x, \sigma^*) + \frac{1}{2} f(y|\theta^\top x, \sigma^*)\big) \bar{f}(x). \quad (2)$$

It is clear that $g_{\text{fit}}(x, y; \theta^*) = g_{\text{true}}(x, y)$. A common approach to obtain an estimator for θ^* is by using maximum likelihood esimation (MLE). However, given that the log-likelihood function of symmetric two-component mixed linear regression is highly non-concave, the MLE does not have a closed-form expression. EM is a popular iterative algorithm to approximate the MLE. Given fitted model (2), simple algebra shows that the EM update for θ can be written as follows:

$$\theta_n^{t+1} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \times \left(\frac{1}{n}\sum_{i=1}^n \tanh\left(\frac{Y_i X_i^\top \theta_n^t}{\sigma^{*2}}\right) Y_i X_i\right), \quad (3)$$

where the hyperbolic function $\tanh(x) := (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x))$ for all $x \in \mathbb{R}$. In order to facilitate the ensuing argument, let us the denote population and finite-sample EM operators by Eqns. 4 and 5, respectively, as given below:

$$M_{mlr}(\theta) := \mathbb{E}[XY \tanh(YX^\top\theta)], \quad (4)$$

$$M_{n,mlr}(\theta) := \left(\frac{1}{n}\sum_i X_i X_i^\top\right)^{-1} \times \left(\frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta)\right). \quad (5)$$

Motivation from experiments: In Figure 1, we present the statistical rate and optimization complexity of EM algorithm under different regimes of SNR. We set $d = 5$ and initialized the estimator in the neighborhood of the true parameters such that $\theta^0 = \theta^* + ru$, where $r = \max\{1, \|\theta^*\|\} \cdot 0.1$ and u is a random unit vector. For measuring the statistical rate, the EM algorithm runs with different size of samples $n \in \{128, 180, 256, ...\}$ (approximately $\sqrt{2}$ times increased) and the final error is averaged over $5,000$ independent runs. The stopping criterion is the change in estimators being less than 0.0001 in l_2 norm. In Figure 1 (a), we observe the standard $n^{-1/2}$ rate in the high SNR regime, and $n^{-1/4}$ rate in the low SNR regime. Interestingly, we can see a clear transition in the statistical rate when SNR $= 0.3$

as n increases. This explains how the low SNR regime is defined $\|\theta^*\| \lesssim (d/n)^{1/4}$: the meaning of low SNR depends on how many samples we have, not on the absolute value that can be computed from a problem instance.

We also look at the optimization complexity in Figure 1 (b, c). We run the EM algorithm with fixed sample size $n = 32768$. Estimation error $\|\theta_n^t - \theta^*\|$ in all iteration steps are averaged over $5,000$ independent runs. In the high SNR regime, note that the y-axis is in log-scale and we can see the linear convergence. In contrast, in the middle-to-low SNR regimes, we can observe that the convergence of the EM algorithm is no longer linear, and significantly slowed down.

2.2 Main results

In this section, we state our main results with the convergence behaviors of the EM algorithm under different regimes of SNR. Our first result assumes a good initialization and focuses on the statistical optimality of the EM algorithm in the last iterations. We can use the standard spectral method to get such a good initialization (see Appendix F.1 for guarantees given by the spectral initialization). Then, with a mild condition on SNR and permission to use a simple variant of EM, our second result shows that EM converges globally to the true parameter with the same optimal statistical rates.

Throughout the paper, we assume that $n \geq Cd$ for sufficiently large constant $C > 0$. Our analysis is divided into two cases when we are in the middle-high SNR regimes and low SNR regime. We state our first main theorem:

Theorem 1. *(a) (Middle-High SNR regimes) Suppose $\|\theta^*\| \geq C_0(d\log^2(n/\delta)/n)^{1/4}$ for some large universal constant $C_0 > 0$. In this regime, suppose we run the EM algorithm starting from well-initialized θ_n^0 such that $\|\theta_n^0\| \geq 0.9\|\theta^*\|$ and $\cos\angle(\theta^*, \theta_n^0) \geq 0.95$. Then, for any $\delta > 0$ there exist universal constants $C_1, C_2 > 0$ such that the EM updates (3) give θ_n^t for θ^* which satisfies*

$$\|\theta_n^t - \theta^*\| \leq C_1 \max\{1, \|\theta^*\|^{-1}\}(d\log^2(n\|\theta^*\|/\delta)/n)^{1/2},$$

with probability at least $1 - \delta$ after $t \geq C_2 \max\{1, \|\theta^*\|^{-2}\}\log(n\|\theta^*\|/d)$ iterations.

(b) (Low SNR regime) When $\|\theta^*\| \leq C_0(d\log^2(n/\delta)/n)^{1/4}$, there exist universal constants $C_3, C_4 > 0$ such that the EM updates (3) initialized with $\|\theta_n^0\| \leq 0.2$ return θ_n^t which satisfies

$$\|\theta_n^t - \theta^*\| \leq C_3(d\log^2(n/\delta)/n)^{1/4},$$

with probability at least $1 - \delta$ after $t \geq C_4\log(\log(n/d))\sqrt{n/(d\log^2(n/\delta))}$ iterations.

The proof sketch of Theorem 1 is in Section 3 while the full proof is in Appendix B. Interestingly, the upper bound given by Theorem 1 matches the known lower bounds given for all SNR regimes in Chen et al. (2014), and explains detailed behavior that interpolates between different separation regimes. Note that, the additional requirement $\|\theta_n^0\| \geq 0.9\|\theta^*\|$ under middle-high SNR regimes is to prevent the analysis to become over-complicated (see Appendix C.3 for the arguments for starting from well-aligned small estimators). Furthermore, the initialization condition $\|\theta_n^0\| \leq 0.2$ in the low SNR regime is not restrictive. In Appendix C.1, we demonstrate that when we initialize with large norm such that $\|\theta_n^0\| \geq 0.2$, in a finite number of steps the norm of EM updates becomes smaller than 0.2.

Next, we present our second result that does not rely on the warm start, but requires slightly more involved mechanisms. We call the following variant of EM as "Easy-EM" operator (Kwon et al., 2019):

$$M_{easy}(\theta) := \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \tanh(Y_i X_i^\top \theta). \tag{6}$$

Note that the only difference is the absence of the inverse of the sample covariance matrix. Our second theorem guarantees the global convergence of the EM algorithm with minimax optimality:

Theorem 2. *Given $C > 0$, suppose that $\|\theta^*\| \leq C$. Let θ_n^0 be a randomly initialized vector in $\mathbb{R}^d$ space such that the direction of θ_n^0 is randomly sampled from a uniform distribution on the unit sphere. The norm of initial estimator can be any non-zero constant such that $\|\theta_n^0\| \geq c(d\log^2(n/\delta)/n)^{1/4}$ for some universal constant $c > 0$.*

(a) In the middle-to-high SNR regimes, there exist universal constants $C_1, C_2, C_3 > 0$ such that when $C_1(d\log^2(n/\delta)/n)^{1/4} \leq \|\theta^*\| \leq C$, with probability at least $1 - \delta$, we have

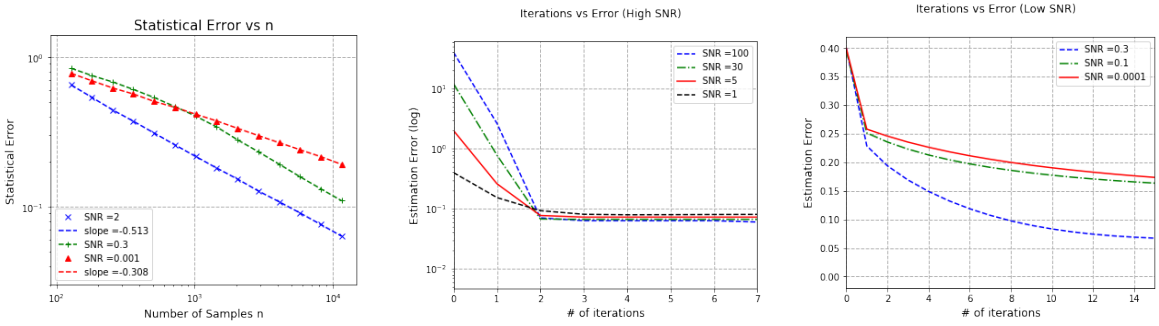
$$\|\theta_n^t - \theta^*\| \leq C_2 \max\{1, \|\theta^*\|^{-1}\}(d\log^2(n/\delta)/n)^{1/2},$$

after we first run the Easy-EM algorithm (6) for $C_3 \max\{1, \|\theta^*\|^{-2}\}\log(d)$ iterations, and then run the standard EM algorithm (4) for $C_3 \max\{1, \|\theta^*\|^{-2}\}\log(n/d)$ iterations.

(b) In the low SNR regime when $\|\theta^*\| \leq C_1(d\log^2(n/\delta)/n)^{1/4}$, there exist universal constants $C_4, C_5 > 0$ such that with probability at least $1 - \delta$, we have

$$\|\theta_n^t - \theta^*\| \leq C_4(d\log^2(n/\delta)/n)^{1/4},$$

after we run either Easy-EM or standard EM for $t \geq C_5\log(\log(n/d))\sqrt{n/(d\log^2(n/\delta))}$ iterations.



(a) (b) (c)

Figure 1. Convergence behavior of the EM algorithm for the fitted model (1) when $d = 5$: (a) statistical rates of EM iterates ($\|\theta_n^t - \theta^*\|$ at the last iteration) in various SNRs (b) linear convergence in high SNR regime (c) slow convergence in low SNR regime.

The proof sketch of Theorem 2 is in Section 3 while the full proof is in Appendix C. A few comments are in order. First, comparing to Theorem 1, we have an additional assumption for $\|\theta^*\|$ being bounded. This is required for a technical reason that arises from giving an uniform control on the deviation of Easy-EM operator in one direction when $\|\theta^*\|$ can be arbitrarily large (see Remark 2 in Appendix C.2 for details). Second, in order to correctly estimate how many iterations we must run Easy-EM, we can check the value of $\frac{1}{n}\sum_{i=1}^n Y_i^2 - 1$, since the expectation of this value is $\|\theta^*\|^2$. We note that Easy-EM is only introduced for a theoretical justification, and in practice we can just run the EM algorithm from a randomly initialized point. Finally, our condition on the norm of initial estimator is to ensure that the initial point is sufficiently far from zero. In practice, we use any constant $\Omega(1)$ for the norm of initial estimator. This is in stark contrast to the initialization of Wu and Zhou (2019) in which only very small initialization of order $\Theta((d/n)^{1/4})$ is allowed, which goes to 0 as $n \to \infty$.

2.3 Tightness of the results

In this section, we discuss in detail the tightness of our results in Theorem 1 and Theorem 2.

Tightness of the result in the high SNR regime: In the high SNR regime, a minimax rate should guarantee exact recovery when the noise variance goes to zero. Our results obtain a statistical rate of $\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}$. Note that, since we have rescaled to $\sigma^* = 1$, we should interpret the statistical rate of EM algorithm in the original scale where it is translated to $(\sigma^*\log(1/\sigma^*))\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}$. Therefore, we still guarantee the exact recovery as $\sigma^* \to 0$. We conjecture that a more careful and thorough analysis can also resolve even the logarithmic dependency on $\|\theta^*\|$, and leave it as future work. As mentioned earlier, there has been much recent interest in establishing the linear convergence and tight finite-sample error in high SNR regime (Yi et al., 2014, 2016; Kwon et al., 2019; Kwon and Caramanis, 2020b). While all previous results are also minimax optimal in all parameters (up to logarithmic factors), as an artifact of their analysis, their results rely on sample-splitting, and thus do not in fact analyze the algorithm that is used in practice. Our results remove this artifact.

A recent work in Ghosh and Ramchandran (2020) has established a super-linear convergence of the EM algorithm in the noiseless setting (a.k.a. Alternating Minimization). We conjecture that their result can be extended to the noisy setting when SNR is high enough (*i.e.*, $\|\theta^*\| \gg 1$). The following lemma on the population EM operator (5) gives a hint for a super-linear convergence in the high SNR regime:

Lemma 1. *If $C\sqrt{\log\|\theta^*\|} \le \|\theta - \theta^*\| \le \|\theta^*\|/10$ for sufficiently large constant $C > 0$, then there exists a constant $c < 10$ such that*

$$\|M_{mlr}(\theta) - \theta^*\| \le c\|\theta - \theta^*\|^2/\|\theta^*\|.$$

The proof of Lemma 1 is in Appendix F.2. This lemma implies that until $\|\theta - \theta^*\|$ drops from $O(\|\theta^*\|)$ to $O(\sqrt{\log\|\theta^*\|})$, the population EM updates converge in a super-linear rate. While interesting, we do not pursue a further extension of super-linear convergence to the noisy setting in this paper.

Tightness of the result in the middle-low SNR regimes: As discussed in the introduction, Kwon et al. (2019) has recently established a convergence of the EM algorithm in SNR regimes for model (2). In particular, according to the result in Kwon et al. (2019), the EM algorithm can achieve arbitrary ϵ accuracy if the sample size n is large enough to compensate a low SNR $\eta := \|\theta^*\|/\sigma^*$, *i.e.*, $\eta^{-6}/\epsilon^2 \lesssim n$. This sub-optimal result is an artifact of the technical approach

used to relate the population and finite-sample EM operators. Specifically, the convergence rate of the population EM operator is given by $1-\eta^2$. The finite-sample analysis then follows by analyzing the uniform deviation of finite-sample operators from population operators, which is in order of magnitude $\sqrt{d/n}$. In order to guarantee the progress toward θ^* in each step as well as to control the accumulation of statistical errors in all iterations, Kwon et al. (2019) required the sample size $n \gtrsim \eta^{-6}$ *per* iteration. The sample-splitting results in even worse total $n \gtrsim \eta^{-8}$ sample complexity in terms of SNR. Furthermore, nothing can be explained when the sample size is less than the threshold η^{-8}. This calls for a more refined analysis of the EM algorithm in middle-to-low SNR regimes.

In the paper, we adopt the localization argument used in the recent works (Dwivedi et al., 2020a,b) where the authors of these works established the convergence behaviors of the EM algorithm under over-specified Gaussian mixtures, namely, when there is no separation of the true parameters. Unlike these previous studies, our analysis is not restricted to strictly over-specified instances, but spans all possible configurations of parameters. The core of our analysis has three parts: (i) refined convergence rate of the population EM operator, namely, the contraction coefficient of the population EM operator is shown to be $1-\max\{\|\theta\|^2-\eta^2, \eta^2\}$, (ii) multi-level application of uniform deviation of finite-sample EM operators from the population EM operators that is proportional to $\|\theta\|\sqrt{d/n}$, and (iii) localization arguments applied to different levels of $\|\theta\|$. The threshold that separates the middle-SNR and low-SNR regimes can be naturally found at $\eta^2 = \sqrt{d/n}$.

Global Convergence of (Easy) EM: Global convergence of the EM algorithm for model (1) has been established in Kwon et al. (2019) using the idea of two-phase analysis where EM first converges in angle, and then converges in l_2 norm. In the initial stage of the EM iterations with a random initialization, Kwon et al. (2019) proposed a simple variant of the EM update (6) to encourage the boosting of angle from $\cos\angle(\theta_n^0, \theta^*) = O(1/\sqrt{d})$. Importantly, our result removes the usage of sample-splitting in Kwon et al. (2019) and tightens the sub-optimal statistical rate of the EM algorithm in middle-to-low SNR regimes as in Theorem 1.

In Wu and Zhou (2019), the authors employed a similar idea of analyzing the growth of the signal strength in the θ^* direction for learning a two symmetric mixture of Gaussian distributions. However, in general the value itself in θ^* direction can indeed decrease if EM starts from large initialization. Therefore, they restricted the initialization to be within a very small radius of $\|\theta_n^0\| \approx (d/n)^{1/4}$ in all SNR (separation) regimes. While it does not degrade the overall computational complexity of the finite-sample EM algorithm, the convergence guarantee with such small initialization is not global in a true sense. Theorem 2 resolves the open issue of small initialization in Wu and Zhou (2019) by analyzing the convergence in angle.

2.4 Towards unknown variance and weight

In this section, we discuss the statistical behavior of the EM algorithm when either the variance $\sigma^* = 1$ or the mixing weight of the true density g_{true} is unknown.

Unknown noise variance: We first discuss the case when the variance σ^* of regression noise is unknown. In this case, the EM updates for θ and σ are as follows:

$$\bar{\theta}_n^{t+1} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \times \left(\frac{1}{n}\sum_{i=1}^n \tanh\left(\frac{Y_i X_i^\top \theta_n^t}{(\bar{\sigma}_n^t)^2}\right) Y_i X_i\right),$$
$$(\bar{\sigma}_n^{t+1})^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - \big((\bar{\theta}_n^{t+1})^\top X_i\big)^2. \tag{7}$$

The EM update $\bar{\theta}_n^{t+1}$ in the unknown variance case depends on σ_n^t, which is updated at each iteration of the EM algorithm. It is different from the update for θ in the known variance case in equation (3). Therefore, the overall analysis of the EM algorithm in the unknown variance case should be re-derived in the population level to get right contraction coefficients of the population EM updates.

We would like to remark that the challenge with more unknowns arises from the convergence analysis in the population level, and the sample complexity analysis is irrelevant to whether we have more unknowns or not. In this section, we provide the statistical behavior of the EM algorithm under the low SNR regime and leave the complete analysis of the EM algorithm for future work. The localization technique used in the low SNR regime of known variance setting remains to be useful for obtaining the convergence and statistical rates of EM in the low SNR regime of unknown variance setting. It leads to the following result with the EM iterates in the low SNR regime.

Theorem 3. *(Low SNR regime of unknown variance case) There exist universal constants $C_0, C_1, C_2, C_3 > 0$ such that when $\|\theta^*\| \le C_0(d\log^2(n/\delta)/n)^{1/4}$, starting from $\|\bar{\theta}_n^0\| \le 0.2$ and $|(\bar{\sigma}_n^0)^2 - 1| \le 0.04$, the EM updates (7) return $(\bar{\theta}_n^t, \bar{\sigma}_n^t)$ which satisfies*

$$\|\bar{\theta}_n^t - \theta^*\| \le C_1(d\log^2(n/\delta)/n)^{1/4},$$

$$|(\bar{\sigma}_n^t)^2 - (\sigma^*)^2| \leq C_2(d\log^2(n/\delta)/n)^{1/2}, \qquad \geq 1-\delta. \quad (8)$$

with probability at least $1-\delta$ *after* $t \geq C_3 \log(\log(n/d))\sqrt{n/(d\log^2(n/\delta))}$ *iterations.*

Unlike in the case of Gaussian mixtures with unknown variances (Dwivedi et al., 2020b), the statistical rate of EM updates for θ is $(d/n)^{1/4}$ for all $d \geq 1$. When $\theta^* = 0$, this coincides with the previous result on the rate of maximum likelihood estimation for over-specified Gaussian mixture of experts (Ho et al., 2019), where it is shown that the MLE rate of estimating θ^* is $n^{-1/4}$ as long as the link functions are algebraically independent, which is the case for the unknown variance setting, and the number of components of Gaussian mixtures of experts is over-specified. The proof of Theorem 3 is given in Appendix F.3.

Unknown mixing weights: The extension to the unknown mixing weight can be more challenging, since the unbalanced mixing weight induces asymmetry in the landscape of the log-likelihood function. The asymmetry completely changes the population landscape of two-component mixed linear regression (e.g., there is a local maxima in the population log-likelihood for a mixture of two Gaussian distributions, which is absent in the symmetric setting (Xu et al., 2018)). It makes the analysis of the EM algorithm challenging even in the two-component settings of mixed linear regression. In high SNR regimes, we can avoid direct analysis of the optimization landscape and still can show the linear convergence of the EM iterates toward true parameters (Kwon and Caramanis, 2020b). However, in middle-to-low SNR regimes, we cannot avoid the analysis of complicated landscape. The extension to unknown mixing weights is an interesting future direction.

3 PROOF SKETCH

In this section, we provide proof sketches of key theorems in the paper.

3.1 Proof Sketch of Theorem 1

We first give a proof sketch of Theorem 1. The full proof of Theorem 1 is in Appendix B. We need the following uniform deviation bound between sample and population EM operators:

Lemma 2. *Given the population and finite-sample EM operators* M_{mlr}, $M_{n,mlr}$ *in equations* (5) *and* (4), *for any given* $r > 0$, *there exists a universal constant* $c > 0$ *such that we have*

$$\mathbb{P}\left(\sup_{\|\theta\| \leq r} \|M_{n,mlr}(\theta) - M_{mlr}(\theta)\| \leq cr\sqrt{d\log^2(n/\delta)/n} \right)$$

While the lemma is a straight-forward consequence of Lemma 11 given in Appendix E, this is the first key result to get a tight statistical rate. The proof of Lemma 2 can be found in Appendix D.3.

High SNR regime: $\|\theta^*\| \geq C$**.** The high-level proof in the high SNR regime follows a specialized proof strategy exploited in Kwon and Caramanis (2020b). The core idea is that for high SNR, most "good" samples are assigned correct (soft but almost hard) labels in E-step, and the portion of "bad" samples is negligibly small. Such an argument first appeared informally in Balakrishnan et al. (2017), and then was formally organized in Kwon and Caramanis (2020b,a) to establish a linear convergence and tight statistical rate. The full proof for the high SNR regime is given in Appendix B.1.

Middle SNR regime: $C_0(d\log^2(n/\delta)/n)^{1/4} \leq \|\theta^*\| \leq C$**.** We consider two cases, when $\|\theta^*\| \geq 1$ and $\|\theta^*\| \leq 1$.

Case (i) $1 \leq \|\theta^*\| \leq C$**:** Given the initialization conditions in Theorem 1, we can show that

$$\|M_{\text{mlr}}(\theta) - \theta^*\| < 0.9\|\theta - \theta^*\|.$$

Furthermore, from the uniform concentration bound (cf. Lemma 11 in Appendix E), we have

$$\|M_{n,mlr}(\theta) - M_{mlr}(\theta)\| \leq \sqrt{d\log^2(n/\delta)/n}$$

with probability at least $1-\delta$. Collecting these results, we can check that

$$\|\theta_n^t - \theta^*\| \lesssim (0.9)^t \|\theta - \theta^*\| + \sqrt{d\log^2(n/\delta)/n}.$$

Case (ii) $C_0(d\log^2(n/\delta)/n)^{1/4} \leq \|\theta^*\| \leq 1$**:** In this case, the result of Lemma 3 in Appendix B shows that

$$\|M_{mlr}(\theta) - \theta^*\| \leq \left(1 - O(\|\theta^*\|^2)\right)\|\theta - \theta^*\|. \quad (9)$$

As Lemma 2 and Corollary 2 in the Appendix make precise, we can infer that in order for the EM algorithm to make progress toward θ^*, we need

$$\|\theta^*\|^2\|\theta - \theta^*\| \gtrsim \|\theta\|\sqrt{d/n}.$$

Intuitively, EM converges to θ^* as long as such a relation holds, and until θ gets close enough to θ^* such that the above equation does not hold. In other words, in the last iterations when $\|\theta\| \approx \|\theta^*\|$, we have

$$\|\theta^*\|^2\|\theta - \theta^*\| \approx \|\theta^*\|\sqrt{d/n},$$

which implies the statistical rate should be on the order of $\|\theta^*\|^{-1}\sqrt{d/n}$. The full proof is given in Appendix B.2.

Low SNR Regime: $\|\theta^*\| \leq C_0(d\log^2(n/\delta)/n)^{1/4}$. In this case, even the standard spectral methods would not give a good initialization since the eigenspace is perturbed too much to be aligned with θ^* (see Lemma 13 in Appendix F.1 for the guarantees given by spectral methods). Instead, we assume the initial estimator to be $\|\theta_n^0\| \leq 0.2$.

The core of idea of the low SNR regime is that EM essentially cannot distinguish the cases between $\theta^* = 0$ and $\theta^* \neq 0$. Therefore, we aim to investigate $\|\theta\|$ instead of the estimation error $\|\theta - \theta^*\|$. If we can show that $\|\theta_n^t\| \leq c_1 \cdot (d/n)^{1/4}$, then given the condition of low SNR regime, we have $\|\theta_n^t - \theta^*\| \leq c_2 \cdot (d/n)^{1/4}$ where c_1, c_2 are some positive constants.

In the low SNR regime, there exist universal constants $c_l, c_u > 0$ such that for $\|\theta\| \leq 0.2$, we have

$$\|\theta\|(1 - 4\|\theta\|^2 - c_l\|\theta^*\|^2) \leq \|M_{\mathrm{mlr}}(\theta)\| \leq \|\theta\|(1 - \|\theta\|^2 + c_u\|\theta^*\|^2).$$

The statistical fluctuation of the finite-sample EM operator given in Lemma 2 shows that $\|M_{n,\mathrm{mlr}}(\theta) - M_{\mathrm{mlr}}(\theta)\| \leq c \cdot \|\theta\|\sqrt{d\log^2(n/\delta)/n}$, for some universal constant c. It is now more clear to see that since $\|\theta^*\|^2 \lesssim \sqrt{d/n}$, the above statistical error will subsume an extra $O(\|\theta^*\|^2)$ term in the contraction rate of the population EM operator. Therefore, the convergence behaviors of the finite-sample EM operator are essentially the same when $\theta^* = 0$ and $\theta^* \neq 0$.

The EM iterations stop improving the estimator when the statistical error becomes larger than the amount that the population EM can proceed:

$$\|\theta\|^2 \approx \sqrt{d\log^2(n/\delta)/n}.$$

Therefore, the statistical rate of the EM algorithm is achieved at $\|\theta\| \lesssim (d/n)^{1/4}$. The rest of the proof in the low SNR regime is a reminiscent of the localization arguments used in Dwivedi et al. (2018, 2020b), and can be found in Appendix B.3.

3.2 Proof Sketch of Theorem 2

The global convergence statement is subsumed into Theorem 1 when the estimator θ enters in the initialization region that Theorem 1 requires. Therefore we can focus on the iterations that θ stays outside of the initialization region. The key idea is to adopt the angle convergence argument presented in Kwon et al. (2019). Note that in low SNR regime, we do not need such an involved argument since the initialization only requires $\|\theta_n^0\| \leq 0.2$ (see Appendix C.1 for an argument why this initialization is easy to satisfied). In middle SNR regime where $(d/n)^{1/4} \lesssim \|\theta^*\| \leq 1$, the key property is the following angle inequality:

$$\cos\angle(M_{\mathrm{mlr}}(\theta), \theta^*) \geq (1 + c\|\theta^*\|^2)\cos\angle(\theta, \theta^*),$$

for some universal constant $c > 0$. We again see that the increase rate is $1 + O(\|\theta^*\|^2)$; however, the cosine value is very small $\Theta(1/\sqrt{d})$ at the initial stage. Then, the second key step is to show that

$$\cos\angle(M_{easy}(\theta) - M_{mlr}(\theta), \theta^*) \leq \epsilon_f/\sqrt{d},$$

for sufficiently small $\epsilon_f \lesssim \sqrt{d/n}$. At a high level, if it holds that $c\|\theta^*\|^2\cos\angle(\theta, \theta^*) \geq 2\epsilon_f/\sqrt{d}$, then we can guarantee that $\cos\angle(M_{easy}(\theta), \theta^*) \geq (1 + c\|\theta^*\|^2/2)\cos\angle(\theta, \theta^*)$. We can conclude that this is true in the middle-SNR regime since $\|\theta^*\|^2 \gtrsim (d/n)^{1/2}$. The argument in high-SNR regime is similar to middle-SNR regime. The formal proof is a bit more involved since we need to ensure that the statistical error in orthogonal directions does not dominate the angle (see Appendix C.2 for more detail).

4 CONCLUSION

In the paper, we completely characterize the convergence behavior of EM under all SNR regimes of symmetric two-component mixed linear regression. We view our results for this model as the first step towards a comprehensive understanding of the EM algorithm for learning weakly separated latent variable models. We now discuss a few future directions naturally arise from our work. First, in more general settings of weakly separated mixture models with k components, it is known that the rate of MLE can be $n^{-O(1/k)}$ in the worst case (Heinrich and Kahn, 2018). Furthermore, EM is known to suffer from very slow convergence in practice for instances with large overlaps. It is an important future direction to characterize the convergence behavior of the EM algorithm in such settings. Second, our results demonstrate that the EM algorithm has sublinear convergence to θ^* under middle and low SNR regimes. It respectively leads to $\|\theta^*\|^{-2}\log(n/d)$ and $\sqrt{n/d}$ number of iterations under middle-to-low SNR regimes, which result in high computational complexity. An important direction is to develop an alternative to EM algorithm that can achieve much cheaper computational complexity and also obtain minimax optimal sample complexity under all SNR regimes of mixed linear regression. Finally, while we prove that the EM algorithm achieves minimax optimal statistical convergence rates for learning two-component mixed linear regression, it is important to further develop uncertainty quantification for the EM iteratates, such as confidence intervals. It necessitates the future study on the central limit theorem of the EM algorithm under all regimes of SNR, which has remained a major open problem in the literature.

References

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

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

J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995.

J. Chen and P. Li. Hypothesis test for normal mixture models: The EM approach. *Annals of Statistics*, 37: 2523–2542, 2009.

S. Chen, J. Li, and Z. Song. Learning mixtures of linear regressions in subexponential time via fourier moments. *arXiv preprint arXiv:1912.07629*, 2019.

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.

C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

R. D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–38, 1997.

R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, and M. I. Jordan. Theoretical guarantees for EM under misspecified Gaussian mixture models. In *NeurIPS 31*, 2018.

R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 48:3161–3182, 2020a.

R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020b.

A. Ghosh and K. Ramchandran. Alternating minimization converges super-linearly for mixed linear regression. *arXiv preprint arXiv:2004.10914*, 2020.

B. Grün, F. Leisch, et al. Applications of finite mixtures of regression models. 2007.

P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46:2844–2870, 2018.

N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016.

N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence rates for Gaussian mixtures of experts. *arXiv preprint arXiv:1907.04377*, 2019.

M. I. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8, 1995.

S. Karmalkar, A. Klivans, and P. Kothari. List-decodable linear regression. In *Advances in Neural Information Processing Systems*, pages 7423–7432, 2019.

J. Kwon and C. Caramanis. The EM algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020a.

J. Kwon and C. Caramanis. EM converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pages 1727–1736, 2020b.

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 *Conference on Learning Theory*, pages 2055–2110, 2019.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

P. Li, J. Chen, and P. Marriott. Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96:411–426, 2009.

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

J. Ma, L. Xu, and M. I. Jordan. Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation*, 12:2881–2907, 2000.

P. Raghavendra and M. Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231, 2016.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7*, 2010.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

Y. Wu and H. H. Zhou. Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv preprint arXiv:1908.10935*, 2019.

J. Xu, D. Hsu, and A. Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*, 2016.

J. Xu, D. J. Hsu, and A. Maleki. Benefits of over-parameterization with EM. In *Advances in Neural Information Processing Systems*, pages 10662–10672, 2018.

B. Yan, M. Yin, and P. Sarkar. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems 30*, 2017.

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.