Alternating Minimization Converges Super-Linearly for Mixed Linear Regression

Avishek Ghosh[*] **Kannan Ramchandran**[*]

[*]Department of Electrical Engineering and Computer Sciences, UC Berkeley
avishek_ghosh@berkeley.edu, kannanr@eecs.berkeley.edu

Abstract

We address the problem of solving mixed random linear equations. We have unlabeled observations coming from multiple linear regressions, and each observation corresponds to exactly one of the regression models. The goal is to learn the linear regressors from the observations. Classically, Alternating Minimization (AM) (which is a variant of Expectation Maximization (EM)) is used to solve this problem. AM iteratively alternates between the estimation of labels and solving the regression problems with the estimated labels. Empirically, it is observed that, for a large variety of non-convex problems including mixed linear regression, AM converges at a much faster rate compared to gradient based algorithms. However, the existing theory suggests similar rate of convergence for AM and gradient based methods, failing to capture this empirical behavior. In this paper, we close this gap between theory and practice for the special case of a mixture of 2 linear regressions. We show that, provided initialized properly, AM enjoys a *super-linear* rate of convergence in certain parameter regimes. To the best of our knowledge, this is the first work that theoretically establishes such rate for AM. Hence, if we want to recover the unknown regressors upto an error (in ℓ_2 norm) of ϵ, AM only takes $\mathcal{O}(\log\log(1/\epsilon))$ iterations. Furthermore, we compare AM with a gradient based heuristic algorithm empirically and show that AM dominates in iteration complexity as well as wall-clock time.

1 INTRODUCTION

We assume that the measurements are coming from the following observation model:

$$y_i = \langle x_i,\, \theta_1^*\rangle z_i + \langle x_i,\, \theta_2^*\rangle (1-z_i) + w_i, \qquad (1)$$

for $i = 1, \ldots, n$, where the covariates are $\{x_i\}_{i=1}^n \in \mathbb{R}^d$ and the unknown regressors are θ_1^* and θ_2^*. z_i here is the (unknown) latent variable taking values $\{0, 1\}$. When $z_i = 1$, the i-th observation comes from the regressor θ_1^*, and $z_i = 0$ implies y_i coming from θ_2^*. w_i here denotes the additive noise. Given the covariate-response pairs $(x_i, y_i)_{i=1}^n$, the goal is to estimate (θ_1^*, θ_2^*) without the knowledge of $\{z_i\}_{i=1}^n$.

Let us provide some motivation for studying the model (1). When measurements are obtained from multiple latent classes and the goal is to estimate the underlying parameters, mixed linear regression is often a reasonable model to assume. The model is introduced by Wedel et al (1995) ([WD95]) and have become a standard framework for applications like health-care [DH00], market segmentation [WK12] and music perception [VT02]. Please refer to Grun et al (2007) ([GL$^+$07]) and the references therein for several other applications of the mixture model. Mixed regression model also serves as a theoretical tool for analyzing benchmark nonconvex optimization algorithms ([CL13]; [KYB19]) or analyzing new algorithms ([CYC14]). Furthermore, the mixed regression model is a close relative of the classical hierarchical mixtures of experts [JJNH91], which has several applications in Generative Adversarial Networks (GAN) and Gated Recurrent Units (GRU) ([MKV18]).

A generic way to approach problems with latent variables is via the EM algorithm or its variants. Alternating Minimization (AM), which can be thought as *hard-EM* is classically used to solve (1). In every iteration AM first guesses the labels, and subsequently solve the linear regression problems with the guesses labels. With Gaussian covariates and a proper initialization, AM provably converges to the optimal parameters at a linear rate. In Yi et al (2014) ([YCS14]), AM

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

for mixed linear regression was proposed and analyzed for the problem of 2 mixtures (similar to (1)), and later it has been extended to the setting with more than 2 mixtures ([YCS16]). A few recent works ([WGNL15], [ZWZG17], [CB18], [CXW$^+$18] [KYB19], [KC19]) also use AM (or its variant EM) to tackle different aspects of the mixed linear regression and related problems.

Another line of work uses gradient descent algorithm to solve (1). Although solving a global non-convex problem (owing to unknown labels) Zhong et al (2016) ([ZJD16]) shows that under certain assumptions, the problem is locally strongly convex and hence gradient descent converges at a linear rate. Later Li et al (2018) ([LL18]) improves the sample and computational complexity via a careful analysis of the gradient descent algorithm. Furthermore, Chaganty et al (2013) ([CL13]) and Sedghi et al (2016) ([SJA16]) use tensor method to solve the mixed linear problem and Chen et al (2014) ([CYC14]) provides a convex relaxation formulation of a mixture of 2 regression problem and proposes a mini-max optimal algorithm. However, the tensor decomposition based method or the nuclear norm based convex relaxation method is computation heavy and slow.

Alternating Minimization is a general purpose algorithm used to solve non-convex problems with latent variables. A few examples of such problems include phase retrieval [NJS13], matrix sensing and completion [JNS13] and max-affine regression [GPGR19]. With proper initialization, it is empirically observed that AM is much faster than gradient based algorithms. For example, in the context of phase retrieval problem, Table 1 of Zhang et al (2016) ([ZL16]) shows that gradient descent takes 36× more iterations and 2.36× more wall-clock time compared to AM. Using a truncation and reshaping technique particular to the phase retrieval problem, Zhang et al (2016) reduces the wall-clock time but still requires 12× iterations over AM. It was also conjectured in Xu et al (1996) ([XJ96]), that EM (i.e., soft-AM) enjoys a super-linear rate of convergence, much like the Newton method for convex optimization. Also, Balakrishnan et al (2017) ([BWY17]) observes a very fast rate of convergence of EM when initialized properly for the problem of estimation from mixture of Gaussians. Later Daskalakis et al (2017) ([DTZ17]) shows that it is sufficient to run EM for only constant number of iterations for the Gaussian mixture problem, thus hinting towards a meteoric speed of convergence. However, to the best of our knowledge, there is no theoretical justification to this fact for either EM or AM.

The goal of this paper is to bridge this gap between theory and practice. We consider the classical AM algorithm to solve the mixed linear regression with 2 components. Via a careful analysis of the underlying empirical process of the AM iteration, we prove that the rate of convergence of AM in this setting is truly *super-linear*, thus explaining the empirical phenomenon mentioned above. We believe that this is the first work to theoretically establish such faster rate of convergence. We also run exhaustive experiments to show the super-linear rate of AM, and compare with a gradient based heuristic. We observe that AM dominates the gradient based heuristic in terms of both iteration complexity and wall-clock time.

Very recently, Shen et al (2019) ([SS19]) uses an iterative least trimmed squares (ILTS) algorithm for the mixture of regressions. Based on residual values, ILTS iteratively estimates the components of mixture one-by-one and removes the corresponding observations in subsequent iterations. Under certain settings, ILTS is provably shown to converge super-linearly. Note that, algorithmically, the classical AM is very different from ILTS. There is no (direct) latent variable estimation phase in ILTS. Also AM estimates the all regressions at once, instead of iterative estimation and trimming. Furthermore, the statistical tools and techniques we use are quite different from [SS19].

Our Contributions: We have the following:

- We analyze the classical AM (Algorithm 1) for a mixture of 2 regressions, and show that provided initialized properly, the rate of convergence is *super-liner* with exponent 3/2 (Theorem 1). Instead of using crude singular value bounds, we fine-tune the underlying empirical process and obtain a better convergence rate. The sample complexity required for our algorithm is $n \geq Cd$ (where C is a universal constant), which is information theoretically optimal ([YCS14]).
- Upon further polishing the analysis of AM iterates, we identify a regime where the rate of convergence is quadratic (with exponent 2), which is identical to Newton method for convex optimization (Theorem 3).
- Via numerical experiments (Section 5), we demonstrate the super-linear convergence of AM. Although we consider a mixture of 2 regression in theory, we show in simulations that more than 2 mixtures of linear regression also enjoys the super linear convergence. Also, we compare the performance of AM with a gradient based heuristic (Algorithm 2). We show that the rate of convergence of the gradient based heuristic is linear. Furthermore, we observe that on average the gradient based heuristic takes 8× iterations and 6× wall-clock time compared to AM.

Algorithm 1 AM for mixed linear regression

```
input Covariate response pairs $(x_i, y_i)_{i=1}^n$, initialization $(\widehat{\theta}_1^{(0)}, \widehat{\theta}_2^{(0)})$, number of rounds $T$
1: Split samples in $T$ disjoint groups $(x_i^{(t)}, y_i^{(t)})_{t=1}^{n/T}$, where $t = 0, 1, \ldots, T-1$
2: for $t = 0, \ldots, T-1$ do
3:     $z_i^{(t)} = \text{argmin}_{j \in \{0,1\}} |y_i^{(t)} - \langle x_i^{(t)}, \widehat{\theta}_j^{(t)} \rangle|$; $i \in [n/T]$
4:     $\widehat{\theta}_j^{(t+1)} = \text{argmin}_{\theta} (y_i^{(t)} - \langle x_i^{(t)}, \theta \rangle)^2 \mathbf{1}\left\{z_i^{(t)} = j\right\}$;
           for $j \in \{0, 1\}$
5: end for
output $(\widehat{\theta}_1^{(T)}, \widehat{\theta}_2^{(T)})$
```

1.1 Notations

We use $\|.\|$ to denote the ℓ_2 norm of a vector unless otherwise specified. $[n]$ denotes the set of integers $\{1, 2, \ldots, n\}$. Throughout the paper, we use $C, C_1, C_2, .., c, c_1, c_2..$ to represent positive universal constants, the value of which may change from instance to instance.

2 AM FOR THE MIXTURE OF REGRESSIONS

In this section, we describe the AM algorithm for parameter estimation where the observations are coming from (1). In particular, we are interested in solving the following least squares problem:

$$L(\theta_1, \theta_2) = \sum_{i=1}^{n} \min_{z_i \in \{0,1\}} \left(y_i - \langle x_i, z_i\theta_1 + (1 - z_i)\theta_2 \rangle \right)^2,$$

$$(\widehat{\theta}_1^{ls}, \widehat{\theta}_2^{ls}) = \text{argmin}_{\theta_1 \in \mathbb{R}^d, \theta_2 \in \mathbb{R}^d} \; L(\theta_1, \theta_2) \qquad (2)$$

where $\{x_i\}_{i=1}^n$ are the covariates and $\{z_i\}_{i=1}^n$ are the latent variables. Note that the above problem is non-convex owing to the presence of $\{z_i\}_{i=1}^n$. Furthermore, (2) is NP-hard for general covariates $\{x_i\}_{i=1}^n$ ([YCS14]). However the problem becomes tractable with structured covariates (e.g., i.i.d Gaussian covariate) and proper initialization. We use Alternating Minimization (AM) to solve the least squares problem of (2), the steps of which are described in Algorithm 1 (also Algorithm 1 of [YCS14]).

The first step of Algorithm 1 is sample-splitting across iterations. The sample split step is standard in the theoretical analysis of AM. For example, [YCS14] and [YCS16] use sample split for mixture of regressions, [NJS13] uses it for phase retrieval and [JNS13] uses it for matrix completion. As illustrated in Section 5, we do not require sample-split in experiments. This assumption is only for theoretical tractability. Also, since the AM converges in super-linear speed, the sample-split will only increase the sample complexity, n, by a multiplicative factor of $\log\log(1/\epsilon)$, where ϵ is the tolerable error (in ℓ_2 norm) in the recovery of (θ_1^*, θ_2^*). In the above-mentioned problems, sample split results in the increase of sample complexity by a multiplicative factor of $\log(1/\epsilon)$, which is much larger compared to the price we pay.

As seen in Algorithm 1, each iteration of AM consists of 2 steps. First, the labels $\{z_i\}_{i=1}^n$ are estimated. This is done by calculating which regressor estimate yields the linear model closer to the observation. Once the label ambiguity is resolved, the problem is now converted to 2 ordinary least squares. The solutions of the least squares yield the next iterate.

3 MAIN RESULTS

We now present the main results of the paper. We characterize the convergence rate of Algorithm 1. We make the following structural assumption on the covariates, which is standard and featured in several previous works ([YCS14, YCS16, NJS13, GPGR19]).

Assumption 1. *The covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d from the standard d-dimensional Gaussian distribution, $(x_i \overset{i.i.d}{\sim} \mathcal{N}(0, I_d))$.*

In this section, we also assume $w_i = 0$ for all $i \in [n]$. We emphasize here that in simulations (Section 5), we observe that our theory perfectly holds even in the presence of noise. However, for analysis we deal with the noiseless scenario only.

Let p_1 and p_2 be the fraction of observations coming from θ_1^* and θ_2^* respectively. We also define the following error metric to quantify the closeness of AM iterates at t-th iteration to the true parameters (θ_1^*, θ_2^*):

$$\mathsf{dist}(\theta_1^{(t)}, \theta_2^{(t)}) := \max\{\|\theta_1^{(t)} - \theta_1^*\|, \|\theta_2^{(t)} - \theta_2^*\|\}.$$

To simplify notation, we drop the superscript from $\theta_i^{(t)}$ for $i = 1, 2$ and consider one iteration of the AM algorithm with (θ_1, θ_2) as input and (θ_1^+, θ_2^+) as output. It is sufficient to show the one step contraction for Algorithm 1. Furthermore, let $n/T := n_1$, where n and T are the sample complexity and the number of iterations of Algorithm 1 respectively. Hence, the number of samples for this particular iteration is n_1. We have the following result.

Theorem 1. *Suppose $n_1 \geq Cd$, and the following*

$$\mathsf{dist}(\theta_1, \theta_2) \leq \frac{\|\theta_1^* - \theta_2^*\|}{2 \log n_1}$$

holds. Then, the inequality

$$\mathsf{dist}^2(\theta_1^+, \theta_2^+) \leq \left[\frac{\log n_1}{4\|\theta_1^* - \theta_2^*\|} \right] \mathsf{dist}(\theta_1, \theta_2)^3$$

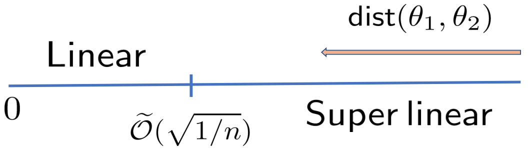

$\mathsf{dist}(\theta_1, \theta_2)$
Linear
0
$\widetilde{\mathcal{O}}(\sqrt{1/n})$
Super linear

Figure 1: Convergence region for iterates of AM algorithm. Super-linear region corresponds to convergence with exponent 3/2.

is satisfied with probability exceeding $1 - c_1 n_1^{-10}$ *provided* $\mathsf{dist}(\theta_1, \theta_2) \geq c_2 \max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}$.

The above theorem shows the super-linear rate of convergence of the dist function. Hence, for an error tolerance (in ℓ_2 norm) of ϵ, we have $T = \mathcal{O}(\log\log(1/\epsilon))$. Note that the super-linear rate holds as long as $\mathsf{dist}(\theta_1, \theta_2) \geq \widetilde{\mathcal{O}}(\sqrt{1/n})$. If $\mathsf{dist}(\theta_1, \theta_2)$ falls below the mentioned threshold, the rate of convergence is no longer super-linear; it falls back to the linear regime (Figure 1). This is because the concentration inequalities we use to prove Theorem 1 cease to produce any meaningful results if $\mathsf{dist}(\theta_1, \theta_2) < \widetilde{\mathcal{O}}(\sqrt{1/n})$.

The proof of Theorem 1 is deferred to Section 6. Super linear convergence is a consequence of two facts:

1. The fraction of non-zero (or *active*) terms that contribute in $\mathsf{dist}(\theta_1^+, \theta_2^+)$ is bounded by $\mathcal{O}(\mathsf{dist}^{1/2}(\theta_1, \theta_2))$.
2. Each active term is $\mathcal{O}(\mathsf{dist}(\theta_1, \theta_2))$.

In the prior works (e.g., [YCS14]), using crude singular value based technique, the fraction of active terms were bounded by $\mathcal{O}(1)$. Since each active terms contribute $\mathcal{O}(\mathsf{dist}(\theta_1, \theta_2))$, a linear rate is obtained. We instead use (sharp) empirical process tools to capture the *active* terms and improve the rate of convergence.

If the desired error tolerance (in ℓ_2 norm) is much less than $\widetilde{\mathcal{O}}(\sqrt{1/n})$, then Theorem 1 fails to characterize the entire behavior of the iterates. In order to so, we appeal to [YCS14, Theorem 1] which yields the following linear rate.

Theorem 2. *Suppose that* $n_1 \geq (C/\min\{p_1, p_2\})d$. *Then, provided* $\mathsf{dist}(\theta_1, \theta_2) \leq c\min\{p_1, p_2\}\|\theta_1^* - \theta_2^*\|$, *the inequality* $\mathsf{dist}(\theta_1^+, \theta_2^+) \leq \frac{1}{2}\mathsf{dist}(\theta_1, \theta_2)$ *holds with probability greater than* $1 - c\exp\{-c_1 d\}$.

3.1 Faster Rates: Quadratic Rate of Convergence

We now prove an improved rate of convergence for AM. In particular, we show that in a particular regime, the convergence rate of AM is quadratic (with exponent 2), which is an improvement over the rate in Theorem 1. We have the following result.

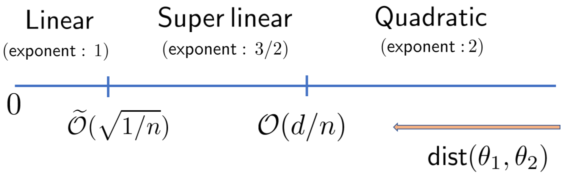


Figure 2: Convergence region for iterates of AM algorithm. Quadratic, Super-linear and linear region of convergence is shown.

Theorem 3. *Suppose that* $n_1 \geq (C/\min\{p_1, p_2\})d$, *and the following*

$$\mathsf{dist}(\theta_1, \theta_2) \leq c\min\{p_1, p_2\}\|\theta_1^* - \theta_2^*\|$$

holds. Then the inequality

$$\mathsf{dist}(\theta_1^+, \theta_2^+) \leq \frac{1}{2}\mathsf{dist}(\theta_1, \theta_2)^2$$

holds with probability exceeding $1 - c_1\exp\{-c_2 d\} - c_3 n_1^{-10}$ *provided*
$\mathsf{dist} \geq \max\left\{C\max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}, \frac{d}{C_1 n_1}\right\}$.

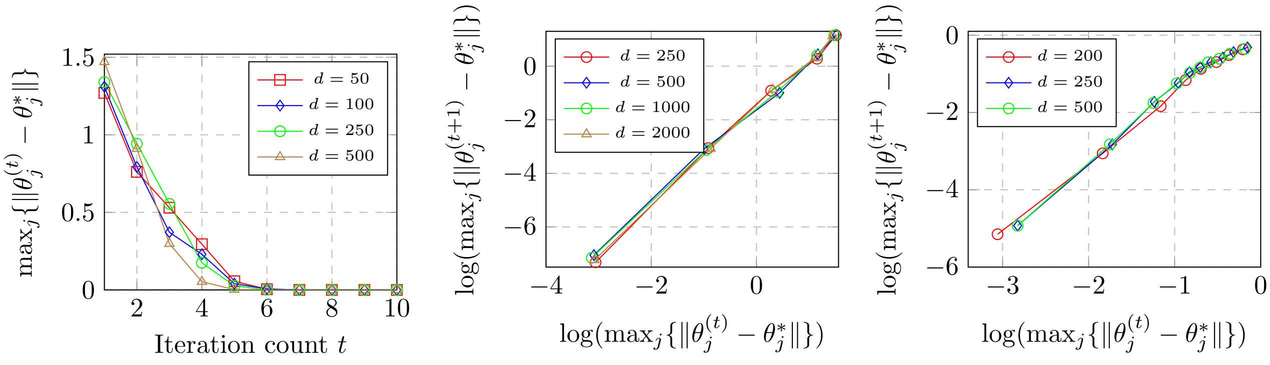
The proof is deferred to the Supplementary material. The gain in rate comes from a careful analysis of the spectrum of a Gaussian random matrix in conjuction with the fact that the fraction of *active* terms is $\mathcal{O}(\mathsf{dist}^{1/2}(\theta_1, \theta_2))$.

Combining Theorem 1 and 3, we are now able to completely characterize the convergence behavior of the iterates of the AM algorithm. It is shown in Figure 2. Until $\mathcal{O}(d/n)$, the convergence is quadratic. Beyond this point the rate slows down but maintains a super-linear rate upto $\widetilde{\mathcal{O}}(\sqrt{1/n})$. After this point, the convergence speed slows down to a linear rate.

4 INITIALIZATION

As seen in Theorem 1,2 and 3, the convergence guarantee of AM requires the initial values of the iterate to be close to the optimal parameters. In particular, we need the initial values to be within a norm ball of constant radius of the optimal parameters.

Usually, spectral methods are employed for initializing AM for a large class of problems. Since we have the covariate-response pairs $(x_i, y_i)_{i=1}^n$, we can compute an appropriate matrix, and the singular value decomposition (SVD) of the matrix yields required initialization. In [CSV13], [NJS13], [Wal18], the spectral method of initialization is used for the phase retrieval problem, and in [GPGR19], it is used for the max-affine regression problem.



(a) (b) (c)

Figure 3: Convergence of the AM with Gaussian covariates—in panel (a), for a mixture of 2 linear regression, we plot the distance to the true parameters $\max_{j\in 1,2}\{\|\theta_j^{(t)} - \theta_1^*\|\}$ over iterations t for different d $(50, 100, 250, 500)$, where we set $n = 6\,d$. Panel (b) shows the super-linear convergence of AM (with exponent $1.7 - 1.8$) for mixed regression with 2 components. Here we choose $d = 250, 500, 1000, 2000$ and $n = 6\,d$. In Panel (c), we consider the problem of mixed linear regression with 3 components. Keeping $n = 15\,d$ and varying $d(200, 250, 500)$, we show that AM retains the super-linear convergence for more than 2 mixtures of linear regression. All the points in the plots are obtained via taking average over 20 trials.

In Algorithm 2 of Yi et al (2014) ([YCS14]), a spectral method for the mixture of 2-component mixture is provided. The algorithm first constructs a matrix $M = \sum_{i=1}^n y_i^2 x_i x_i^\top$. Then, using SVD, the subspace spanned by the eigenvectors corresponding to the top 2 eigenvalues is obtained. It is shown in [YCS14] that the optimal parameters (θ_1^*, θ_2^*) lie in this subspace. Finally, *griding* the 2-dimensional subspace yields in the required initial values of the iterates of AM.

For the mixture of regression with more than 2 components, Yi et al (2016) ([YCS16]) uses a tensor decomposition technique. After obtaining the subspace from appropriate matrix, a tensor is constructed from the covariate-response pairs $(x_i, y_i)_{i=1}^n$. Decomposing the tensor results in the required initialization.

In this paper, we use a slightly stricter initialization (by a log factor) of [YCS14] to get the conditions of Theorem 1,2 and 3. When $n/(\log n)^2 \sim d(\log d)^2$, we get this initialization by using [YCS14, Proposition 2]. Since we work with 2 components, employing a tensor decomposition method of [YCS16] is unnecessary because griding a 2-dimensional subspace requires very light computation.

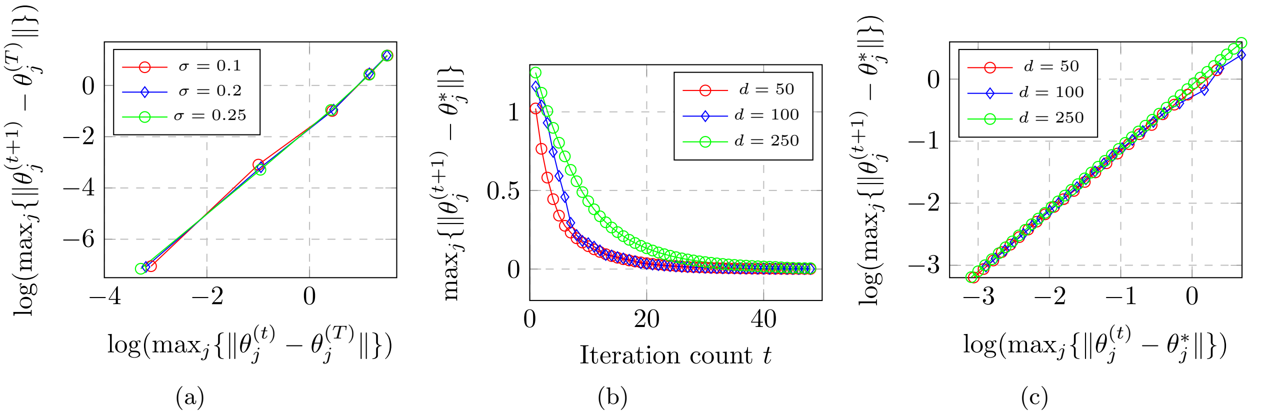
5 SIMULATIONS

In this section, we validate the theory presented in Section 3. Additionally, we handle the setting where the observations, $\{y_i\}_{i=1}^n$ are corrupted with additive noise. Furthermore, we consider the problem of mixed linear regression with more than 2 components. Note that as mentioned in Section 3, in all our experiments, we do not require the sample-split step of Algorithm 1. Finally, we compare the performance of AM with a gradient based heuristic.

For the following experiments, we sample the covariates $\{x_i\}_{i=1}^n$ in an i.i.d fashion from the standard d-dimensional Gaussian distribution. We randomly select the d-dimensional true parameters $\{\theta_1^*, \theta_2^*\}$. The observations are then obtained via equation (1).

Convergence of AM for 2 mixture: We first show the convergence of Algorithm 1 for a mixture of 2 linear regression in the noiseless setting. The results are shown in Figure 3. In panel (a) of Figure 3, we plot $\max\{\|\theta_1^{(t)} - \theta_1^*\|, \|\theta_1^{(t)} - \theta_2^*\|\}$ with respect to iterations of Algorithm 1, where $\{\theta_1^{(t)}, \theta_2^{(t)}\}$ is the output at the t-th iterate of the algorithm. We consider different values of d (specifically $50, 100, 250$ and 500) and choose the sample size $n = 6\,d$. We observe that the iterates converge to 0 very quickly, and hence Algorithm 1 guarantees perfect recovery of $\{\theta_1^*, \theta_2^*\}$. In fact, we see that AM takes at most 6 steps to converge.

Super linear convergence for 2 mixture: In Figure 1 (b), we characterize the rate of convergence of Algorithm 1. Here, we plot $\log\left(\max\{\|\theta_1^{(t+1)} - \theta_1^*\|, \|\theta_2^{(t+1)} - \theta_2^*\|\}\right)$ with respect to $\log\left(\max\{\|\theta_1^{(t)} - \theta_1^*\|, \|\theta_2^{(t)} - \theta_2^*\|\}\right)$. Note that the slope of this plot quantifies the rate of convergence of the underlying iterative algorithm that produces the iterates $\{\theta_1^{(j)}, \theta_2^{(j)}\}$ for $j = 0, 1, \ldots$. Algorithms with



(a)

(b)

(c)

Figure 4: Noisy AM and comparison with gradient based heuristic—in panel (a), we plot the (log) optimization error $\log(\max_j\{\|\theta_j^{(t)} - \theta_j^{(T)}\|\})$ and show that AM retain the super-linear speed of convergence even in the presence of noise. Here we fix $d = 250$ and $n = 6\,d$ and vary over σ $(0.1, 0.2, 0.25)$. In Panel (b) we show that gradient based heuristic (Algorithm 2 indeed recovers the true parameters. We fix $n = 6\,d$ and vary $d (= 50, 100, 200)$. We also tune the step-size to fasten Algorithm 2. In Panel (c) we demonstrate that the rate of convergence of Algorithm 2 is truly linear. We plot $\log(\max_j\{\|\theta_j^{(t+1)} - \theta_j^*\|\})$ with respect to $\log(\max_j\{\|\theta_j^{(t)} - \theta_j^*\|\})$ and the slope of the line is close to 1 implying linear rate of convergence.

linear rate of convergence will result in slope 1. Any slope strictly greater than 1 implies super-linear convergence.

We set $d = 250, 500, 1000$ and 2000 and take $n = 6\,d$. The results are shown in Figure 3 (b). We observe that the slope of the line is around $1.7 - 1.8$, which validates our theory of super-linear convergence. In Theorem 1 and 3, we prove that the exponent of convergence is 1.5 and 2 under different settings. However, in the simulations, we observe the exponent of (super-linear) convergence is a constant between 1.5 and 2.

Super linear convergence for more than 2 mixtures: We now show empirically that our theory of super-linear convergence holds for mixture of more than 2 linear regression. For this setting we use an extension of Algorithm 1 tailored to more than 2 mixtures. This is precisely Algorithm 3 of Yi et al (2016) ([YCS16]). In Figure 3 (c), we consider a mixture of 3 linear regression. We consider $d = 200, 250$ and 500 and take $n = 15\,d$. Similar to the 2 mixture setting, we are interested in the rate of convergence and hence we plot $\log\left(\max_{j\in\{1,2,3\}}\{\|\theta_j^{(t+1)} - \theta_j^*\|\}\right)$ with respect to $\log\left(\max_{j\in\{1,2,3\}}\{\|\theta_j^{(t)} - \theta_j^*\|\}\right)$. From Figure 3 (c), we see that the plot is linear with slope around $1.7 - 1.85$. Hence, AM retains the super linear speed of convergence for an arbitrary number of mixtures of linear regression.

Algorithm 1 with noisy observations: We now consider the setting where $w_i \neq 0$ in equation (1). In particular we assume $w_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$. Under this setting, we can only recover $\{\theta_1^*, \theta_2^*\}$ upto an error floor depending on (σ, n, d). To demonstrate super-linear convergence in this setting, we first estimate the error floor by letting Algorithm 1 run for $T = 50$ iterations. We now construct the *optimization-error* $\max_{j\in\{1,2\}}\{\|\theta_j^{(t)} - \theta_j^{(T)}\|\}$. In Figure 4 (a), we plot the variation of the (logarithm of) *optimization-error* in the $t+1$-th iteration with respect to t-th iteration. We notice that the dependence is linear (for different values of σ $(= 0.1, 0.2, 0.25)$) with a slope of $1.8 - 1.85$. This observation validates the fact that Algorithm 1 retains the super-linear convergence in the presence of noise.

Comparison with gradient based heuristic: Finally, we compare Algorithm 1 with a gradient based heuristic. The heuristic algorithm is described in Algorithm 2.

In Algorithm 2, we simply replace the least-squares step by a gradient step. We study this algorithm empirically in the noiseless setting for a mixture of 2 linear regressions. In Figure 4 (b), we show that Algorithm 2 indeed converges. We set $d = 50, 100, 250$ and take $n = 6\,d$. Furthermore, we increase the step size γ until the algorithm starts oscillating and choose the largest stepsize for which Algorithm 2 converges. We observe that even with the step-size tuning, Algorithm 2 takes a lot of iterations compared to Algorithm 1 (Figure 3 (a)) in identical settings.

We now characterize the rate of convergence of Algo-

Algorithm 2 Gradient based heuristic for mixed linear regression

```
input Covariate response pairs $(x_i, y_i)_{i=1}^n$, initialization $(\widehat{\theta}_1^{(0)}, \widehat{\theta}_2^{(0)})$, step size $\gamma$, number of rounds $T$
for $t = 0, \ldots, T-1$ do
    $z_i^{(t)} = \mathrm{argmin}_{j\in\{0,1\}} |y_i - \langle x_i, \widehat{\theta}_j^{(t)}\rangle|$; $i \in [n]$
    $\widehat{\theta}_j^{(t+1)} = \widehat{\theta}_j^{(t)} - \gamma\nabla\left((y_i - \langle x_i, \theta\rangle)^2 \mathbf{1}\left\{z_i^{(t)} = j\right\}\right)$;
        for $j \in \{0, 1\}$
end for
output $(\widehat{\theta}_1^{(T)}, \widehat{\theta}_2^{(T)})$
```

Table 1: Comparison between AM (Algorithm 1) and GD based heuristic (Algorithm 2) in terms of iteration complexity and wall-clock time.

DIMENSION (d)	ALGORITHM	ITERATION	WALL-CLOCK (SEC)
50	AM	5	0.015
50	GD	45	0.125
100	AM	5	0.094
100	GD	47	0.659
250	AM	6	0.953
250	GD	48	6.546

rithm 2. In order to do so, similar to the previous scenarios, we plot $\log\left(\max_{j\in 1,2}\{\|\theta_j^{(t+1)} - \theta_j^*\|\}\right)$ with respect to $\log\left(\max_{j\in 1,2}\{\|\theta_j^{(t)} - \theta_j^*\|\}\right)$. This is shown in Figure 4 (c). We observe that the dependence is linear with slope roughly equal to 1. This shows that the rate of convergence of the gradient based heuristic is truly linear.

Finally, we compare the iteration complexity and wall-clock time of AM (Algorithm 1) with the gradient based heuristic (Algorithm 2). The results are tabulated in Table 1. Here, we fix a target precision of 0.001. We observe that compared to AM, Algorithm 2 takes 9× iterations to recover the true parameter within the given precision. We also compare the algorithms in terms of wall-clock time. We found that on average gradient based heuristic takes 6× time compared to AM, even when we tune and choose the largest step-size. These results imply that AM is a much faster algorithm compared to the gradient based heuristic in terms of both number of iterations and total wall-clock time.

6 Proof of Theorem 1

We now prove Theorem 1. To simplify notation, we use the shorthand $\mathsf{dist} := \mathsf{dist}(\theta_1, \theta_2)$ and $\mathsf{dist}^+ := \mathsf{dist}(\theta_1^+, \theta_2^+)$. Also, we drop the superscript in $x_i^{(t)}$ and $y_i^{(t)}$. Recall that the sample complexity for this iteration is $n/T = n_1$.

Here we retain the notation of [YCS14]. To that end, let us denote the set of indices J_1^* and J_2^* corresponding to observations coming from θ_1^* and θ_2^* respectively. Similarly, we define J_1 and J_2 corresponding to the iterate of AM for θ_1 and θ_2 respectively. Hence, we have

$$J_1^* = \{i \in [n_1] : y_i = \langle x_i, \theta_1^*\rangle\}$$

and similarly, using the criteria of Algorithm 1, we have

$$J_1 = \{i \in [n_1] : (y_i - \langle x_i, \theta_1\rangle)^2 < (y_i - \langle x_i, \theta_2\rangle)^2\}.$$

We can define J_2^* and J_2 in a similar way. We also define a diagonal matrix $W \in \mathbb{R}^{n_1 \times n_1}$ such that $W_{ii} = 1$ if $i \in J_1$ and 0 if $i \in J_2$. Similarly, we define W^* such that $W_{ii}^* = 1$ if $i \in J_1^*$ and 0 otherwise. With this new notation, it immediately follows that θ_1^+ is the least squares solution to $Wy = WX\theta$, and hence

$$\theta_1^+ = (X^\top W X)^{-1} X^\top W y$$

where the n_1 dimensional vector $y = [y_1 \ldots y_{n_1}]^\top$ and we use the fact that $W^2 = W$. Similarly we observe that θ_2^+ is the least squares solution of $(I - W)y = (I - W)X\theta$.

With this, the observation vector y can be written as

$$y = W^* X \theta_1^* + (I - W^*) X \theta_2^*,$$

and hence substituting y in the closed form expression for θ_1^+, we obtain

$$\theta_1^+ - \theta_1^* = (X^\top W X)^{-1} X^\top (WW^* - W) X (\theta_1^* - \theta_2^*).$$

Let $\mathcal{S} = J_1 \cap J_2^*$. We have

$$\begin{aligned}
&\|X(\theta_1^+ - \theta_1^*)\|^2 \\
&= \|X(X^\top W X)^{-1} X^\top (WW^* - W) X (\theta_1^* - \theta_2^*)\|^2 \qquad (3) \\
&= \|X_{J_1} X_{J_1}^\dagger (WW^* - W) X (\theta_1^* - \theta_2^*)\|^2 \qquad (4) \\
&\leq \|X_{\mathcal{S}}(\theta_1^* - \theta_2^*)\|^2 \qquad (5) \\
&= \sum_{i\in\mathcal{S}} \langle x_i, \theta_1^* - \theta_2^*\rangle^2
\end{aligned}$$

where equation (3) follows from substituting $\theta_1^+ - \theta_1^*$, equation (4) follows from the definition of J_1. Equation (5) follows from the facts that equation 4 is non-zero only when $i \in \mathcal{S}$ and $X_{J_1} X_{J_1}^\dagger$ is a projection matrix and hence non-expansive. Continuing, we obtain

$$\begin{aligned}\|X(\theta_1^+ - \theta_1^*)\|^2 &\leq \sum_{i\in\mathcal{S}} \langle x_i, \theta_1^* - \theta_2^*\rangle^2 \\ &= \sum_{i\in\mathcal{S}} \langle x_i, \theta_1^* - \theta_1 + \theta_1 - \theta_2^*\rangle^2 \\ &\leq \sum_{i\in\mathcal{S}} \left(2\langle x_i, \theta_1^* - \theta_1\rangle^2 + 2\langle x_i, \theta_2^* - \theta_1\rangle^2\right)\end{aligned}$$

Now, recall that if $i \in \mathcal{S}$, $y_i = \langle x_i, \theta_2^*\rangle$. Also, we have

$$\begin{aligned}&(y_i - \langle x_i, \theta_1\rangle)^2 < (y_i - \langle x_i, \theta_2\rangle)^2 \\ &\Rightarrow \langle x_i, \theta_2^* - \theta_1\rangle^2 < \langle x_i, \theta_2^* - \theta_2\rangle^2.\end{aligned}$$

Substituting the above, we have

$$\begin{aligned}\|X(\theta_1^+ - \theta_1^*)\|^2 \leq\ & 2\sum_{i\in\mathcal{S}} \langle x_i, \theta_1^* - \theta_1\rangle^2 \\ &+ 2\sum_{i\in\mathcal{S}} \langle x_i, \theta_2^* - \theta_2\rangle^2 \qquad (6)\end{aligned}$$

We now concentrate on the first term of the right hand side of equation (6). Recall that from the definition of dist, we have $\|\theta_1 - \theta_1^*\|^2 \leq \mathsf{dist}^2$. Hence, we obtain

$$\sum_{i\in\mathcal{S}} \langle x_i, \theta_1^* - \theta_1\rangle^2 \leq (\mathsf{dist}^2) \left(\sum_{i\in\mathcal{S}} \langle x_i, \frac{\theta_1 - \theta^*}{\|\theta_1 - \theta^*\|}\rangle^2\right)$$

Let us define the unit vector $u = \frac{\theta_1 - \theta^*}{\|\theta_1 - \theta^*\|}$. Since we re-sample at each iteration of the AM algorithm, x_i is independent of u. Furthermore, conditioned on the fact that $i \in \mathcal{S}$, the distribution of x_i is no longer Gaussian. To this end, [YCS16, Lemma 15(b)] shows that the distribution of x_i is c-sub-Gaussian, implying that $\langle x_i, u\rangle$ is a c sub-Gaussian random variable (here c is a constant) with $\mathbb{E}[\langle x_i, u\rangle^2] \leq C$, where C is a constant. Moreover, using [Ver18, Lemma 2.7.6], the distribution of $\langle x_i, u\rangle^2$ is $(\tilde{c_1}, \tilde{c_2})$ sub-exponential, where $\tilde{c_1}$ and $\tilde{c_2}$ are constants. Similar argument holds for the second term in the right hand side of equation (6).

We now use the following Lemma which gives a high probability upper-bound on $|\mathcal{S}|$.

Lemma 1. *We have*

$$|\mathcal{S}| \leq C_1 \frac{p_2 n_1\, \mathsf{dist}}{\|\theta_1^* - \theta_2^*\|}$$

with probability exceeding $1 - n_1^{-10}$ *provided* $\mathsf{dist} \geq c\max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}$.

We first take this lemma for granted and conclude the proof of the theorem. Let $k_0 := C_1 \frac{p_2 \mathsf{dist}\, n_1}{\|\theta_1^* - \theta_2^*\|}$. We have

$$\mathbb{P}\left(\sum_{i=1}^{|\mathcal{S}|} \langle x_i, u\rangle^2 \geq t\right) \leq \mathbb{P}(\sum_{i=1}^{k_0} \langle x_i, u\rangle^2 \geq t) + \mathbb{P}(|\mathcal{S}| \geq k_0)$$

Provided $\mathsf{dist} \geq c\max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}$ and choosing $t = \frac{3k_0 \log n_1}{2}$, sub-exponential concentration ([Wai19, Chapter 2]) along with Lemma 1 yields

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^{|\mathcal{S}|} \langle x_i, u\rangle^2 \geq t) &\leq \mathbb{P}(\sum_{i=1}^{k_0} \langle x_i, u\rangle^2 \geq \frac{3k_0 \log n_1}{2}) + c_1 n_1^{-10} \\ &\leq c n_1^{-10} + c_1 n_1^{-10}.\end{aligned}$$

Similar expression holds for the second term of equation (6). Substituting this in equation (6), we obtain

$$\|X(\theta_1^+ - \theta_1^*)\|^2 \leq 6C_1 (\mathsf{dist})^3 \frac{p_2 n_1 \log n_1}{\|\theta_1^* - \theta_2^*\|}. \qquad (7)$$

We now convert this prediction error to estimation error via exploiting the spectral properties of the Gaussian random matrix X. We have

$$\|X(\theta_1^+ - \theta_1^*)\|^2 \geq \lambda_{\min}(X^\top X) \|\theta_1^+ - \theta_1^*\|^2.$$

Since $X \in \mathbb{R}^{n\times d}$ is a Gaussian random matrix, using [Ver10], the minimum singular value is

$$\sigma_{\min}(X) \geq \sqrt{n_1} - \sqrt{d} - t$$

with probability exceeding $1 - \exp\{-ct^2\}$.

Using the fact that $n_1 \geq Cd$, and substituting $t = \sqrt{n_1}/\sqrt{C}$, we obtain

$$\lambda_{\min}(X^\top X) \geq c\, n_1$$

with probability exceeding $1 - \exp\{-c_1\, n_1\}$. Putting everything together, we have

$$\|\theta_1^+ - \theta_1^*\|^2 \leq \frac{6C_1 p_2 \log n_1}{c\|\theta_1^* - \theta_2^*\|} \mathsf{dist}^3 \leq \frac{1}{4} \frac{\log n_1}{\|\theta_1^* - \theta_2^*\|} \mathsf{dist}^3$$

where we use the fact that $p_2 \leq 1$, and choose appropriate constants c and C_1.

Similarly we prove an upper bound for $\|\theta_2^+ - \theta_2^*\|^2$, and hence the theorem follows.

7 CONCLUSION

We prove the super linear convergence of AM for noiseless mixture of 2 linear regressions. However, in experiments, we see that the super linear rate retains for noisy setting and even for more than 2 mixtures. Providing theoretical guarantees in these settings will be our immediate future works. In experiments we also observe that sample-split is unnecessary and the exponent of convergence is around $1.75 - 1.8$. Is the exponent really 1.5, or a finer analysis can improve this? We leave this questions as our future endeavors.

References

[BWY17] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[CB18] Sheng Chen and Arindam Banerjee. An improved analysis of alternating minimization for structured multi-response regression. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6617–6628, USA, 2018. Curran Associates Inc.

[CL13] Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pages 1040–1048, 2013.

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[CXW$^+$18] Jinghui Chen, Pan Xu, Lingxiao Wang, Jian Ma, and Quanquan Gu. Covariate adjusted precision matrix estimation via nonconvex optimization. In *International Conference on Machine Learning*, pages 922–931, 2018.

[CYC14] Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 560–604, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[DH00] Partha Deb and Ann M. Holmes. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics*, 9(6):475–489, 2000.

[DTZ17] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *30th Annual Conference on Learning Theory*, 2017.

[GL$^+$07] Bettina Grün, Friedrich Leisch, et al. Applications of finite mixtures of regression models. *URL: http://cran. r-project. org/web/packages/flexmix/vignettes/regression-examples. pdf*, 2007.

[GPGR19] Avishek Ghosh, Ashwin Pananjady, Aditya Guntuboyina, and Kannan Ramchandran. Max-affine regression: Provable, tractable, and near-optimal statistical estimation. *arXiv preprint arXiv:1906.09255*, 2019.

[JJNH91] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[KC19] Jeongyeol Kwon and Constantine Caramanis. EM converges for a mixture of many linear regressions. *CoRR*, abs/1905.12106, 2019.

[KYB19] Jason M. Klusowski, Dana Yang, and W. D. Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Trans. Inf. Theor.*, 65(6):3515–3524, June 2019.

[LL18] Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. *arXiv preprint arXiv:1802.07895*, 2018.

[MKV18] Ashok Vardhan Makkuva, Sreeram Kannan, and Pramod Viswanath. Globally consistent algorithms for mixture of experts. *CoRR*, abs/1802.07417, 2018.

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[SJA16] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231, 2016.

[SS19] Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. *arXiv preprint arXiv:1902.03653*, 2019.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[VT02] Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[Wal18] Irène Waldspurger. Phase retrieval with random Gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, May 2018.

[WD95] Michel Wedel and Wayne S DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–55, 1995.

[WGNL15] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional em algorithm: Statistical optimization and asymptotic normality. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2521–2529. Curran Associates, Inc., 2015.

[WK12] Michel Wedel and Wagner A Kamakura. *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media, 2012.

[XJ96] Lei Xu and Michael I Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

[YCS14] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

[YCS16] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *CoRR*, abs/1608.05749, 2016.

[ZJD16] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2190–2198. Curran Associates, Inc., 2016.

[ZL16] Huishuai Zhang and Yingbin Liang. Reshaped wirtinger flow for solving quadratic system of equations. In *Advances in Neural Information Processing Systems*, pages 2622–2630, 2016.

[ZWZG17] Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, and Quanquan Gu. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4180–4188. JMLR. org, 2017.

8 SUPPLEMENTARY MATERIAL

8.1 Proof of Lemma 1:

We first upper bound the quantity $\mathbb{E}|\mathcal{S}|$. Then invoking Hoeffding's inequality, we obtain a high probability bound on $|\mathcal{S}|$. We have

$$\begin{aligned}\mathbb{E}|\mathcal{S}| &= \mathbb{E}\sum_{i=1}^{n_1} \mathbf{1}\left\{i \in J_1 \cap J_2^*\right\} \\ &= \sum_{i=1}^{n_1} \mathbb{P}(i \in J_2^*)\mathbb{P}(i \in J_1 | i \in J_2^*)\end{aligned}$$

The event $i \in J_1$, conditioned on the event $i \in J_2^*$ is equivalent to the event

$$\langle x_i, \theta_1 - \theta_2^* \rangle^2 < \langle x_i, \theta_2 - \theta_2^* \rangle^2.$$

Hence, we have

$$\begin{aligned}\mathbb{E}(\mathbf{1}\left\{i \in J_1 \cap J_2^*\right\}) =& \mathbb{P}(i \in J_2^*)\times \\ & \mathbb{P}\left(\langle x_i, \theta_1 - \theta_2^* \rangle^2 < \langle x_i, \theta_2 - \theta_2^* \rangle^2\right).\end{aligned}$$

Now, from the initialization condition, we have

$$\mathsf{dist} \leq \frac{\|\theta_1^* - \theta_2^*\|}{2},$$

and hence $\|\theta_1 - \theta_2^*\| > \|\theta_2 - \theta_2^*\|$. Using this fact and invoking Lemma 1 of [YCS14] yields

$$\mathbb{P}\left(\langle x_i, \theta_1 - \theta_2^* \rangle^2 < \langle x_i, \theta_2 - \theta_2^* \rangle^2\right) \leq \frac{\|\theta_2 - \theta_2^*\|}{\|\theta_1 - \theta_2^*\|}.$$

Also, we have

$$\begin{aligned}\|\theta_1 - \theta_2^*\| &= \|\theta_1 - \theta_1^* + \theta_1^* - \theta_2^*\| \\ &\geq \|\theta_1^* - \theta_2^*\| - \|\theta_1 - \theta_1^*\| \geq \frac{1}{2}\|\theta_1^* - \theta_2^*\|\end{aligned}$$

where the last inequality follows from the initialization condition. Substituting this, we obtain

$$\begin{aligned}\mathbb{P}\left(\langle x_i, \theta_1 - \theta_2^* \rangle^2 < \langle x_i, \theta_2 - \theta_2^* \rangle^2\right) &\leq \frac{2\|\theta_2 - \theta_2^*\|}{\|\theta_1^* - \theta_2^*\|} \\ &\leq \frac{2 \;\mathsf{dist}}{\|\theta_1^* - \theta_2^*\|}.\end{aligned}$$

So, we finally have

$$\mathbb{E}|\mathcal{S}| \leq \frac{2 p_2 n_1 \;\mathsf{dist}}{\|\theta_1^* - \theta_2^*\|}.$$

We now invoke the Hoeffding's inequality to obtain

$$\mathbb{P}\bigg(||\mathcal{S}| - \mathbb{E}|\mathcal{S}|| > n_1 t\bigg) \leq 2\exp\{-c n_1 t^2\}$$

Substituting $t = C\frac{p_2 \,\mathsf{dist}}{\|\theta_1^* - \theta_2^*\|}$, we obtain

$$|\mathcal{S}| \leq C_1 \frac{p_2 n_1 \;\mathsf{dist}}{\|\theta_1^* - \theta_2^*\|}$$

with probability exceeding $1 - c_1 n_1^{-10}$ as long as $\mathsf{dist} \geq C \max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}$.

8.2 Proof of Theorem 3

We follow the same setup and notations as in the proof of Theorem 1. This proof borrows a few technical details from the proof of [YCS14, Theorem 1]. Recall that we have

$$\theta_1^+ - \theta_1^* = (X^\top W X)^{-1} X^\top (W W^* - W) X (\theta_1^* - \theta_2^*).$$

Hence, we obtain $\|\theta_1^+ - \theta_1^*\| \leq AB$, where

$$A = \|(X^\top W X)^{-1}\| \quad \text{and,}$$
$$B = \|X^\top (W W^* - W) X (\theta_1^* - \theta_2^*)\|.$$

Bounding A: We just use the bound in [YCS14, Theorem 1] to control A, which yields the following. If $n_1 \geq Cd$, we have

$$A \leq \frac{C}{n_1}$$

with probability exceeding $1 - c\exp\{-c_1 d\}$.

Bounding B: Let

$$Q = X^\top (W W^* - W) X.$$

Using the proof steps of [YCS14, Theorem 1], we obtain

$$B \leq 2\sigma_{\max}(Q)\, \mathsf{dist}.$$

Note that the Q is non-zero when $i \in J_1 \cap J_2^*$. Using Lemma 1 of [YCS14] in conjunction with [Ver10] yields

$$\sigma_{\max}(Q) \leq c \max\left(d, |J_1 \cap J_2^*|\right).$$

In the proof of Lemma 1, we observe that

$$\mathbb{E}|J_1 \cap J_2^*| \leq \frac{2p_2 n_1\, \mathsf{dist}}{\|\theta_1^* - \theta_2^*\|} \leq C n_1\, \mathsf{dist}.$$

Hence applying Hoeffding's inequality yields

$$|J_1 \cap J_2^*| \geq C_1 n_1\, \mathsf{dist}$$

with probability exceeding $1 - c n_1^{-10}$ provided $\mathsf{dist} \geq C \max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}$.

Now, additionally if $\mathsf{dist} \geq (\frac{1}{C_1})\frac{d}{n}$, we obtain

$$|J_1 \cap J_2^*| \geq d.$$

With this, we have

$$B \leq c|J_1 \cap J_2^*|\, \mathsf{dist} \leq c_1 n\, \mathsf{dist}^2.$$

where we use Lemma 1 once again in the last step. Putting everything together, we obtain

$$\|\theta_1^+ - \theta_1^*\| \leq \frac{C}{n_1} c_1 n_1\, \mathsf{dist}^2 \leq \frac{1}{2}\, \mathsf{dist}^2,$$

provided

$$\mathsf{dist} \geq \max\left\{ C \max\{p_1, p_2\}\sqrt{\frac{\log n_1}{n_1}}, \frac{d}{C_1 n_1} \right\}.$$

The constants are chosen such that $C\, c_1 \leq 1/2$. Similarly we show an equivalent upper bound on $\|\theta_2^+ - \theta_2^*\|$, thus yielding the theorem.

Note that, the condition $\mathsf{dist} \geq \frac{d}{C_1 n_1}$ on the dist function is quite strong since we are interested in the information theoretic minimum sample regime with $n_1 = \mathcal{O}(d)$.