CTR D-STOP

Technical Report 145

Global Convergence of EM Algorithm for Mixtures of Two Component Linear Regression

Research Supervisor
Constantine Caramanis
Wireless Networking & Communications Group

Project Title: Clustering and Classification

October 2018

Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin

CTR CENTER FOR TRANSPORTATION RESEARCH

WNCG Wireless Networking & Communications Group

D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

1. Report No. D-STOP/2018/145	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Global Convergence of EM Algorithm for Mixtures of Two Component Linear Regression		5. Report Date October 2018	
		6. Performing Organization Code	
7. Author(s) Jeongyeol Kwon Constantine Caramanis		8. Performing Organization Report No. Report 145	
9. Performing Organization Name and Address Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 3925 W. Braker Lane, 4th Floor Austin, Texas 78759		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No. DTRT13-G-UTC58	
12. Sponsoring Agency Name and Address United States Department of Transportation University Transportation Centers 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered	
		14. Sponsoring Agency Code	
15. Supplementary Notes Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program. Project Title: Clustering and Classification			
16. Abstract The Expectation-Maximization algorithm is perhaps the most broadly used algorithm for inference of latent variable problems. A theoretical understanding of its performance, however, largely remains lacking. Recent results established that EM enjoys global convergence for Gaussian Mixture Models. For Mixed Regression, however, only local convergence results have been established, and those only for the high SNR regime. We show here that EM converges for mixed linear regression with two components (it is known not to converge for three or more), and moreover that this convergence holds for random initialization.			
17. Key Words		18. Distribution Statement No restrictions. This document is available to the public through NTIS (http://www.ntis.gov): National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161	
19. Security Classif.(of this report) Unclassified	20. Security Classif.(of this page) Unclassified	21. No. of Pages	22. Price

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers.

Global Convergence of EM Algorithm for Mixtures of Two Component Linear Regression

Jeongyeol Kwon
kwonchungli@utexas.edu

Constantine Caramanis
constantine@utexas.edu

The University of Texas at Austin

October 23, 2018

Abstract

The Expectation-Maximization algorithm is perhaps the most broadly used algorithm for inference of latent variable problems. A theoretical understanding of its performance, however, largely remains lacking. Recent results established that EM enjoys global convergence for Gaussian Mixture Models. For Mixed Regression, however, only local convergence results have been established, and those only for the high SNR regime. We show here that EM converges for mixed linear regression with two components (it is known not to converge for three or more), and moreover that this convergence holds for random initialization.

1 Introduction

The expectation-maximization (EM) algorithm is widely used for inference in the presence of missing data, often modeled as latent variables. It is a general-purpose technique for computing the maximum likelihood solution for such problems [1, 2]. In general, solving the max-likelihood problem in the presence of missing data is an intractable (NP-hard) problem due to the non-convexity of the likelihood function. EM is an iterative procedure that computes successively tighter lower bounds on the (typically non-convex) likelihood function. Its appeal stems from its success as observed in a broad array of problems in practice, and its computational simplicity: it is essentially no more complex than solving the ML problem in the setting of no missing variables. Despite its simplicity and widespread use, very little is understood about EM. Recent results have demonstrated that in the high SNR regime (and under some additional regularity assumptions), EM converges locally (e.g., [3, 4, 5, 6, 7]). For the special case of Gaussian Mixtures with two components, very recent work [8] has demonstrated that a two-phase version of EM converges from a random initialization point. As far as we are aware, no comparable result is known for mixed linear regression – in other words, prior to this work, no results guarantee global convergence of EM. Nevertheless, the EM algorithm is widely used to solve mixed linear regression in practice (e.g., see [9, 10]).

We show that EM for mixed linear regression (MLR) with two components converges globally. There is no need for any special initialization. Moreover, our proof reveals (a bound on) the rate of convergence of EM as a function of how far it is from the optimal solution. Locally, we recover and expand on past results (e.g., [6, 4]), as these not only required an initialization step, but only demonstrated that EM converges in the high SNR regime. We explain the connections to prior art in more detail in Section 1.2.

1.1 Basic Setup and the EM Algorithm

Mixed linear regression (MLR) models the setting where different subsets of the response variables are generated by different regressors. In the case of two components, which we consider here, data $(\boldsymbol{x}_i, y_i) \in R^d \times R$ are generated by:

$$y_i = \boldsymbol{\beta}^*_{z_i} \boldsymbol{x}_i + e_i, \qquad i = 1, ..., n,$$

where z_i are the hidden or latent variables that take values $z_i \in \{1, 2\}$, and thus play the role of labels, denoting that a sample $(\boldsymbol{x}_i, y_i)$ satisfies a (noisy) linear equation based on $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$. Finding the true parameters $\boldsymbol{\beta}^*_1, \boldsymbol{\beta}^*_2$ is known to be NP-hard in general [6], even in the absence of noise. Accordingly, a common assumption in the literature requires both $\boldsymbol{x}_i$ and e_i to be sampled independently according to a Gaussian distribution, *i.e.*, $\boldsymbol{x}_i \sim \mathcal{N}(0, I_d)$ and $e_i \sim \mathcal{N}(0, \sigma^2)$. We assume, moreover, that the hidden variables are balanced, and independent of everything else.

At each iteration, the EM algorithm performs two steps, known as the E-step and M-step; these can be written as follows:

$$\begin{aligned} \text{E-step :} \qquad & Q(\boldsymbol{\beta}|\boldsymbol{\beta}_t) = \mathbb{E}_X[\sum_z p(z|X; \boldsymbol{\beta}_t) \log f(X, z; \boldsymbol{\beta})], \\ \text{M-step :} \qquad & \boldsymbol{\beta}_{t+1} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\beta}_t), \end{aligned} \tag{1}$$

where f is the probability distribution function that generates the data. Specifically, the E-step forms the likelihood function by assigning conditional probability to hidden labels given samples, based on the current estimator of true parameters. Subsequently, the M-step maximizes the expectation built at the E-step to find a new estimator. The EM algorithm alternates over these two steps iteratively until it converges. Due to the intuitive appeal of updating weights, and then updating the estimates of the $\boldsymbol{\beta}_i$, and also due to the computational tractability of each iteration, the EM algorithm has been widely used in many different applications.

When we take f to be the Gaussian distribution, the log becomes the squared loss function, and hence the M-step becomes the familiar (weighted) least squared loss minimization problem. In the finite-sample setting, the expectation over X is the empirical distribution over the samples $\{\boldsymbol{x}_i\}$ observed, and the EM update has a closed form expression: for sample-based EM operator with current estimator β, the update becomes

$$\text{(EM)} \qquad \tilde{\beta}' = (\frac{1}{n}\sum_{i=1}^n x_i x_i^T)^{-1} (\frac{1}{n}\sum_{i=1}^n \tanh(\frac{\beta^T x_i}{\sigma^2} y_i) y_i x_i). \tag{2}$$

In the setting where the covariates have identity covariance (or have been normalized to have identity covariance), it is also interesting, as we explain further below, to consider a further simplified version of EM that replaces $(\frac{1}{n}\sum_{i=1}^n x_i x_i^T)$ with its expectation. In this case, for the finite sample setting, the EM update, or "easy EM update" as we call it, takes the following form:

$$\text{(Easy-EM)} \qquad \tilde{\beta}'' = (\frac{1}{n}\sum_{i=1}^n \tanh(\frac{\beta^T x_i}{\sigma^2} y_i) y_i x_i). \tag{3}$$

The contribution of this work is to analyze these two iterations in the finite-sample setting, and thereby to provide guarantees for their convergence from a random initialization point.

1.2 Related Work and Main Contributions

Despite the popularity of EM in practice, our knowledge of when it converges to the true solution is still limited, as mentioned above. In general, it is known that EM algorithm may settle in a bad local optimum unless it starts from well initialized point [2]. Examples that illustrate this poor performance are, however, contradictory, at least in spirit, to the observations of EM's practical performance, where global convergence from random initialization has been observed and indeed conjectured for some time.

In MLR, the local convergence of EM has been recently established, i.e., it is known that the EM algorithm does converge to the global optimum if it starts from a small enough neighborhood around the true parameter [3, 4, 5, 6, 7]. In other words, if we assume we have the good fortune of starting from a point sufficiently close to the desired solution, the EM algorithm converges to it. However, the EM algorithm has been successful in practice for this particular problem even when started from randomly initialized point which is not necessarily close to the true parameter.

EM is also a very popular algorithm for a the somewhat simpler problem of clustering; specifically, its behavior has recently been studied for the problem of learning a Gaussian mixture model(GMM). Motivated by [4], two works in [8, 11] on global analysis of EM for the mixture of two Gaussians have recently delivered results that guarantee convergence of EM for this specific problem, from a random initialization.

For both GMM and MLR, the EM update involves a quadratic minimization problem, hence the problems are related. As we detail below in Section 2, there are some similarities between EM for the GMM problem, and EM for the MLR problem which we focus on here. Therefore we are spared from reinventing the wheel: some useful lemmas and proof techniques from analysis on GMM can be reused in our proof. We highlight these carefully in the details that follow. We also point out the numerous differences, that posed novel challenges to demonstrating the global convergence of EM for MLR.

MLR is in and of itself an interesting problem, and in fact, until work in [6, 12], there were no efficient results for the solution of the problem. The work in [12] developes a lifted convex formulation approach and using information theoretic arguments, obtains tight minimax bounds on solving MLR. A good initialization strategy for EM based on Stein's second-order lemma was proposed in [6], though this seems to rely on the noiseless setting which they study. The above two papers have focused on the two component case. Recent work has extended the focus to more components. Work in [13, 14] develops gradient descent based algorithms. In parallel, the work in [7, 15, 16] is based on tensor decomposition, which estimates third order moments to recover true parameters.

Recently, the work in [4] proposed a novel framework to analyze EM algorithm in general, and showed a local convergence result for MLR with two symmetric components as an application. A better local region was suggested in [5], where the convergence is guaranteed inside a region where the angle formed by the initialization with the true parameter is small enough. Still, all known results remain inherently local, and in particular, are not satisfied by random initialization, even when a norm bound on the true parameter is known.

Still, the question of whether EM converges from a random initialization, remains open. Our main contribution is to resolve this point affirmatively.

Main Contributions. We prove the global convergence of the EM algorithm, i.e., it converges with probability 1, when initialized from a random initialization point. We first establish this result for the infinite sample limit, i.e., population EM. We then develop concentration inequalities to couple the finite-sample version of EM with population EM, thereby providing a finite sample analysis. Though the general approach and most of the technical details differ, this approach of coupling the finite sample analysis with the population version is inspired by [4]. As we comment on in greater detail below, this coupling is strong enough to yield information theoretically optimal sample complexity dependence on the dimension, but our results still leave room for improvement on the precise dependence on the signal-to-noise ratio.

1.3 A Roadmap and Proof Outline

We provide a brief outline of the main steps of the paper. Because of our balanced (and independent) hidden label assumption, solving a standard regression in the infinite data setting (i.e., ignoring the fact that the problem is a mixture) returns the mean of the two parameters, $\frac{\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*}{2}$. Even with finite samples, we can get a center within small sampling error. This allows the re-centering of the data. Therefore, throughout this paper, we consider a symmetric model, *i.e.*, z_i is randomly selected with probability $\frac{1}{2}$ for each, and $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1^* = -\boldsymbol{\beta}_2^*$.

Analysis on Global Convergence of Population EM.

- **Decreasing Angle**. For both noisy and noiseless settings, previous work on mixed regression ([6, 4, 3, 17]) relies on demonstrating that the distance between the current iterate and the true solution, contracts at every iteration. Specifically, prior work has relied on demonstrating a fixed contraction with respect to l_2 distance, between the estimator and the true parameter. All techniques following this approach, are able to show that such a contraction holds only when the iterates are close enough to the true parameter, i.e., only under a sufficiently good initialization. Indeed, this is not surprising, for as pointed out in [5], the EM update may in fact result in *larger distance* from the solution at some specific points in the sequence of iterations. Thus, a proof of global convergence solely based on contraction in distance is fundamentally impossible. We needed another approach to prove global convergence.

  The first step in our proof is to show that the angle between the estimate and the true solution is always decreasing, unless we start from exactly orthogonal vector (a measure zero event). We show that the sine of the angle is globally contracting from the first EM update, regardless of the initial distance from the true parameter, $\boldsymbol{\beta}^*$. Consequently, EM quickly enters a local region where the direction of the current parameter estimate is well aligned to that of the true solution, $\boldsymbol{\beta}^*$. We then show that once the angles are well aligned, the estimate is close enough so that subsequent iterations indeed yield a contraction in distance.

- **Low SNR**. Even in this local region about the true parameter, previous results ([6, 4, 3, 17]) have additionally assumed high (or infinite) SNR; that is, the standard deviation, σ, of the additive noise, is of the same scale or smaller than the norm of the true parameter. Indeed, the low-SNR regime poses additional challenges, precisely because the SNR does impact the convergence rate. In our analysis, we reveal the explicit convergence rate as a function of the noise level, and tracking this dependence allows us to demonstrate convergence independently of the magnitude of σ. In particular, after the angle has become small enough, we show that the l_2 distance error decreases in rate $\max(c_1, 1 - c_2\eta^2)$ where $c_1 < 1, c_2 > 0$ is a constant that depends on the norm of initial guess and angle, and $\eta = \frac{||\boldsymbol{\beta}^*||}{\sigma}$ denotes the SNR.

- **Escaping Nearly Orthogonal Region**. As issued in [8], random initialization in d-dimensional space is highly likely to yield a vector whose projection in the direction of $\boldsymbol{\beta}^*$ is $O(1/\sqrt{d})$. In that region, the convergence behavior of sine or distance can be very subtle. We concern how many iterations should we run the algorithm until the estimator and $\boldsymbol{\beta}^*$ have enough correlation.

  In population EM, we show that we can escape from the nearly orthogonal region by simply running EM algorithm a few more steps. Specifically, in $O(\max(1, \eta^{-2}) \log d)$ number of EM iterations, we get an estimator whose projection into $\boldsymbol{\beta}^*$ direction is larger than a constant.

Finite Sample EM Analysis .

- After analyzing and proving global convergence for population EM, we provide our results on finite-sample based EM. In [4], a concentration bound for a sample-based estimator was provided in l_2 distance. Since our argument is based on contraction of angle, we extend the concentration result to bound cosine and sine of the angle. Then, we conclude that, starting from random initial guess in d-dimensional space, with $n = \tilde{O}(\max(1, poly(\eta^{-2}))d/\epsilon^2)$ fresh samples in each iteration, after $T = O(\max(1, \eta^{-2}) \max(\log d, \log(1/\epsilon)))$ iterations, we reach the l_2 error bounded by $O(\epsilon)$.

- **Small** ϵ. For the initial iterations of EM when the cosine between the estimate (or random initialization) and the true parameter can be as small as $1/\sqrt{d}$, our results have a dependence on ϵ^2. As long as ϵ is small (smaller than $1/\sqrt{d}$), this ϵ^2 term has no impact on the final result.
- **Arbitrary** ϵ. For ϵ significantly larger than $1/\sqrt{d}$, the ϵ^2 term may be significant. If instead we iterate using our "Easy-EM" iteration (i.e., Eq. (3)), then our analysis requires no such ϵ^2 term. Therefore our results show that one can either run Easy-EM until convergence (this information-theoretically optimal dependence on dimension, d and number of samples, n), or run Easy-EM until the cosine of the angle of the current estimate and the true parameter is larger than ϵ, and subsequently run EM.

Paper Organization The remainder of this paper is organized as follows. In Section 2, we derive a closed form equation of population EM in standard MLR setting. Section 3 is devoted to summarize our results on global convergence of population EM, and give a (sketch of) proof for each theorem. Then analysis on finite-sample based EM is provided in Section 4. All technical proofs that are not given in the main paper are deferred to the Appendix to facilitate readability.

2 Population EM Update

This section derives a closed form expression for the population EM operator. This serves as a starting point of our subsequent analysis and proof of convergence.

2.1 Basic Notation

We begin with establishing the notation we use throughout the remainder of the paper. We use X, Y to denote random variables that follow MLR with two components as defined in introduction. Then, $\boldsymbol{x}_i, y_i$ are samples of X and Y, respectively. Thanks to our symmetrization of the problem, the true parameters of the mixture are $-\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^*$, and so rather than referring to the pair of parameters, we express our results in terms of convergence to $\boldsymbol{\beta}^*$. Accordingly, at the t^{th} iteration of the algorithm, $\boldsymbol{\beta}_t \in \mathbb{R}^d$ is the current estimate of $\boldsymbol{\beta}^*$, the true parameter to be recovered. If we are interested in understanding the impact of a single iteration, we drop the subscript t and we use $\boldsymbol{\beta}$ in place of $\boldsymbol{\beta}_t$, and $\boldsymbol{\beta}'$ in place of $\boldsymbol{\beta}_{t+1}$. We use θ_t to denote the angle formed by $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^*$; similarly, θ_{t+1} corresponds to the angle formed by $\boldsymbol{\beta}_{t+1}$ and $\boldsymbol{\beta}^*$. Similarly to $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t+1}$, we often use θ and θ' for θ_t and θ_{t+1}. We assume without loss of generality that the initial angle with $\boldsymbol{\beta}^*$, θ_0 is in $[0, \pi/2)$. We exclude $\pi/2$ because it is a measure zero event for our random initialization. An initialization falling in the remainder of the circle has precisely the same behavior, but with convergence to $-\boldsymbol{\beta}^*$ in place of $\boldsymbol{\beta}^*$. Rather than add an additional subscript, we use the same notation in the infinite and finite sample analysis, as the context makes the distinction clear.

Every norm operator $\|\cdot\|$ without subscript is taken as the l_2 norm. We use σ to denote the standard deviation of the additive noise in each sample. An important parameter in the sequel is the signal-to-noise ratio (SNR); we use η to denote this:

$$\eta = \frac{\|\boldsymbol{\beta}^*\|}{\sigma}.$$

2.2 The EM Update

As in [4], the population EM operator for the problem we consider is

$$\boldsymbol{\beta}_{t+1} = 2\mathbb{E}[w_{\boldsymbol{\beta}_t}(X,Y)YX], \tag{4}$$

where $w_{\boldsymbol{\beta}_t}(X,Y)$ is defined as

$$w_{\boldsymbol{\beta}_t}(X,Y) = \frac{\exp\left(-\frac{(Y-\boldsymbol{\beta}_t^\top X)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(Y+\boldsymbol{\beta}_t^\top X)^2}{2\sigma^2}\right) + \exp\left(-\frac{(Y-\boldsymbol{\beta}_t^\top X)^2}{2\sigma^2}\right)}, \tag{5}$$

and the expectation is taken over the joint distribution of $X \times Y$.

We can check that the conditional distribution, $Y|X \sim 0.5\mathcal{N}(\boldsymbol{\beta}^{*\top}X, \sigma^2) + 0.5(-\boldsymbol{\beta}^{*\top}X, \sigma^2)$, is symmetric in sign. Thus, we can further simplify Eq. (4) by substituting it and dividing common factors in $w_{\boldsymbol{\beta}_t}$. This yields

$$\boldsymbol{\beta}_{t+1} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I)}\left[\left(\mathbb{E}_{y|\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}^\top\boldsymbol{\beta}^*,\sigma^2)}\left[\tanh\left(\frac{\boldsymbol{\beta}_t^\top\boldsymbol{x}}{\sigma^2}y\right)y\right]\right)\boldsymbol{x}\right]. \tag{6}$$

We focus on one update of population EM to see how each iteration yields the next estimator. First, we change the basis by choosing $\boldsymbol{v}_1 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$, the unit vector in the direction of the current estimator, and $\boldsymbol{v}_2$ to be the orthogonal complement of $\boldsymbol{v}_1$ in span$\{\boldsymbol{\beta}, \boldsymbol{\beta}^*\}$. We let $\boldsymbol{v}_3, ..., \boldsymbol{v}_d$ be a completion to an orthonormal basis for the full parameter space, $\mathbb{R}^d$, along with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. By the spherical symmetry of the distribution of $\boldsymbol{x}$, we have

$$\boldsymbol{\beta}' = \mathbb{E}_{\alpha_i}\left[\mathbb{E}_{y|\alpha_i}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}y\right)y\right]\sum_i \alpha_i\boldsymbol{v}_i\right], \tag{7}$$

where the expectation is taken over $\alpha_i \sim \mathcal{N}(0,1)$, and $y|\alpha_i \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)$, and we defined $b_1 = \langle \boldsymbol{\beta}, \boldsymbol{v}_1\rangle = \|\boldsymbol{\beta}\|$, $b_1^* = \langle \boldsymbol{\beta}^*, \boldsymbol{v}_1\rangle$, and $b_2^* = \langle \boldsymbol{\beta}^*, \boldsymbol{v}_2\rangle$. Without loss of generality, we assume $b_1, b_1^*, b_2^* \geq 0$.

The inner expectation over y does not have any dependence on α_i for $i \geq 3$. Therefore, taking expectation over α_i for $i \geq 3$ yields 0, which implies $\boldsymbol{\beta}'$ is also on the plane spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2$. It enables us to rewrite it as $\boldsymbol{\beta}' = b_1'\boldsymbol{v}_1 + b_2'\boldsymbol{v}_2$ where

$$b_1' = \mathbb{E}_{\alpha_1,\alpha_2}\left[\mathbb{E}_{y|\alpha_1,\alpha_2}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}y\right)y\right]\alpha_1\right], \tag{8}$$

$$b_2' = \mathbb{E}_{\alpha_1,\alpha_2}\left[\mathbb{E}_{y|\alpha_1,\alpha_2}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}y\right)y\right]\alpha_2\right], \tag{9}$$

where the expectation is similarly taken over $\alpha_i \sim \mathcal{N}(0,1)$, and $y|\alpha_i \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)$.

Before we move on to the convergence analysis, we further simplify b_1' and b_2' with the following lemma.

Lemma 1. *Let the variables b_1^*, b_2^*, b_1, b_1', b_2' be as defined above. Further, define $\sigma_2^2 = \sigma^2 + b_2^{*2}$. Then, we can derive the following simplified equations*

$$\begin{aligned} b_1' &= b_1^* S + R, \\ b_2' &= b_2^* S, \end{aligned} \tag{10}$$

where S and R are defined as

$$S = \mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right] \tag{11}$$

$$R = (\sigma^2 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right], \tag{12}$$

where expectation is taken over $\alpha_1 \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(0,\sigma_2^2)$.

Lemma 1 is the consequence of several applications of Stein's lemma [18], and the proof can be found in Appendix A.1. From this lemma, we can deduce that the new $\boldsymbol{\beta}'$ forms a smaller angle with $\boldsymbol{\beta}^*$. To see this, compare the cotangent of angles which $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}^*$ form with base axis, $\frac{b_1'}{b_2'} \geq \frac{b_1^*}{b_2^*}$. Since $\boldsymbol{\beta}'$ lies on the plane spanned by $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$, it always lies between the two vectors. Note that $R \geq 0$ since it is the expectation of positive values. Lemma 7 in the appendix shows that $S \geq 0$.

3 Main Results on Population EM

In this section, we provide our main results on the global convergence of the population EM algorithm for mixed linear regression. As described in the introduction, we adopt a different strategy from previous works that have established local convergence for mixed regression. Rather than trying to show the distance to the optimal parameter is contracting in all regions, we initially show the angle decreases. Specifically, as long as we begin from an initial vector not exactly orthogonal to $\boldsymbol{\beta}^*$, then (i) the sine of angle between $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^*$ geometrically decreases, (ii) when the angle has become less than $\pi/8$, we show that $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|$ is bounded by geometrically decreasing error. The amount of decrease in each case depends on the angle and SNR.

In d-dimensional space, a random initialization is highly likely to yield a vector whose inner product with $\boldsymbol{\beta}^*$ is $\Theta(1/\sqrt{d})$. Recall that this was also discussed as a technical challenge in [8]. In Subsection 3.2, we build our result on the growth of the cosine. We use this to show that $O(\log d)$ iterations of the EM algorithm suffice to bring the current iterate within a range of $O(1)$ inner product with $\boldsymbol{\beta}^*$.

Then Section 3.3 completes the result, proving linear convergence in a neighborhood of the optimal parameter.

3.1 Convergence of Sine

As the problem is symmetric in sign, without loss of generality, we assume that the inner product of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}^*$ is positive where $\boldsymbol{\beta}_0$ is the initial guess. Then, we prove sine of the angle geometrically converges to 0. This is reminiscent of the proof for Theorem 3 in [11] where they used a similar logic to show the convergence of population EM for the non-centered mixture of two Gaussians. However, that work does not provide an explicit rate of convergence. This makes it difficult to analyze the exact behavior of the angle at each iteration – something that is critical in order to port the population-level results to the finite sample setting.

The next result proves the convergence, and also provides a convergence rate, for the sine value of the angle between the current estimate and the true solution.

Theorem 1 (Convergence of sine)**.** *Let $0 \leq \theta < \frac{\pi}{2}$ be the angle between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Similarly, we denote by θ' the angle between $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}^*$. Then for every population EM iteration,*

$$\sin\theta' \leq \kappa \sin\theta,$$

where $\kappa = \left(\sqrt{1 + 2\frac{b_1^{*2}}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}}\right)^{-1} < 1$.

Proof. (Sketch) We provide a brief proof sketch here, and defer the details to the appendix. From Lemma 1, straightforward algebra yields

$$\sin\theta' = \sin\theta \frac{R}{\sqrt{R^2 + S^2\|\boldsymbol{\beta}^*\|^2 + 2SRb_1^*}}.$$

Our strategy is to find a good lower bound for $\frac{S}{R}$. Applying Stein's Lemma yields $\frac{S}{R} \geq \frac{b_1^*}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}$, which in turn gives the desired result by plugging in to the inequality. See Appendix A.2 for details. □

Remark (Convergence Rate)**.** *The constant κ can be rewritten in terms of η and θ, as:*

$$\kappa = \left(\sqrt{1 + \frac{2\eta^2}{1+\eta^2} \cos^2 \theta} \right)^{-1}.$$

From this expression, we can infer that the convergence rate increases as the angle decreases. We note that in the high SNR regime ($\eta \gg 1$), κ can be much smaller than 1 (depending on the initialization angle). In low SNR ($\gamma \ll 1$) regime, however, the convergence rate cannot be faster than $1 - O(\eta^2)$, regardless of the initial angle.

While we do not (and cannot) guarantee the contraction in distance between the estimator and true parameter, we want the norm of the estimator to remain bounded so that it does not blow up. The following lemma guarantees that while EM may increase the norm of $\boldsymbol{\beta}$, the growth can be controlled, i.e., we have bounds for $\|\boldsymbol{\beta}'\|$.

Lemma 2. *For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have*

$$\|\boldsymbol{\beta}'\| \leq 3\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}. \tag{13}$$

3.2 Convergence of Cosine

The previous section establishes the convergence of sine. While we could use this to conclude that eventually EM pushes any random initialization into a local neighborhood of the optimal solution, we need to further strengthen the results to establish a good bound on the number of steps of EM required to achieve this.

We accomplish this by turning to the cosine of the angle, and obtaining bounds on how quickly it grows with each EM iteration. In particular, we show that if we start from $\cos\theta_0 = \Theta(1/\sqrt{d})$, then $t = O(\log(d)\max(1, \eta^{-2}))$ iterations of EM is sufficient to guarantee $\cos\theta_t = O(1)$.

Theorem 2. *As long as $\frac{\pi}{2} > \theta \geq \frac{\pi}{3}$, one population EM iteration yields*

$$\cos\theta' \geq \kappa \cos\theta,$$

where $\kappa = \sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}}$.

If $\cos\theta_0 = \Theta(1/\sqrt{d})$, after $t = O(\log(d)\max(1, \eta^{-2}))$ iterations, we get $\theta_t < \pi/3$ or $\cos\theta_t \geq \frac{1}{2}$.

We defer the proof to Appendix A.3.

3.3 Convergence of Distance

We have established that $\sin\theta_t$ is geometrically decreasing, and $\cos\theta_t$ geometrically increasing, and in particular that after at most $O(\log(d)\max(1, \eta^{-2}))$ iterations, the angle between the current iterate and the optimal solution is no more than $\pi/8$. We now turn our attention to the convergence to the true $\boldsymbol{\beta}^*$. The next result shows that the distance does in fact contract, once the angle has become smaller than $\pi/8$.

Theorem 3 (Convergence in Distance)**.** *Assume that $\theta < \pi/8$, and define $\sigma_2^2 = \sigma^2 + b_2^{*2}$. If $b_2^* < \sigma$ or $\frac{\sigma_2^2}{\sigma^2} b_1 < b_1^*$, then*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \leq \kappa\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \kappa(16\sin^3\theta)\|\boldsymbol{\beta}^*\|\frac{\eta^2}{1+\eta^2}, \tag{14}$$

where $\kappa = \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)^2}{\sigma_2^2}}\right)^{-1}$.

Otherwise, we get

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \leq 0.6\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{15}$$

Proof. (Sketch) We show this result holds by finding a bound for the difference in each coordinate separately after one EM update. That is, we obtain bounds for $|b_1' - b_1^*|$ and $|b_2' - b_2^*|$ respectively. While this is a standard way to get an upper bound for $\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\|$, the challenge comes from the fact that bounding $|b_1' - b_1^*|$ solely in terms of $|b_1 - b_1^*|$ is not always possible.

To see that, consider the case when $b_1 = b_1^*$. If b_2^* were 0, then b_1^* becomes a fixed point in equation (8), yielding $b_1' = b_1^*$ as desired. However, if $b_2^* > 0$, then b_1^* is not a fixed point and b_1' will be moved slightly apart from it. As a result, $|b_1' - b_1^*| > |b_1 - b_1^*| = 0$.

Instead, we obtain a bound using $|b_1 - b_1^*|$ and b_2^*. In Appendix A.4, we prove the following inequality holds:

$$|b_1' - b_1^*| \leq \kappa^3 \left|(b_1 - b_1^*) + \frac{{b_2^*}^2}{\sigma^2} b_1\right|, \tag{16}$$

using the similar logic used in [8]. While this does not directly yield (14), an intuition can be obtained from it. If $b_2^* = 0$, i.e., if $\boldsymbol{\beta}$ is perfectly aligned to $\boldsymbol{\beta}^*$, we easily get a contraction in distance. It can be then combined with the fact that $b_2^* = \|\boldsymbol{\beta}^*\| \sin\theta$ decreases geometrically as shown in the previous subsection to conclude that EM converges to $\boldsymbol{\beta}^*$.

However, another technical difficulty arises when $\sigma \to 0$, i.e., in the noiseless case, as the right-hand side becomes infinitely large, making (16) useless. Therefore, we divide the cases by when $b_2^* < \sigma$ and $b_2^* \geq \sigma$ to avoid pitfalls in the limit of small σ. See Appendix A.4 for a detailed proof. □

We note that this result does not conflict with any previous result on local convergence of EM. In fact, if we start from smaller angle and assume noise is small enough, we can guarantee locally constant rate of contraction in distance.

Corollary 1 (Local Convergence with Small Noise)**.** *If $\boldsymbol{\beta}$ is a point close enough to $\boldsymbol{\beta}^*$ and η is sufficiently large, then*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \leq 0.6\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{17}$$

Proof. First note that the proximity of $\boldsymbol{\beta}$ implies small angle θ. Large η and small θ gives a proper upper bound on κ. Suppose we set a condition such that κ is guaranteed to be less than 0.5. For instance, we can set $2 \leq \eta$ and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| < \|\boldsymbol{\beta}^*\|/16$. This condition leads us to

$$b_1^* \geq \frac{15}{16}\|\boldsymbol{\beta}^*\|, \quad b_1 \geq \frac{15}{16}\|\boldsymbol{\beta}^*\|,$$
$$\sigma_2^2 \leq \sigma^2(1 + \frac{1}{16}\eta^2),$$

which altogether gives a numerical bound to $\kappa = \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)^2}{\sigma_2^2}}\right)^{-1}$ less than 0.5.

Then small $\sin\theta < 1/16$ restricts the second term in (14) to be smaller than $0.1\|\boldsymbol{\beta}^*\| \sin\theta \leq 0.1\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$. Then they altogether become less than $0.6\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$. □

To conclude on the convergence of distance in arbitrary noise level, we bound the error after t iterations of population EM.

Corollary 2. *Assume we start from $\theta_0 < \pi/8$. After t iterations of population EM, there exists some constant $\kappa < 1$ such that,*

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\| < \kappa^t \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + t\kappa^t \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}. \tag{18}$$

In particular, the result is satisfied if we take κ to be the maximum among

$$0.6, \sqrt{\left(1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}\right)^{-1}}, \sqrt{1 - \frac{0.8\eta^2}{1+\eta^2}}. \tag{19}$$

Remark. *For the convergence rate, observe that it depends on the norm of initial guess and the SNR, η. In the Appendix, we show that the convergence rate only becomes faster as the EM algorithm proceeds. Depending on the SNR, the convergence rate is either constant or $1 - O(\eta^2)$, as was in the case of sine. Therefore, the required number of iterations t is $O(\max(1, \eta^{-2}) \log(1/\epsilon))$ to achieve an ϵ-optimal solution. Note that $\frac{t\eta^2}{1+\eta^2} = \tilde{O}(1)$.*

3.4 Summary of Phases in Convergence Behavior

Collecting all results we have, we can summarize the behavior of population EM algorithm in three phases.

1. Starting from randomly initialized vector in d-dimensional space, after $O(\max(1, \eta^{-2}) \log d)$ iterations, we reach the angle below $\pi/3$.

2. Starting from a vector whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/3$, after $O(\max(1, \eta^{-2}))$ iterations we have an estimator whose angle with $\boldsymbol{\beta}^*$ is less than $\pi/8$.

3. Starting from a vector which is well aligned with $\boldsymbol{\beta}^*$ such that the angle between two vectors is less than $\pi/8$, after $O(\max(1, \eta^{-2}) \log(1/\epsilon))$ iterations, we have an error less than $O(\epsilon)$.

4 Finite Sample Analysis

We now turn to prove the convergence of finite-sample based EM. As discussed in the outline, our approach is to couple the finite sample EM to the population EM.

Along the way, we also prove the convergence of the Easy-EM algorithm. As we discuss in length below, this is interesting on its own, but also useful in the setting where ϵ is chosen as an $O(1)$ quantity, rather than when we take it very small.

We define additional notation that we use in this section. When we focus on one finite-sample based EM iteration, we use $\boldsymbol{\beta}$ to denote our current estimator, $\boldsymbol{\beta}'$ to denote the result from one step of the *population EM operator*, and $\tilde{\boldsymbol{\beta}}'$ to denote the result from one step of the finite-sample EM operator. We use $\tilde{\theta}'$ to represent the angle formed by $\tilde{\boldsymbol{\beta}}'$ and $\boldsymbol{\beta}^*$. When we consider the sequence of estimators from finite-sample based EM, we use $\boldsymbol{\beta}_t$ for the estimator at the t^{th} iteration.

We consider sample-splitting as an analysis technique, as it renders subsequent steps of the EM algorithm independent. As with the many other papers that have used this technique, we believe it is an artifact of the analysis, but we as well are unable to find a way to remove it. Recall that the closed form equation for sample-based EM operator with current estimator $\boldsymbol{\beta}$ is as given in Eq. (2), which we reproduce here:

$$\tilde{\boldsymbol{\beta}}' = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \tanh\left(\frac{\boldsymbol{\beta}^\top \boldsymbol{x}_i}{\sigma^2} y_i\right) y_i \boldsymbol{x}_i\right).$$

For simplicity, let us assume that the problem is normalized, *i.e.*, $\|\boldsymbol{\beta}^*\| = 1$. In this normalized setting, $\sigma^2 = \eta^{-2}$. We use η^{-2} when we control statistical error in this section, in order to state the dependence of statistical error and sample complexity on SNR more explicitly. All statistical error in l_2 norm will therefore implicitly have a dependence on $\|\boldsymbol{\beta}^*\|$.

Work in [4] establishes a bound between the population EM update and the finite sample EM update. Specifically, starting from $\boldsymbol{\beta}$, then with probability at least $1-\delta$, the update $\tilde{\boldsymbol{\beta}}'$ from one iteration of finite-sample EM with n samples is controlled by $\boldsymbol{\beta}'$, the population-EM update, as follows:

$$\|\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \leq c\sqrt{(\sigma^2 + \|\boldsymbol{\beta}^*\|^2)\frac{d}{n}\log(1/\delta)}.$$

Equivalently, that with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ samples, we have $\|\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \leq \epsilon$.

Recall, however, that to make our analysis global, we need to make use of contraction of sine and cosine. We thus need to couple these sufficiently strongly to the finite sample setting. That is, we must show that cosine and sine of the angle are concentrated around the respective quantities from population EM. For that purpose, we need a more fine-grained concentration result. This is the content of the next theorem.

Theorem 4. *Consider one iteration of sample-based EM algorithm. There exist absolute constants* $c_1, c_2 > 0$*, such that statistical error in a fixed direction* $\boldsymbol{\beta}^*$ *can be bounded with probability at least* $1-\delta$*, by*

$$|(\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^\top \boldsymbol{\beta}^*| \leq \sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\left(c_1\sqrt{\frac{1}{n}\log(1/\delta)} + c_2\frac{d}{n}\log(1/\delta)\right). \tag{20}$$

Corollary 3. *Suppose that the norm of the estimator* $\|\boldsymbol{\beta}\|$ *is larger than* $\frac{\|\boldsymbol{\beta}^*\|}{10}$*. Then, with* $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ *samples for one finite-sample based EM iteration, we have*

$$\cos\tilde{\theta}' \geq \kappa(1-10\epsilon)\cos\theta - O\left(\max\left(\frac{\epsilon}{\sqrt{d}}, \epsilon^2\right)\right), \tag{21}$$

$$\sin^2\tilde{\theta}' \leq \kappa'\sin^2\theta + O(\epsilon), \tag{22}$$

with $\kappa = \sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \geq 1$*, and* $\kappa' = (1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta)^{-1} < 1$.

The theorem implies that the statistical error is very small in a fixed direction $\boldsymbol{\beta}^*$. We note the extra factor ϵ^2 in the bound. Technically, this arises from controlling the impact of the inverse of sample covariance matrix $(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}$. This term is negligible when we select ϵ small enough, namely, $\epsilon < 1/\sqrt{d}$. In fact, the proofs of Theorem 4 and Corollary 3 (in the Appendix), demonstrate that the ϵ^2 term becomes negligible as soon as $\epsilon < \langle\boldsymbol{\beta}_0, \boldsymbol{\beta}^*\rangle$, i.e., as soon as ϵ is smaller than the dot product of the estimate with the true solution. Thus the term is negligible either if ϵ is very small (smaller than $1/\sqrt{d}$), or if the current iterate is good. In Section 4.1 we show that Easy-EM exhibits very similar convergence behavior, without the appearance (at all) of the ϵ^2 term. Therefore, one can simply run Easy-EM to ϵ-convergence, as guaranteed by the results of Section 4.1, or one could run Easy-EM and then switch to EM.

For now, we assume that one of the conditions described above holds, and thus we can assume that the ϵ^2 term can be safely ignored.

The condition for the norm of initial guess is to guarantee that the overall statistical error ϵ cannot dominate the angle. We show that once the norm of estimator is greater than $\frac{\|\boldsymbol{\beta}^*\|}{10}$, the EM algorithm maintains the norm of our next estimator larger than $\frac{\|\boldsymbol{\beta}^*\|}{10}$.

Lemma 3. *If* $\|\boldsymbol{\beta}\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$*, then after one finite-sample EM update with* $n = O(\max(1, poly(\eta^{-2}))$ $(d/\epsilon^2))$ *samples,* $\|\tilde{\boldsymbol{\beta}}'\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$.

Thus, if our initial guess is sufficiently large in norm, we keep having estimator large enough so that we can ignore ϵ in norm. It can be easily satisfied by setting the norm of initial estimator large enough. We note here that $\frac{\|\boldsymbol{\beta}^*\|}{10}$ is a pessimistic choice of lower bound for the norm of the estimator, and in practice it quickly increases to $\|\boldsymbol{\beta}^*\|$ even if we start from small initial vector. The proof for Theorem 4 and Lemma 3 are given in Appendix C.

With (21) and (22), we can give our results on sample-splitting finite-sample based EM for our problem. Recall that the convergence rates of sine and distance are $1 - O(\eta^2)$ in low SNR regime. Therefore, the statistical error has to be smaller than η^2 in order to guarantee that every iteration improves the estimator. It makes the sample complexity heavily dependent on η, which becomes $O(\eta^{-6}d/\epsilon^2)$ in low SNR regime. We revisit this high dependency on SNR after we state our main theorem. For the following results, we will slightly abuse the notation ϵ so that it is in fact $\epsilon_1 \min(1, \eta^2)$ where $\epsilon_1 > 0$ is the desired statistical error we control, instead of ϵ.

Lemma 4 (Increasing Cosine in Finite-Sample EM)**.** *Let $\boldsymbol{\beta}_0$ be the initial guess and θ_t be the angle formed by $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^*$ and assume $\|\boldsymbol{\beta}_0\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$. We run sample-splitting sample-based EM algorithm, each step with $n/T = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ samples, $T = O(\max(1, \eta^{-2}) \log d)$ iterations, and $\epsilon = \epsilon_1 \min(1, \eta^2)$. We also take $\epsilon_1 > 0$ small enough such that there exists a constant $\kappa = (1 - 10\epsilon)\sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}} > 1$. As long as $\theta_t > \pi/3$ for $t \leq T$, with high probability,*

$$\cos\theta_T \geq \kappa^\top \cos\theta_0 - \frac{\kappa^\top - 1}{\kappa - 1} O\big(\frac{\epsilon}{\sqrt{d}}\big). \tag{23}$$

In particular, when $\cos\theta_0 = \Theta\big(\frac{1}{\sqrt{d}}\big)$, we get $\cos\theta_T \geq \frac{1}{2} - O(\epsilon_1)$.

Lemma 5 (Convergence of Sine in Finite-Sample EM)**.** *Suppose we get a $\boldsymbol{\beta}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/3$ from previous phase. We run sample-splitting sample-based EM with sample complexity $n/T = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ and error $\epsilon = \epsilon_1 \min(1, \eta^2)$. Then with high probability, with a constant $\kappa = \big(\sqrt{1 + \frac{0.5\eta^2}{1+\eta^2}}\big)^{-1} < 1$,*

$$\sin^2\theta_T \leq \kappa^{2T} \sin^2\theta_0 + O(\epsilon_1). \tag{24}$$

After $T = O(\max(1, \eta^{-2}))$ iterations, we have $\sin^2\theta_T \leq \sin^2\frac{\pi}{8} + O(\epsilon_1)$.

Remark. *We will take $\epsilon_1 > 0$ small enough such that as long as $\sin^2\theta$ is not too small, in each iteration*

$$\big(1 + \frac{0.5\eta^2}{1+\eta^2}\big)^{-1} \sin^2\theta + O(\epsilon) \leq \sin^2\theta,$$

i.e., θ' keeps remaining less than previous angle with high probability. Therefore, we are convinced that our estimator stays in the second phase of convergence. Note that we set $\epsilon = \epsilon_1 \min(1, \eta^2)$ such that it is guaranteed with sufficiently small ϵ_1.

Finally, suppose we have reached the angle below $\pi/8$. We provide a convergence guarantee in l_2 distance for sample based EM.

Lemma 6 (Convergence of Distance in Finite-Sample EM)**.** *Suppose we get $\boldsymbol{\beta}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/8$ from previous phase. We run sample-splitting EM with $n/T = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ and $\epsilon = \epsilon_1 \min(1, \eta^2)$, getting*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| \leq \kappa^\top \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + T\kappa^\top \frac{\eta^2}{1+\eta^2} + O(\epsilon_1), \tag{25}$$

where κ is the maximum among (19) as in Corollary 2.

After $T = O(\max(1, \eta^{-2}) \log(1/\epsilon_1))$ iterations, we get $O(\epsilon_1)$ optimal error.

Collecting all the lemmas we have stated in this section, we can conclude this section with the following theorem for sample-splitting sample-based EM.

Theorem 5. *Suppose we start from initial vector whose correlation with $\boldsymbol{\beta}^*$ is at least $\Omega(\frac{1}{\sqrt{d}})$, with norm larger than at least $\frac{\|\boldsymbol{\beta}^*\|}{10}$ in d-dimensional space. We run sample-splitting sample-based EM algorithm with $O(\max(1, poly(\eta^{-2}))\ (d/\epsilon^2)\log(T/\delta))$ fresh samples in each iteration, where ϵ is the desired statistical error less than $1/\sqrt{d}$. After $T = O(\max(1, \eta^{-2})\ \max(\log d, \log(1/\epsilon)))$ iterations, we get*

$$\mathbb{P}(\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| \le \epsilon) \ge 1 - \delta.$$

The overall sample complexity in order to achieve ϵ error is $n = \tilde{O}(\max(1, poly(\eta^{-2}))(d/\epsilon^2))$. In the high SNR regime, this is the optimal rate of sample complexity up to logarithmic factors. In the low SNR regime, we can see that EM algorithm is very sensitive to SNR and we need presumably large number of samples. This high dependency on SNR is to guarantee the statistical error be less than η^2, as we are trying to convince that every iteration improves the estimator. It seems to be the nature of EM algorithm as we have seen similarly high dependence on SNR in GMM settings [8]. Nevertheless, once enough number of samples are given to offset low SNR, the statistical error is still optimal up to logarithmic factors in dimension and error.

4.1 Initialization of EM with non-small ϵ

As mentioned, we believe the ϵ^2 in equation (21) is simply an artifact of our analysis. Therefore here, we analyze Easy-EM. Recall that the Easy-EM update is given by:

$$\tilde{\boldsymbol{\beta}}'' = (\frac{1}{n}\sum_{i=1}^{n} \tanh(\frac{\boldsymbol{\beta}^\top x_i}{\sigma^2} y_i) y_i x_i).$$

We note that this too is an unbiased estimator of population EM operator.

We show that Easy-EM behaves in an almost identical fashion as EM. Because of the absence of the inverse of the empirical covariance matrix, our analysis does not require the ϵ^2 term. Thus, for large ϵ, one can obtain the same guarantees as the main theorem above, either by simply using this Easy-EM algorithm until convergence, or by using Easy-EM, and then transitioning to EM.

Our main result is as follows.

Theorem 6. *Consider one iteration of Easy-EM algorithm. There exist absolute constants $c_1, c_2 > 0$, such that with probability at least $1 - \delta$,*

$$\|\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}'\| \le c_1\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\sqrt{\frac{d}{n}\log(1/\delta)},$$
$$|(\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}')^\top \boldsymbol{\beta}^*| \le c_2\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\sqrt{\frac{1}{n}\log(1/\delta)}.$$

Furthermore, suppose that the norm of current estimator $\|\boldsymbol{\beta}\|$ is larger than $\frac{\|\boldsymbol{\beta}^*\|}{10}$. Then, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ samples for one Easy-EM iteration, we have

$$\cos\tilde{\theta}'' \ge \kappa(1 - 10\epsilon)\cos\theta - O(\frac{\epsilon}{\sqrt{d}}),$$
$$\sin^2\tilde{\theta}'' \le \kappa'\sin^2\theta + O(\epsilon),$$

with $\kappa = \sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \ge 1$, and $\kappa' = \left(1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta\right)^{-1} < 1$.

The only difference between Theorem 4 and 6 is that we do not have an extra factor ϵ^2 in equation (6). Thus, we do not require ϵ to be smaller than $1/\sqrt{d}$ anymore, and Lemmas 4, 5, and 6, and Theorem 5 can be identically applied to this simplified version of EM.

The reason we obtain the same results for Easy-EM and standard finite-sample EM is that the statistical error we use for both versions of EM relies on the concentration of $\tilde{\boldsymbol{\beta}}''$ around $\boldsymbol{\beta}'$. The inverse of the sample covariance matrix in standard EM presents a technical analytical challenge under the framework where we are trying to control the behavior of finite-sample EM through coupling with population EM; as practical experience also supports, the effect of the inverse of the sample covariance matrix is beneficial. For instance, using a different approach that does not have a global analog, we know that we have an exact (but local) recovery guarantee when finite-sample EM is directly analyzed in noiseless case [6]. It would be interesting to explore precisely how the standard EM has advantages over Easy-EM in more general settings. We leave it as a future work.

To conclude this section, we propose that when ϵ is larger than $1/\sqrt{d}$, we run Easy-EM for $O(\log(\epsilon/\sqrt{d})\max(1,\eta^{-2}))$ iterations to get $\cos\theta$ larger than ϵ. At this point, both $\frac{\epsilon}{\sqrt{d}}$ and ϵ^2 become small enough compared to cosine. Then we return to standard version of finite-sample EM, and run it until it converges. We summarize our proposition as follows.

Summary of Finite-Sample EM In all iterations, we use $n = O(\max(1, poly(\eta^{-2}))d/\epsilon^2)$ fresh samples.

1. Starting from randomly initialized vector with large enough norm in d-dimensional space, compare the statistical error ϵ to $1/\sqrt{d}$. If it is smaller than $1/\sqrt{d}$, then go to step 2. Otherwise, run Easy-EM for $O(\log(\epsilon/\sqrt{d})\max(1,\eta^{-2}))$ iterations to get $\cos\theta \geq O(\epsilon)$.

2. Run finite-sample based EM for $O(\min(\log d, \log(1/\epsilon))\max(1,\eta^{-2}))$ iterations to get $\cos\theta \geq 1/2$.

3. Run finite-sample based EM for $O(\max(1,\eta^{-2}))$ iterations to get $\sin\theta \leq \sin(\pi/8)$.

4. Run finite-sample based EM for $O(\max(1,\eta^{-2})\log(1/\epsilon))$ iterations to get $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq O(\epsilon)$.

5 Conclusion

We studied the convergence behavior of both population EM and sample-based EM, and at least for two mixture of symmetric linear regression models we have shown that EM algorithm can recover the true parameters. In particular, we found that EM converges to true parameters globally without any specialized initialization algorithm in large sample limits. The explicit convergence rates of cosine, sine, and distances were presented, from which we concluded $O(\max(1,\eta^{-2})\max(\log d, \log(1/\epsilon)))$ steps are required to get l_2 error less than ϵ. This is the first result that has shown the global convergence of EM, as well as the convergence in low SNR regime.

In finite sample case, we showed that EM enjoys the same convergences behavior, though it may need the aid of Easy-EM in the first few steps, given the number of samples $\tilde{O}(\max(1, poly(\eta^{-2}))d/\epsilon^2)$ in each iteration. It would be interesting to explore whether we can remove the dependency on Easy-EM steps, as well as if we can improve the sample complexity in terms of SNR. Extensions of this work could be analyzing the performance of EM when the weight of each component is not equal or there are more than two components, as well as applying our results to high-dimensional setting.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

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

[3] Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Proc. Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

[4] 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, February 2017.

[5] Jason M. Klusowski, Dana Yang, and W. D. Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *arXiv:1704.08231*, April 2017.

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

[7] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

[8] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710, 2017.

[9] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

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

[11] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

[12] Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604, 2014.

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

[14] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pages 2190–2198, 2016.

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

[16] 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.

[17] Chong Wu, Can Yang, Hongyu Zhao, and Ji Zhu. On the convergence of the em algorithm: A data-adaptive analysis. *arXiv preprint arXiv:1611.00519*, 2016.

[18] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

Appendix A Proofs for Main Results on Population EM

A.1 Proof of Lemma 1

Lemma 1. *Let the variables b_1^*, b_2^*, b_1, b_1', b_2' be as defined above. Further, define $\sigma_2^2 = \sigma^2 + {b_2^*}^2$. Then, we can derive the following simplified equations*

$$\begin{aligned} b_1' &= b_1^* S + R, \\ b_2' &= b_2^* S, \end{aligned} \tag{10}$$

where S and R are defined as

$$S = \mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right] \tag{11}$$

$$R = (\sigma^2 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right], \tag{12}$$

where expectation is taken over $\alpha_1 \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(0, \sigma_2^2)$.

Proof. We start with second coordinate b_2'. We will occasionally omit variables for expectation when it is clear over which variable the expectation is taken. We can rewrite the equation (9) as

$$b_2' = \mathbb{E}[g(\alpha_1, \alpha_2)\alpha_2],$$

where $g(\alpha_1, \alpha_2) = \mathbb{E}_{y\sim\mathcal{N}(0,\sigma^2))}[\tanh(\frac{b_1\alpha_1}{\sigma^2}(y + \alpha_1 b_1^* + \alpha_2 b_2^*))(y + \alpha_1 b_1^* + \alpha_2 b_2^*)]$. Apply Stein's lemma with respect to α_2 yields

$$\begin{aligned} b_2' = \mathbb{E}[g(\alpha_1, \alpha_2)\alpha_2] &= \mathbb{E}\left[\frac{\partial}{\partial\alpha_2} g(\alpha_1, \alpha_2)\right], \\ \frac{\partial}{\partial\alpha_2} g(\alpha_1, \alpha_2) &= b_2^*\mathbb{E}_{y\sim\mathcal{N}(0,\sigma^2))}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^* + \alpha_2 b_2^*)\right) + \right. \\ &\qquad \left. \frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^* + \alpha_2 b_2^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^* + \alpha_2 b_2^*)\right)\right] \\ &\overset{(a)}{=} b_2^*\mathbb{E}_{y\sim\mathcal{N}(0,\sigma_2^2))}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right]. \\ \therefore b_2' &= b_2^* S, \end{aligned}$$

where in (a), we replaced $y + \alpha_2 b_2^*$ with a new Gaussian variable as they are the sum of two Gaussian variables.

For the first coordinate b_1', we take the similar strategy but we arrange it in a different way. First, we rewrite equation (8) as

$$b_1' = \mathbb{E}_{\alpha_1\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{y\sim\mathcal{N}(0,\sigma_2^2))}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)(y + \alpha_1 b_1^*)\right]\alpha_1\right], \tag{26}$$

where we again replaced $y + \alpha_2 b_2^*$ with one Gaussian variable. Then observe that another application of Stein's lemma yields

$$\begin{aligned}
\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\alpha_1^2\right] &= \mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right) + \left(\frac{2b_1^* b_1}{\sigma^2}\alpha_1 + \frac{b_1}{\sigma^2}y\right)\alpha_1 \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right] \\
&= \mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right] \\
&\quad + b_1^* \mathbb{E}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right]. \qquad (27)
\end{aligned}$$

On the other hand,

$$\mathbb{E}_{\substack{\alpha_1 \sim \mathcal{N}(0,1) \\ y \sim \mathcal{N}(0,\sigma_2^2))}}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\alpha_1 y\right] = \sigma_2^2 \mathbb{E}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right],$$

where we applied Stein's lemma for y this time. Plugging the above two equations into (26), we get

$$b_1' = b_1^* S + R,$$

completing the proof. □

A.2 Proof of Theorem 1

Theorem 1 (Convergence of sine)**.** *Let $0 \le \theta < \frac{\pi}{2}$ be the angle between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Similarly, we denote by θ' the angle between $\boldsymbol{\beta'}$ and $\boldsymbol{\beta}^*$. Then for every population EM iteration,*

$$\sin\theta' \le \kappa \sin\theta,$$

where $\kappa = \left(\sqrt{1 + 2\frac{b_1^{*2}}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}}\right)^{-1} < 1.$

Proof. From equation (10), we can compute cosine and sine.

$$\cos\theta' = \frac{<\boldsymbol{\beta}^*, \boldsymbol{\beta'}>}{||\boldsymbol{\beta}^*|| ||\boldsymbol{\beta'}||} = \frac{S||\boldsymbol{\beta}^*||^2 + Rb_1^*}{||\boldsymbol{\beta}^*||\sqrt{R^2 + S^2||\boldsymbol{\beta}^*||^2 + 2SRb_1^*}}, \qquad (28)$$

$$\begin{aligned}
\sin\theta' &= \frac{Rb_2^*}{||\boldsymbol{\beta}^*||\sqrt{R^2 + S^2||\boldsymbol{\beta}^*||^2 + 2SRb_1^*}} \\
&= \sin\theta \frac{1}{\sqrt{1 + (S/R)^2||\boldsymbol{\beta}^*||^2 + 2(S/R)b_1^*}} \\
&\le \sin\theta \frac{1}{\sqrt{1 + 2(S/R)b_1^*}}. \qquad (29)
\end{aligned}$$

Now we are left with proving $\frac{S}{R}b_1^* \ge \frac{b_1^{*2}}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}$, which gives us the claimed result by plugging it into (29). To see that, we first observe

$$S = \underbrace{\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}y \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right]}_{A} + \underbrace{b_1^* \mathbb{E}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)\right]}_{(\frac{b_1^*}{\sigma^2 + ||\boldsymbol{\beta}^*||^2})R}.$$

Since $R \geq 0$ as it is the expectation of positive function, if A is greater than 0, then we get the desired result. Another application of Stein's lemma yields

$$\begin{aligned}\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)y^2\right] &= \sigma_2^2 \mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2} y \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right] \\ &= \sigma_2^2 A.\end{aligned}$$

We can rewrite the left side as

$$\begin{aligned}\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)y^2\right] &= \frac{1}{2}\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)y^2\right] + \frac{1}{2}\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(-y+\alpha_1 b_1^*)\right)y^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(-y+\alpha_1 b_1^*)\right) + \tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right)y^2\right] \\ &\geq 0,\end{aligned}$$

where in the last inequality, we used the fact that $\tanh(c+x) + \tanh(-c+x) \geq 0$ when $x \geq 0$ for any real value c. Consequently, $A \geq 0$ and we complete the proof. □

A.3 Proof of Theorem 2

Theorem 2. *As long as $\frac{\pi}{2} > \theta \geq \frac{\pi}{3}$, one population EM iteration yields*

$$\cos\theta' \geq \kappa \cos\theta,$$

where $\kappa = \sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}}$.

If $\cos\theta_0 = \Theta(1/\sqrt{d})$, after $t = O(\log(d)\max(1, \eta^{-2}))$ iterations, we get $\theta_t < \pi/3$ or $\cos\theta_t \geq \frac{1}{2}$.

Proof. Recall that from the proof in Theorem 1, we have

$$\cos\theta' = \frac{S||\boldsymbol{\beta}^*||^2 + Rb_1^*}{||\boldsymbol{\beta}^*||\sqrt{R^2 + 2SRb_1^* + S^2||\boldsymbol{\beta}^*||^2}}, \qquad \text{and} \qquad \frac{S}{R} \geq \frac{b_1^*}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}.$$

Starting from these two equations, we can get a lower bound of $\cos\theta'$ in terms of $\cos\theta$ and σ. First observe that

$$\begin{aligned}\cos\theta' &= \frac{(S/R)||\boldsymbol{\beta}^*||^2 + b_1^*}{||\boldsymbol{\beta}^*||\sqrt{1 + 2(S/R)b_1^* + (S/R)^2||\boldsymbol{\beta}^*||^2}} \\ &\overset{(a)}{\geq} \frac{b_1^*(1 + \frac{||\boldsymbol{\beta}^*||^2}{||\boldsymbol{\beta}^*||^2+\sigma^2})}{||\boldsymbol{\beta}^*||\sqrt{1 + {b_1^*}^2\frac{1}{||\boldsymbol{\beta}^*||^2+\sigma^2}(2 + \frac{||\boldsymbol{\beta}^*||^2}{||\boldsymbol{\beta}^*||^2+\sigma^2})}} \\ &\overset{(b)}{\geq} \cos\theta\sqrt{1 + \frac{{b_2^*}^2}{k(\sigma^2)^{-1} + {b_1^*}^2}},\end{aligned}$$

where $k(\sigma^2) = \frac{1}{||\boldsymbol{\beta}^*||^2+\sigma^2}(2 + \frac{||\boldsymbol{\beta}^*||^2}{||\boldsymbol{\beta}^*||^2+\sigma^2})$. (a) comes from the following:

$$\begin{aligned}\frac{(S/R)||\boldsymbol{\beta}^*||^2 + b_1^*}{||\boldsymbol{\beta}^*||\sqrt{1 + 2(S/R)b_1^* + (S/R)^2||\boldsymbol{\beta}^*||^2}} &= \sqrt{\frac{(S/R)^2||\boldsymbol{\beta}^*||^2 + 2(S/R)b_1^* + {b_1^*}^2/||\boldsymbol{\beta}^*||^2}{1 + 2(S/R)b_1^* + (S/R)^2||\boldsymbol{\beta}^*||^2}} \\ &= \sqrt{1 - \frac{{b_2^*}^2/||\boldsymbol{\beta}^*||^2}{1 + 2(S/R)b_1^* + (S/R)^2||\boldsymbol{\beta}^*||^2}},\end{aligned}$$

which shows us that $\cos\theta'$ is an increasing in (S/R) and therefore lower bounded by the lowest possible value of (S/R).

From (b), we can infer that the amount of increase gets smaller as the angle gets smaller. Thus, we can further bound it with straight-forward algebra by

$$\cos\theta\sqrt{1+\frac{{b_2^*}^2}{k(\sigma^2)^{-1}+{b_1^*}^2}} \geq \cos\theta\sqrt{1+\frac{\sin^2\theta}{\cos^2\theta+\frac{1}{2}(1+\eta^{-2})}} \tag{30}$$

$$\geq \cos\theta\sqrt{1+\frac{\eta^2}{\frac{2}{3}+\eta^2}}, \tag{31}$$

where the last inequality is established since we assumed $\theta \geq \pi/3$. □

A.4 Proof of Theorem 3

Before we prove Theorem 3, we state two lemmas that are essential in our proof. Let all the symbols be as defined in Section 2. Recall that

$$S = \mathbb{E}_{\substack{\alpha_1\sim\mathcal{N}(0,1)\\ y\sim\mathcal{N}(0,\sigma_2^2)}}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)+\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right]$$

$$R = (\sigma^2+||\boldsymbol{\beta}^*||^2)\mathbb{E}_{\substack{\alpha_1\sim\mathcal{N}(0,1)\\ y\sim\mathcal{N}(0,\sigma_2^2)}}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right].$$

Lemma 7. $1-\left(\sqrt{1+\frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1,b_1^*)b_1^*}{\sigma_2^2}}\right)^{-1} \leq S \leq 1.$

Proof. From equation (27) in proof of lemma 1, we get

$$S = \mathbb{E}\left[\alpha_1^2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)-\frac{b_1 b_1^*}{\sigma^2}\alpha_1^2\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right]$$

$$\leq \mathbb{E}\left[\alpha_1^2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right] \leq \mathbb{E}[\alpha_1^2] = 1,$$

where we used $\tanh'(x) \geq 0$ and $\tanh(x) \leq 1$ for any x.

For the lower bound of S, we can apply the lemmas 1, 2 from [8].

Lemma 1 in [8] *Let* $\alpha, \beta \geq 0$ *and* $X \sim \mathcal{N}(\alpha, \sigma^2)$, *then* $\mathbb{E}[\tanh'(\beta X/\sigma^2)X] \geq 0$.

Lemma 2 in [8] *Let* $\alpha, \beta \geq 0$ *and* $X \sim \mathcal{N}(\alpha, \sigma^2)$, *then* $\mathbb{E}[\tanh(\beta X/\sigma^2)] \geq 1-\exp[-\frac{\min(\alpha,\beta),\alpha}{2\sigma^2}]$.

We can apply these two lemmas by setting $\alpha = \alpha_1 b_1^*$, $\beta = \alpha_1\frac{{b_2^*}^2}{\sigma^2}b_1$ (when $\alpha_1 < 0$, we can get the same result due to the symmetry of the expression in sign). It yields

$$S \geq \mathbb{E}_{\alpha_1}\left[1-\exp\left[-\frac{\alpha_1^2 b_1^*\min(b_1^*,\frac{\sigma_2^2}{\sigma^2}b_1)}{2\sigma_2^2}\right]\right]$$

$$= 1-\frac{1}{\sqrt{1+\frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1,b_1^*)b_1^*}{\sigma_2^2}}}.$$

□

Lemma 8. *b_1' is increasing in b_1. Furthermore, in the limit $b_1 \to \infty$,*

$$\lim_{b_1 \to \infty} b_1' = \frac{2}{\pi}(b_1^* \tan^{-1}(\frac{b_1^*}{\sigma_2}) + \sigma_2). \tag{32}$$

Proof. First, we show that b_1' is increasing in b_1. From (8), differentiate it with respect to b_1 yields

$$\frac{db_1'}{db_1} = \mathbb{E}\left[\tanh'(\frac{b_1\alpha_1}{\sigma^2}y)y^2\alpha_1^2\right] \geq 0. \tag{33}$$

Next, we show the limit value of b_1'. Recall that $b_1' = b_1^* S + R$. Again from Stein's lemma, R can be rewritten as

$$R = \frac{\sigma^2 + ||\boldsymbol{\beta}^*||^2}{\sigma_2^2}\mathbb{E}_{\alpha_1,y}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*)\right)y\alpha_1\right].$$

In the limit $b_1 \to \infty$, tanh function becomes sign function. Therefore,

$$\begin{aligned}
\mathbb{E}_{\alpha_1,y}[\text{sign}(\alpha_1(y + \alpha_1 b_1^*))y\alpha_1] &= \frac{1}{\pi}\int_0^\infty 2\frac{\alpha_1}{\sigma_2}e^{-\frac{\alpha_1^2}{2}}\Big(\int_{\alpha_1\boldsymbol{\beta}_1^*}^\infty ye^{-\frac{y^2}{2\sigma_2^2}}dy\Big)d\alpha_1 \\
&= \frac{2}{\pi}\int_0^\infty \alpha_1\sigma_2 e^{-\frac{\alpha_1^2(b_1^*)^2}{2\sigma_2^2}}e^{-\frac{\alpha_1^2}{2}}d\alpha_1 \\
&= \frac{2}{\pi}\sigma_2/(1 + (b_1^*/\sigma_2)^2),
\end{aligned}$$

$$\therefore \lim_{b_1 \to \infty} R = \frac{2}{\pi}\sigma_2.$$

Now we find a limit value of S. In the limit, $\lim_{c \to \infty} cx \tanh'(cx) = 0$ for all x. Therefore,

$$\begin{aligned}
\lim_{b_1 \to \infty} S = \mathbb{E}[\text{sign}(\alpha_1(y + \alpha_1 b_1^*))] &= \frac{1}{\pi}\int_0^\infty\int_{-\alpha_1 b_1^*}^{\alpha_1 b_1^*} e^{-\frac{y^2}{2\sigma_2^2}}e^{-\frac{\alpha_1^2}{2}} \\
&= \frac{2}{\pi}\int_0^\infty\int_0^{\alpha_1 b_1^*/\sigma_2} e^{-\frac{y^2}{2}}e^{-\frac{\alpha_1^2}{2}} = \frac{2}{\pi}\tan^{-1}(b_1^*/\sigma_2).
\end{aligned}$$

Combining the results, we get the desired lemma. □

Now we are ready to prove Theorem 3.

Theorem 3 (Convergence in Distance)**.** *Assume that $\theta < \pi/8$, and define $\sigma_2^2 = \sigma^2 + b_2^{*2}$. If $b_2^* < \sigma$ or $\frac{\sigma_2^2}{\sigma^2}b_1 < b_1^*$, then*

$$\|\boldsymbol{\beta'} - \boldsymbol{\beta}^*\| \leq \kappa\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \kappa(16\sin^3\theta)\|\boldsymbol{\beta}^*\|\frac{\eta^2}{1 + \eta^2}, \tag{14}$$

where $\kappa = \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)^2}{\sigma_2^2}}\right)^{-1}$.

Otherwise, we get

$$\|\boldsymbol{\beta'} - \boldsymbol{\beta}^*\| \leq 0.6\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{15}$$

Proof of Theorem 3. First, difference in second coordinate is easily bounded.

$$(b_2^* - b_2') = (1-S)b_2^* \leq \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)b_1^*}{\sigma_2^2}}\right)^{-1} b_2^*. \tag{34}$$

We therefore focus on giving a bound for $|b_1' - b_1^*|$.

We start from the following observation. Suppose $b_1 = \frac{\sigma^2}{\sigma_2^2}b_1^*$. From equation (26), we have

$$b_1' = \mathbb{E}_{\alpha_1}[\mathbb{E}_{y\sim\mathcal{N}(\alpha_1 b_1^*, \sigma_2^2)}[\tanh(\frac{\alpha_1 b_1^*}{\sigma_2^2}y)y]\alpha_1] = \mathbb{E}_{\alpha_1}[\alpha_1^2 b_1^*] = b_1^*. \tag{35}$$

Also from Lemma 8, b_1' is increasing in b_1. We will separate the cases based on this point.

1. $b_1 \leq \frac{\sigma^2}{\sigma_2^2}b_1^*$:

$$\begin{aligned} b_1' - \frac{\sigma_2^2}{\sigma^2}b_1 &= \mathbb{E}_{\alpha_1}\left[\alpha_1\left(\mathbb{E}_{y\sim\mathcal{N}(\alpha_1 b_1^*, \sigma_2^2)}\left[\tanh\left(\frac{\alpha_1(\frac{\sigma_2^2}{\sigma^2}b_1)}{\sigma_2^2}y\right)y\right] - \mathbb{E}_{y\sim\mathcal{N}(\alpha_1(\frac{\sigma_2^2}{\sigma^2}b_1), \sigma_2^2)}\left[\tanh\left(\frac{\alpha_1(\frac{\sigma_2^2}{\sigma^2}b_1)}{\sigma_2^2}y\right)y\right]\right)\right] \\ &\overset{(a)}{\geq} (b_1^* - \frac{\sigma_2^2}{\sigma^2}b_1)\mathbb{E}\left[\alpha_1^2 \min_{\mu\in(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)} \frac{\partial}{\partial\mu}\left(\mathbb{E}\left[\tanh\left(\frac{\alpha_1(\frac{\sigma_2^2}{\sigma^2}b_1)}{\sigma_2^2}(y+\mu)\right)(y+\mu)\right]\right)\right] \\ &\overset{(b)}{\geq} (b_1^* - \frac{\sigma_2^2}{\sigma^2}b_1)\mathbb{E}\left[\alpha_1^2\left(1 - \exp\left(-\frac{\alpha_1^2\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)^2}{2\sigma_2^2}\right)\right)\right], \end{aligned}$$

   where in (a) we used mean-value theorem, and in (b) we applied lemma 1, 2 in [8]. In turn, we have

$$b_1^* - b_1' \leq \kappa^3\big(b_1^* - \frac{\sigma_2^2}{\sigma^2}b_1\big) \leq \kappa^3(b_1^* - b_1), \tag{36}$$

   where we have $\kappa = (\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)^2}{\sigma_2^2}})^{-1}$ and plugging the relation $b_1 \leq \frac{\sigma_2^2}{\sigma^2}b_1 \leq b_1^*$ into the above.

   Finally, we have $\left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)b_1^*}{\sigma_2^2}}\right)^{-1} \leq \kappa$. Combining them altogether, we have

$$||\boldsymbol{\beta}^* - \boldsymbol{\beta'}|| \leq \kappa||\boldsymbol{\beta}^* - \beta||.$$

2. $b_1 > \frac{\sigma^2}{\sigma_2^2}b_1^*$, $\sigma > b_2^*$: Following the exactly same procedure above, we have

$$b_1' - b_1^* \leq \kappa^3(\frac{\sigma_2^2}{\sigma^2}b_1 - b_1^*) = \kappa^3(b_1 - b_1^*) + \kappa^3\frac{{b_2^*}^2}{\sigma^2}b_1. \tag{37}$$

   By the condition in this case, $\kappa = \big(\sqrt{1 + \frac{{b_1^*}^2}{\sigma_2^2}}\big)^{-1} = \sqrt{\frac{\sigma^2 + {b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}}$. We divided cases into two parts.

   (i) Suppose $b_1 > 2b_1^*$, or $b_1 < 2(b_1 - b_1^*)$. Then,

$$\begin{aligned} b_1' - b_1^* &\leq \kappa^3(b_1 - b_1^*)(1 + 2\frac{{b_2^*}^2}{\sigma^2}) \\ &= \kappa(b_1 - b_1^*)(\frac{\sigma^2 + {b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2})(1 + \frac{2{b_2^*}^2}{\sigma^2}) \\ &= \kappa\underbrace{(\frac{\sigma^2 + {b_2^*}^2}{\sigma^2 + {b_1^*}^2 + {b_2^*}^2}\frac{\sigma^2 + 2{b_2^*}^2}{\sigma^2})}_{A}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*). \end{aligned}$$

Check if A is less than 1. To see that,

$$\begin{aligned}
&\sigma^2(\sigma^2 + {b_1^*}^2 + {b_2^*}^2) - (\sigma^2 + (b_2^*)^2)(\sigma^2 + 2(b_2^*)^2) \\
&= \sigma^2({b_1^*}^2 - 2{b_2^*}^2) - 2{b_2^*}^4 \overset{(a)}{\geq} \sigma^2({b_1^*}^2 - 4{b_2^*}^2) \overset{(b)}{\geq} 0,
\end{aligned}$$

where (a) comes from $b_2^* < \sigma$ and (b) comes from $\tan\frac{\pi}{8} < 1/2$.

$$\therefore b_1' - b_1^* \leq \kappa(b_1 - b_1^*)$$

(ii) $b_1 < 2b_1^*$. We will assume $b_1 \frac{{b_2^*}^2}{\sigma^2} \geq (\frac{1}{\kappa^2} - 1)(b_1 - b_1^*)$. Otherwise, we can easily get $b_1' - b_1^* \leq \kappa(b_1 - b_1^*)$ similarly by plugging it into equation (37).

$$\begin{aligned}
(b_1' - b_1^*)^2 &\leq \kappa^6(b_1 - b_1^*)^2 + \kappa^6\left(2(\frac{b_2^*}{\sigma})^2 b_1(b_1 - b_1^*) + (\frac{b_2^*}{\sigma})^4 b_1^2\right) \\
&\leq \kappa^6(b_1 - b_1^*)^2 + \kappa^6(\frac{b_2^*}{\sigma})^4 b_1^2\left(2(\frac{\kappa^2}{1-\kappa^2}) + 1\right) \\
&= \kappa^6(b_1 - b_1^*)^2 + \underbrace{\kappa^6(\frac{b_2^*}{\sigma})^4 b_1^2\left(\frac{2\sigma^2 + 2{b_2^*}^2 + {b_1^*}^2}{{b_1^*}^2}\right)}_{B}.
\end{aligned}$$

We bound B. We rearrange terms as below:

$$\begin{aligned}
B &= \kappa^6(\frac{b_2^*}{\sigma})^4 b_1^2\left(\frac{2\sigma^2 + 2{b_2^*}^2 + {b_1^*}^2}{{b_1^*}^2}\right) \\
&= \kappa^2(\frac{b_2^*}{\sigma})^4 b_1^2\left(\frac{2\sigma^2 + 2{b_2^*}^2 + {b_1^*}^2}{{b_1^*}^2}\right)\left(\frac{\sigma^2 + {b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}\right)^2 \\
&= \kappa^2{b_2^*}^4(\frac{b_1^2}{{b_1^*}^2})\left(\frac{2\sigma^2 + 2{b_2^*}^2 + {b_1^*}^2}{\sigma^2 + {b_2^*}^2 + {b_1^*}^2}\right)\left(\frac{(\sigma^2 + {b_2^*}^2)^2}{\sigma^4}\right)\frac{1}{\sigma^2 + ||\boldsymbol{\beta}^*||^2} \\
&\leq \kappa^2{b_2^*}^4 * 4 * 2 * 4 * (\frac{1}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}) \\
&= \kappa^2\frac{32{b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}{b_2^*}^2
\end{aligned}$$

Therefore, we get $(b_1' - b_1^*)^2 \leq \kappa^2(b_1 - b_1^*)^2 + \kappa^2\frac{32{b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}{b_2^*}^2$. Combining it with $(b_2' - b_2^*)^2 \leq \kappa^2(b_2 - b_2^*)^2$ yields

$$||\boldsymbol{\beta}' - \boldsymbol{\beta}^*||^2 \leq \kappa^2||\boldsymbol{\beta} - \boldsymbol{\beta}^*||^2 + \kappa^2\frac{32{b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}{b_2^*}^2$$

Now using $\sqrt{a^2 + b^2} \leq a + \frac{b^2}{2a}$,

$$\begin{aligned}
||\boldsymbol{\beta}' - \boldsymbol{\beta}^*|| &\leq \kappa||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| + \kappa\frac{16{b_2^*}^2}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}\frac{b_2^*}{||\boldsymbol{\beta} - \boldsymbol{\beta}^*||}b_2^* \\
&\leq \kappa||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| + \kappa(16\sin^3\theta)||\boldsymbol{\beta}^*||\frac{\eta^2}{1 + \eta^2},
\end{aligned}$$

where we used $\frac{b_2^*}{||\boldsymbol{\beta} - \boldsymbol{\beta}^*||} \leq 1$.

3. $b_1 > \frac{\sigma^2}{\sigma_2^2} b_1^*$, $\sigma < b_2^*$:

   This condition leads us to a special analysis, a constant rate of contraction in local region with high SNR.

   First note that, $b_1' \geq b_1^*$ and its difference $(b_1' - b_1^*)$ is increasing in b_1. Therefore, invoking lemma 8 yields

$$\begin{aligned} b_1' - b_1^* &\leq \frac{2}{\pi}(\sigma_2 + b_1^* \tan^{-1}(\frac{b_1^*}{\sigma_2})) - b_1^* \\ &\leq \frac{2}{\pi}(\sigma_2 + b_1^* \tan^{-1}(\frac{b_1^*}{b_2^*})) - b_1^* \\ &\leq \frac{2}{\pi}(\sqrt{2} - \theta \cot \theta) b_2^*, \end{aligned}$$

   where we used $\sigma_2^2 = \sigma^2 + {b_2^*}^2 \leq 2{b_2^*}^2$, $\tan^{-1}(\frac{b_1^*}{b_2^*}) = \frac{\pi}{2} - \theta$, and $b_1^* = b_2^* \cot \theta$.

   One can easily check that $\theta \cot \theta$ is decreasing in $[0, \frac{\pi}{2}]$. Therefore, we can further bound it:

$$b_1' - b_1^* \leq \frac{2}{\pi}(\sqrt{2} - \frac{\pi}{8} \cot \frac{\pi}{8}) b_2^* \leq 0.3 b_2^*.$$

   On the other side,

$$\begin{aligned} b_2^* - b_2' = (1 - S) b_2^* &\leq \frac{b_2^*}{\sqrt{1 + (b_1^*/\sigma_2)^2}} \\ &\leq \frac{b_2^*}{\sqrt{1 + \frac{1}{2}(b_1^*/b_2^*)^2}} = \frac{b_2^*}{\sqrt{1 + \frac{\cot^2 \frac{\pi}{8}}{2}}} \leq 0.51 b_2^*. \end{aligned}$$

   Combining the result, we get

$$||\boldsymbol{\beta}' - \boldsymbol{\beta}^*|| \leq 0.6 b_2^* \leq 0.6 ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||,$$

   as claimed.

$\square$

Proof of Corollary 2

Corollary 2. *Assume we start from $\theta_0 < \pi/8$. After t iterations of population EM, there exists some constant $\kappa < 1$ such that,*

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\| < \kappa^t \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + t\kappa^t \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1 + \eta^2}. \tag{18}$$

In particular, the result is satisfied if we take κ to be the maximum among

$$0.6, \ \sqrt{(1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2})^{-1}}, \ \sqrt{1 - \frac{0.8\eta^2}{1 + \eta^2}}. \tag{19}$$

Proof. We first show that κ is only decreasing as iteration goes on. It is enough to show that after one EM iteration, $b_1' \geq \min(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*)$, and b_1^* is increasing as the iteration is going on.

If $\frac{\sigma_2^2}{\sigma^2} b_1$ is larger than b_1^*, b_1' becomes larger than b_1^* as we can conclude from Lemma 8 and (35). If $\frac{\sigma_2^2}{\sigma^2} b_1$ were less than b_1^*, then the corresponding $\frac{\sigma_2^2}{\sigma^2} b_1$ at the next iteration is larger than it, as it is inferred from (36). The fact that $b_1^* = ||\boldsymbol{\beta}^*|| \cos\theta_t$ is increasing is obvious from the fact that angle is always decreasing.

Now we will fix κ, the contraction rate at the first iteration. We compare the following quantities:

$$0.6, \left(\sqrt{1 + \frac{2b_1^{*2}}{\sigma^2 + ||\boldsymbol{\beta}^*||^2}}\right)^{-3}, \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*)^2}{\sigma_2^2}}\right)^{-1}.$$

each of which can be rewritten as

$$0.6, \left(\sqrt{1 + \frac{2\eta^2\cos^2\theta_0}{1+\eta^2}}\right)^{-3}, \left(\sqrt{1 + (1+\eta^2\sin^2\theta_0)\frac{||\boldsymbol{\beta}_0||^2}{\sigma^2}}\right)^{-1}, \left(\sqrt{1 + \frac{\eta^2\cos^2\theta_0}{1+\eta^2\sin^2\theta_0}}\right)^{-1}.$$

Since we start from $\theta_0 < \pi/8$, we can plug $\theta_0 = \pi/8$ above and simplify the candidates as (19). We will pick the maximum among these values and fix κ.

Next, we rewrite the equation now with subscript t on each variable:

$$\begin{aligned}
||\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*|| &\leq \kappa||\boldsymbol{\beta}_t - \boldsymbol{\beta}^*|| + \kappa(16\sin^3\theta_t)\frac{\eta^2}{1+\eta^2} \\
&\leq \kappa^2||\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}^*|| + 2\kappa^2(16\sin^3\theta_{t-1})\frac{\eta^2}{1+\eta^2} \\
&... \\
&\leq \kappa^t||\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*|| + t\kappa^t(16\sin^3\theta_0)\frac{\eta^2}{1+\eta^2} \\
&\leq \kappa^t||\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*|| + t\kappa^t\frac{\eta^2}{1+\eta^2},
\end{aligned}$$

where for the last inequality, we used $\theta_0 < \pi/8$. □

Appendix B Proofs for Finite-Sample Based EM

Throughout this section, we use the concentration results that with probability $1 - \delta/T$ in each EM iteration, $||\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'|| \leq \epsilon$ from [4] as well as Theorem 4.

Now we are ready to prove lemmas on finite-sample based EM in three convergence phases.

Proof of Lemma 4

Lemma 4 (Increasing Cosine in Finite-Sample EM)**.** *Let $\boldsymbol{\beta}_0$ be the initial guess and θ_t be the angle formed by $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^*$ and assume $\|\boldsymbol{\beta}_0\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$. We run sample-splitting sample-based EM algorithm, each step with $n/T = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ samples, $T = O(\max(1, \eta^{-2})\log d)$ iterations, and $\epsilon = \epsilon_1 \min(1, \eta^2)$. We also take $\epsilon_1 > 0$ small enough such that there exists a constant $\kappa = (1 - 10\epsilon)\sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}} > 1$. As long as $\theta_t > \pi/3$ for $t \leq T$, with high probability,*

$$\cos\theta_T \geq \kappa^\top \cos\theta_0 - \frac{\kappa^\top - 1}{\kappa - 1} O(\frac{\epsilon}{\sqrt{d}}). \qquad (23)$$

In particular, when $\cos\theta_0 = \Theta(\frac{1}{\sqrt{d}})$, we get $\cos\theta_T \geq \frac{1}{2} - O(\epsilon_1)$.

Proof. From equation (21) with sufficiently small ϵ, we have

$$\begin{aligned}\cos\theta_T &\geq \kappa\cos\theta_{T-1} - O(\frac{\epsilon}{\sqrt{d}})\\ &\geq \kappa^2\cos\theta_{T-2} - (1+\kappa)O(\frac{\epsilon}{\sqrt{d}})\\ &...\\ &\geq \kappa^T\cos\theta_0 - (1+\kappa+\kappa^2+...+\kappa^{T-1})O(\frac{\epsilon}{\sqrt{d}})\\ &\geq \kappa^T\cos\theta_0 - \frac{\kappa^T-1}{\kappa-1}O(\frac{\epsilon}{\sqrt{d}}),\end{aligned}$$

where each inequality holds with probability at least $1-\delta/T$, and all inequalities hold with probability $1-\delta$ by taking a union bound. □

Proof of lemma 5

Lemma 5 (Convergence of Sine in Finite-Sample EM)**.** *Suppose we get a $\boldsymbol{\beta}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/3$ from previous phase. We run sample-splitting sample-based EM with sample complexity $n/T = \tilde{O}(\max(1,\eta^{-2})d/\epsilon^2)$ and error $\epsilon = \epsilon_1\min(1,\eta^2)$. Then with high probability, with a constant $\kappa = \big(\sqrt{1+\frac{0.5\eta^2}{1+\eta^2}}\big)^{-1} < 1$,*

$$\sin^2\theta_T \leq \kappa^{2T}\sin^2\theta_0 + O(\epsilon_1). \tag{24}$$

After $T = O(\max(1,\eta^{-2}))$ iterations, we have $\sin^2\theta_T \leq \sin^2\frac{\pi}{8} + O(\epsilon_1)$.

Proof. Similarly,

$$\begin{aligned}\sin^2\theta_T &\leq \kappa^2\sin^2\theta_{T-1} + O(\epsilon)\\ &\leq \kappa^4\sin^2\theta_{T-2} + (1+\kappa^2)O(\epsilon)\\ &...\\ &\leq \kappa^{2T}\sin^2\theta_0 + (1+\kappa^2+\kappa^4+...+\kappa^{2(T-1)})O(\epsilon)\\ &\leq \kappa^{2T}\sin^2\theta_0 + \frac{1}{1-\kappa^2}O(\epsilon),\end{aligned}$$

with probability $1-\delta$.

Finally,

$$\frac{1}{1-\kappa^2}O(\epsilon) = \frac{\min(1,\eta^2)}{1-\kappa^2}O(\epsilon_1) = \min(1,\eta^2)\frac{1+1.5\eta^2}{0.5\eta^2}O(\epsilon_1) = O(\epsilon_1),$$

which yields the desired result. □

Proof of Lemma 6

Lemma 6 (Convergence of Distance in Finite-Sample EM)**.** *Suppose we get $\boldsymbol{\beta}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/8$ from previous phase. We run sample-splitting EM with $n/T = \tilde{O}(\max(1,\eta^{-2})d/\epsilon^2)$ and $\epsilon = \epsilon_1\min(1,\eta^2)$, getting*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| \leq \kappa^\top\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + T\kappa^\top\frac{\eta^2}{1+\eta^2} + O(\epsilon_1), \tag{25}$$

where κ is the maximum among (19) as in Corollary 2.

After $T = O(\max(1,\eta^{-2})\log(1/\epsilon_1))$ iterations, we get $O(\epsilon_1)$ optimal error.

Proof. We start from Theorem 3. Note that the chosen κ satisfies $\sin^3\theta_t \le \kappa^t \sin^3\theta_0 + \frac{1}{1-\kappa}O(\epsilon)$, which can shown similarly as Lemma 5.

$$\begin{aligned}
||\boldsymbol{\beta}_T - \boldsymbol{\beta}^*|| &\le \kappa||\boldsymbol{\beta}_{T-1} - \boldsymbol{\beta}^*|| + O(\epsilon) + \kappa(16\sin^3\theta_{T-1})\frac{\eta^2}{1+\eta^2} \\
&\le \kappa^2||\boldsymbol{\beta}_{T-2} - \boldsymbol{\beta}^*|| + (1+\kappa)O(\epsilon) + \frac{16\eta^2}{1+\eta^2}(\kappa^2\sin^3\theta_{T-2} + \kappa\sin^3\theta_{T-1}) \\
&\ldots \\
&\le \kappa^T||\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*|| + \frac{1}{1-\kappa}O(\epsilon) + \frac{16\eta^2}{1+\eta^2}(\kappa^T\sin^3\theta_0 + \kappa^{T-1}\sin^3\theta_1 + \ldots + \kappa\sin^3\theta_{T-1}) \\
&\le \kappa^T||\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*|| + \frac{1}{1-\kappa}O(\epsilon) + \frac{16\eta^2}{1+\eta^2}(T\kappa^T\sin^3\theta_0 + \frac{\kappa+\kappa^2+\ldots+\kappa^T}{1-\kappa}O(\epsilon)) \\
&\le \kappa^T||\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*|| + \frac{1}{1-\kappa}O(\epsilon) + T\kappa^T\frac{\eta^2}{1+\eta^2} + \frac{16\eta^2}{1+\eta^2}\frac{1}{(1-\kappa)^2}O(\epsilon) \\
&= \kappa^T||\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*|| + T\kappa^T\frac{\eta^2}{1+\eta^2} + \frac{1}{1-\kappa}O(\epsilon) + \frac{1}{(1-\kappa)^2}\frac{\eta^2}{1+\eta^2}O(\epsilon).
\end{aligned}$$

Finally, check that $1-\kappa$ is $O(\min(1,\eta^2))$. Then the statistical error is in $O(\epsilon_1)$, as desired. □

Appendix C Proofs for Auxiliary Results

C.1 Proof of Lemma 2

Lemma 2. *For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have*

$$||\boldsymbol{\beta}'|| \le 3\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2}. \tag{13}$$

Proof. From Lemma 8, we know that $b_1' \le b_1^* + \frac{2}{\pi}\sqrt{\sigma^2 + b_2^{*2}}$. On the other side, from lemma 7 we have $b_2' \le b_2^*$. Therefore,

$$\begin{aligned}
b_1' &\le b_1^* + \frac{2}{\pi}\sqrt{\sigma^2 + b_2^{*2}} \\
&\le ||\boldsymbol{\beta}^*|| + \frac{2}{\pi}\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2} \\
&\le 2\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2}, \\
b_2' &\le ||\boldsymbol{\beta}^*||.
\end{aligned}$$

Combining the bound for each, we get $||\boldsymbol{\beta}'|| \le 3\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2}$. □

C.2 Proof of Lemma 3

Lemma 3. *If $\|\boldsymbol{\beta}\| \ge \frac{\|\boldsymbol{\beta}^*\|}{10}$, then after one finite-sample EM update with $n = O(\max(1, poly(\eta^{-2}))$ $(d/\epsilon^2))$ samples, $\|\tilde{\boldsymbol{\beta}}'\| \ge \frac{\|\boldsymbol{\beta}^*\|}{10}$.*

Proof. We divide the cases by varying θ. Note that n is now proportional to $poly(\eta^{-2})$, and we control the number of samples so that statistical error in norm is $||\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'|| \le O(\epsilon)\min(1,\eta^2)$. We first show that population EM operator $||\boldsymbol{\beta}'||$ is larger enough than $\frac{||\boldsymbol{\beta}^*||}{10}$, therefore $||\boldsymbol{\beta}'|| - ||\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'||$ is greater than $\frac{||\boldsymbol{\beta}^*||}{10}$.

$\cos\theta \geq 0.2, \sin\theta \geq 0.2$: Suppose $||\boldsymbol{\beta}|| \geq \frac{||\boldsymbol{\beta}^*||}{10}$. If $\cos\theta \geq 0.2$ or $b_1^* \geq \frac{||\boldsymbol{\beta}^*||}{5}$, then as shown in the proof of Corollary 2, $||\boldsymbol{\beta}'|| \geq \min(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*) \geq \min((1+\eta^2\sin^2\theta)\frac{||\boldsymbol{\beta}^*||}{10}, 0.2||\boldsymbol{\beta}^*||)$. We take small enough ϵ, we have $||\tilde{\boldsymbol{\beta}}'|| \geq ||\boldsymbol{\beta}'|| - \epsilon \geq \frac{||\boldsymbol{\beta}^*||}{10}$.

$\cos\theta \leq 0.2$: Recall that $||\boldsymbol{\beta}'|| \geq b_1' = \mathbb{E}[\tanh(\frac{b_1\alpha_1}{\sigma^2}(\alpha_1 b_1^* + y))(\alpha_1 b_1^* + y)\alpha_1]$, where $\alpha_1 \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(0, \sigma_2^2)$. We first claim that $b_1' \geq \mathbb{E}[\tanh\left(\frac{b_1}{\sigma^2}\alpha_1 y\right)\alpha_1 y]$, *i.e.*, lower bounded by setting $b_1^* = 0$. In order to show that, we differentiate b_1' with respect to b_1^*, which yields

$$\mathbb{E}[\alpha_1^2 \tanh(\frac{b_1\alpha_1}{\sigma^2}(\alpha_1 b_1^* + y))] + \mathbb{E}[\frac{\alpha_1^3 b_1}{\sigma^2}(\alpha_1 b_1^* + y)\tanh'(\frac{b_1\alpha_1}{\sigma^2}(\alpha_1 b_1^* + y))].$$

However,

$$\begin{aligned}&\mathbb{E}[\alpha_1^2 \tanh(\frac{b_1\alpha_1}{\sigma^2}(\alpha_1 b_1^* + y))] = \\ &\qquad \frac{1}{\pi\sigma_2}\int_0^\infty \alpha_1^2 e^{-\alpha_1^2/2}\int_0^\infty \tanh(\frac{b_1\alpha_1}{\sigma^2}y)(e^{-\frac{(y-\alpha_1 b_1^*)^2}{2\sigma_2^2}} - e^{-\frac{(y+\alpha_1 b_1^*)^2}{2\sigma_2^2}})dyd\alpha_1 \geq 0.\end{aligned}$$

Simiarly,

$$\begin{aligned}&\mathbb{E}[\frac{\alpha_1^3 b_1}{\sigma^2}(\alpha_1 b_1^* + y)\tanh'(\frac{b_1\alpha_1}{\sigma^2}(\alpha_1 b_1^* + y))] = \\ &\frac{1}{\pi\sigma_2}\int_0^\infty \frac{\alpha_1^3 b_1}{\sigma^2}e^{-\alpha_1^2/2}\int_0^\infty y\tanh'(\frac{b_1\alpha_1}{\sigma^2}y)(e^{-\frac{(y-\alpha_1 b_1^*)^2}{2\sigma_2^2}} - e^{-\frac{(y+\alpha_1 b_1^*)^2}{2\sigma_2^2}})dyd\alpha_1 \geq 0.\end{aligned}$$

Now it becomes clear that b_1' is increasing in b_1^*, thus the claim is verified.

Next, we bound $\mathbb{E}[\tanh(\frac{b_1}{\sigma^2}\alpha_1 y)\alpha_1 y]$.

$$\begin{aligned}\mathbb{E}[\tanh(\frac{b_1}{\sigma^2}\alpha_1 y)\alpha_1 y] &= \frac{2}{\pi\sigma_2}\int_0^\infty\int_0^\infty \alpha_1 y\tanh(\frac{b_1}{\sigma^2}\alpha_1 y)e^{-\frac{y^2}{2\sigma_2^2}}e^{-\frac{\alpha_1^2}{2}}d\alpha_1 dy \\ &= \frac{2}{\pi}\sigma_2\int_0^\infty\int_0^\infty \alpha_1 y\tanh(\frac{b_1}{\sigma^2}\sigma_2\alpha_1 y)e^{-\frac{y^2}{2}}e^{-\frac{\alpha_1^2}{2}}d\alpha_1 dy \\ &\geq \frac{2}{\pi}\sigma_2\int_0^\infty\int_0^\infty \alpha_1 y\tanh(\frac{b_1}{\sigma}\alpha_1 y)e^{-\frac{y^2}{2}}e^{-\frac{\alpha_1^2}{2}}d\alpha_1 dy.\end{aligned}$$

Now suppose if $\frac{b_1}{\sigma} \geq \frac{1}{2}$. We can get a numerical result for the integration

$$\int_0^\infty\int_0^\infty xy\tanh(\frac{1}{2}xy)e^{-\frac{y^2}{2}}e^{-\frac{x^2}{2}}dxdy,$$

which is greater than 0.5. Thus we can conclude $b_1' \geq \frac{1}{\pi}\sigma_2 \geq \frac{1}{\pi}b_2^*$, which is much greater than $||\boldsymbol{\beta}^*||/10$ when $\sin\theta \geq \sqrt{1-0.2^2}$.

If $\frac{b_1}{\sigma}$ is less than 1/2, then we use the Taylor bound for $\tanh(x) \geq x - \frac{x^3}{3}$ to get

$$\begin{aligned}&\frac{2}{\pi}\sigma_2\int_0^\infty\int_0^\infty \alpha_1 y\tanh(\frac{b_1}{\sigma}\alpha_1 y)e^{-\frac{y^2}{2}}e^{-\frac{\alpha_1^2}{2}}d\alpha_1 dy \\ &\quad \geq \frac{2}{\pi}\sigma_2\int_0^\infty\int_0^\infty \alpha_1 y(\frac{b_1}{\sigma}\alpha_1 y - \frac{1}{3}(\frac{b_1}{\sigma}\alpha_1 y)^3)e^{-\frac{y^2}{2}}e^{-\frac{\alpha_1^2}{2}}d\alpha_1 dy \\ &\quad = b_1\frac{\sigma_2}{\sigma}(1 - 3\frac{b_1^2}{\sigma^2}) \geq b_1\sqrt{1 + \frac{24}{25}\eta^2}(1 - 3\frac{b_1^2}{\sigma^2}). \qquad (38)\end{aligned}$$

If $\eta = \frac{||\boldsymbol{\beta}^*||}{\sigma} \geq 5$, then since we assumed $\frac{b_1}{\sigma} < 1/2$, we have $b_1\sqrt{1 + \frac{24}{25}\eta^2}(1 - 3\frac{b_1^2}{\sigma^2}) \geq \frac{5}{4}b_1$. Otherwise, suppose $b_1 = ||\boldsymbol{\beta}^*||/10$, then we have $b_1' \geq b_1\sqrt{1 + \frac{24}{25}\eta^2}(1 - \frac{3}{100}\eta^2)$. When $1 \leq \eta \leq 5$, we have $b_1' \geq \frac{5}{4}b_1$. When $0 \leq \eta \leq 1$, we have $b_1' \geq b_1(1 + 0.3\eta^2)$. Since by (33) we know b_1' is increasing as b_1 increases, and $||\boldsymbol{\beta}'|| \geq b_1'$. Therefore, we conclude that sufficiently ϵ guarantees $||\tilde{\boldsymbol{\beta}}'|| \geq \frac{||\boldsymbol{\beta}^*||}{10}$.

$\sin\theta \leq 0.2$: Assume $b_1 = \frac{||\boldsymbol{\beta}^*||}{10} < \frac{\sigma^2}{\sigma_2^2}b_1^*$. Otherwise we can do as in the first case. From equation (36), we have

$$b_1' \geq b_1 + (1 - \kappa^3)(b_1^* - b_1),$$

where $\kappa = (\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)^2}{\sigma_2^2}})^{-1} \geq \sqrt{1 + \frac{b_1^2}{\sigma^2}}^{-1}$. Since $b_1^* - b_1 \geq \frac{||\boldsymbol{\beta}^*||}{2}$ in this case, we have $b_1' \geq b_1 + \frac{\eta^2}{100+\eta^2}\frac{||\boldsymbol{\beta}^*||}{2}$. Similarly as in other cases, since b_1' is increasing in $b_1 = ||\beta||$, with sufficiently small ϵ we have $||\tilde{\boldsymbol{\beta}}'|| \geq \frac{||\boldsymbol{\beta}^*||}{10}$ whenever $||\beta|| \geq \frac{||\boldsymbol{\beta}^*||}{10}$. □

C.3 Proof of Theorem 4

Theorem 4. *Consider one iteration of sample-based EM algorithm. There exist absolute constants $c_1, c_2 > 0$, such that statistical error in a fixed direction $\boldsymbol{\beta}^*$ can be bounded with probability at least $1 - \delta$, by*

$$|(\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^\top \boldsymbol{\beta}^*| \leq \sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\left(c_1\sqrt{\frac{1}{n}\log(1/\delta)} + c_2\frac{d}{n}\log(1/\delta)\right). \tag{20}$$

Proof. The error for which we are interested in giving a bound is

$$\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}' = \underbrace{(\frac{1}{n}\sum_{i=1}^n x_i x_i^T)^{-1}}_{\hat{\Sigma}^{-1}}\underbrace{(\frac{1}{n}\sum_{i=1}^n(2w_\beta(x_i, y_i) - 1)y_i x_i)}_{\hat{\mu}} - \underbrace{2\mathbb{E}[w_\beta(x, y)yx]}_{\mu}, \tag{39}$$

where w_β is defined in equation (5). Now we fix some $v \in R^d$ such that $||v|| = 1$, and give a bound for $|(\tilde{\boldsymbol{\beta}}' - \beta)^T v|$. First observe that,

$$\begin{aligned}
|(\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}')^T v| &= |(\hat{\Sigma}^{-1}\hat{\mu} - \mu)^T v| \\
&= |(\hat{\mu} - \mu)^T v + \mu^T(\hat{\Sigma}^{-1} - I)v + (\hat{\mu} - \mu)^T(\hat{\Sigma}^{-1} - I)v| \\
&\leq |\underbrace{(\hat{\mu} - \mu)^T v}_{A}| + |\underbrace{\mu^T(\hat{\Sigma}^{-1} - I)v}_{B}| + |\underbrace{(\hat{\mu} - \mu)^T(\hat{\Sigma}^{-1} - I)v}_{C}|.
\end{aligned}$$

We will bound A, B and C separately. For simplicity, we will assume the problem is normalized, *i.e.*, $||\boldsymbol{\beta}^*|| = 1$.

Bounding A: We follow the same procedure that is in [4] with slight refinement in the argument. The main difference is not taking an union bound over the covering set at the end of the stage, since we are only interested in one fixed direction.

We start from symmetrization argument. For symmetrization, we first introduce (x_i', y_i'), independent copy of (x_i, y_i). Then,

$$\begin{aligned}
&\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle - \sum_{i=1}^n(2w_\beta(x'_i,y'_i)-1)y'_i\langle x'_i,v\rangle| \geq t/2\big)\\
&\qquad \geq \mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle - \langle \mu,v\rangle| \geq t, |\frac{1}{n}\sum_{i=1}^n(2w_\beta(x'_i,y'_i)-1)y'_i\langle x'_i,v\rangle - \langle \mu,v\rangle| \leq t/2\big)\\
&\qquad = \mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle - \langle \mu,v\rangle| \geq t\big)\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x'_i,y'_i)-1)y'_i\langle x'_i,v\rangle - \langle \mu,v\rangle| \leq t/2\big)\\
&\qquad \geq \frac{1}{2}\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle - \langle \mu,v\rangle| \geq t\big),
\end{aligned}$$

where the first inequality comes from the fact that condition in the second line implies one in the first line. For the last inequality we used Chevyshev's inequality as the following:

$$\begin{aligned}
\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x'_i,y'_i)-1)y'_i\langle x'_i,v\rangle - \langle \mu,v\rangle| \geq t/2\big) &\leq \frac{\mathrm{Var}((2w_\beta(x'_i,y'_i)-1)y'_i\langle x'_i,v\rangle)}{nt^2/4}\\
&\leq \frac{\mathbb{E}[(2w_\beta(x'_i,y'_i)-1)^2(y'\langle x',v\rangle)^2]}{nt^2/4} \leq \frac{\mathbb{E}[(\sigma^2+\langle x',\boldsymbol{\beta}^*\rangle^2)\langle x',v\rangle^2]}{nt^2/4}\\
&\leq \frac{(\sigma^2+3||\boldsymbol{\beta}^*||^2)}{nt^2/4} \leq \frac{1}{2},
\end{aligned}$$

where in the last two inequalities we used $\mathbb{E}[\langle x',u\rangle^2\langle x',v\rangle^2] \leq 3||u||^2||v||^2$ from lemma 7 in [4] and the fact that we will use the number of samples such that $nt^2 > c_3(\sigma^2+||\boldsymbol{\beta}^*||^2)$ for some c_3.

Then, we introduce Rademacher variables $\{\varepsilon_i\}$ to conclude that

$$\begin{aligned}
&\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle - \frac{1}{n}\sum_{i=1}^n(2w_\beta(x'_i,y'_i)-1)y'_i\langle x'_i,v\rangle| \geq t\big)\\
&\qquad\qquad\qquad \leq 2\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle| \geq t/2\big),
\end{aligned}$$

which in turn yields $\mathbb{P}(A \geq 4t) \leq 4\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle| \geq t\big)$.

Now let us define two events $E_1 = \{\frac{1}{n}\sum_{i=1}^n\langle x_i,v\rangle^2 \leq 2\}$ and $E_2 = \{\frac{1}{n}\sum_{i=1}^n\varepsilon_i(\langle x_i,\boldsymbol{\beta}^*\rangle\langle x_i,v\rangle - \langle\boldsymbol{\beta}^*,v\rangle) \leq t/2\}$. Let $E = E_1 \cap E_2$. Then,

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle \geq t\big) \leq \mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle \geq t|E\big) + P(E^c)$$

In order to apply Chernoff's bound, we bound the expectation given the event.

$$\mathbb{E}\big[\exp\big(\frac{\lambda}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i,y_i)-1)y_i\langle x_i,v\rangle\big)|E\big] \leq \mathbb{E}\big[\exp\big(\frac{\lambda}{n}\sum_{i=1}^n\varepsilon_i y_i\langle x_i,v\rangle\big)|E\big],$$

where we used the fact that $e^a + e^{-a} \leq e^b + e^{-b}$ whenever $|a| < |b|$, and $|2w_\beta(x_i,y_i)-1| < 1$. Conditioned on x_i, $y_i \sim \varepsilon'_i\langle x_i,\boldsymbol{\beta}^*\rangle + w_i$, where $w_i \sim \mathcal{N}(0,\sigma^2)$ and ε'_i is a latent variable taking value

+1 or −1 with equal probability. Then,

$$\begin{aligned}
\mathbb{E}\big[\exp\big(\frac{\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i\langle x_i, v\rangle\big)|E\big] &= \mathbb{E}\big[\exp\big(\frac{\lambda}{n}\sum_{i=1}^n(\varepsilon_i\langle x_i, \boldsymbol{\beta}^*\rangle\langle x_i, v\rangle + \varepsilon_i w_i\langle x_i, \boldsymbol{\beta}^*\rangle)\big)|E\big] \\
&\le \{\mathbb{E}\big[\exp\big(\frac{2\lambda}{n}\sum_{i=1}^n(\varepsilon_i\langle x_i\boldsymbol{\beta}^*\rangle\langle x_i, v\rangle)|E\big]\ \mathbb{E}\big[\exp\big(\frac{2\lambda}{n}\sum_{i=1}^n\varepsilon_i w_i\langle x_i, v\rangle\big)|E\big]\}^{1/2} \\
&\le \{\exp(\lambda t)\ \mathbb{E}\big[\exp\big(\frac{2\lambda}{n}\sum_{i=1}^n\varepsilon_i\langle\boldsymbol{\beta}^*, v\rangle\big)|E\big]\ \mathbb{E}\big[\exp\big(\frac{2\lambda^2}{n^2}\sigma^2\sum_{i=1}^n\langle x_i, v\rangle^2\big)|E\big]\}^{1/2} \\
&\le \{\exp\big(\lambda t + \frac{2\lambda^2}{n}\langle\boldsymbol{\beta}^*, v\rangle^2 + \frac{4\lambda^2}{n}\sigma^2\big)\}^{1/2} \\
&\le \exp\big(\frac{\lambda t}{2} + \frac{2\lambda^2}{n}(\sigma^2 + ||\boldsymbol{\beta}^*||^2)\big),
\end{aligned}$$

where we used Cauchy-Schwartz inequality first, then Gaussian-tail bound, and sub-Gaussian bound for rademacher variables and definition of E. Using this, we use the Chernoff bound to get

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i, y_i) - 1)y_i\langle x_i, v\rangle \ge t|E\big) \le \exp\big(\frac{2\lambda^2}{n}(||\boldsymbol{\beta}^*||^2 + \sigma^2) - \frac{\lambda t}{2}\big),$$

from which optimal choice of λ gives

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(2w_\beta(x_i, y_i) - 1)y_i\langle x_i, v\rangle \ge t|E\big) \le \exp\big(-\frac{nt^2}{32(||\boldsymbol{\beta}^*||^2 + \sigma^2)}\big).$$

Now we are left with bounding $P(E^c)$. As it is less than $P(E_1^c) + P(E_2^c)$, we are bounding probability of these two events. As each $\langle x_i, v\rangle^2$ is a Chi-square random variable, using the sub-exponential tail bound, we have $P(E_1^c) \le e^{-n/8}$. For bounding probability of E_2^c, we first introduce two random variables $Z_1 = \langle x_i, u\rangle$ and $Z_2 = \langle x_i, v\rangle$ where $u = \boldsymbol{\beta}^*/||\boldsymbol{\beta}^*||$. Note that Z_1, Z_2 is jointly Gaussian with zero-mean and covariance $\left[\begin{smallmatrix} 1 & \langle u,v\rangle \\ \langle u,v\rangle & 1\end{smallmatrix}\right]$. We are bounding

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n\varepsilon_i(Z_1Z_2 - \langle u, v\rangle) > \frac{t}{2||\boldsymbol{\beta}^*||}) \le \frac{\mathbb{E}[\exp(\lambda\varepsilon(Z_1Z_2 - \langle u, v\rangle)]^n}{e^{\lambda nt/2||\boldsymbol{\beta}^*||}}$$

First observe that $Z_2|Z_1 \sim \mathcal{N}(\langle u, v\rangle Z_1, 1 - \langle u, v\rangle^2)$. Using it to get an expectation over Z_2 conditioned on Z_1, we get

$$\mathbb{E}[\exp(\lambda\varepsilon Z_1Z_2)] = \mathbb{E}[\exp(\frac{\lambda^2}{2}(1 - \langle u, v\rangle^2) + \lambda\varepsilon\langle u, v\rangle)Z_1^2].$$

Since Z_1^2 is Chi-square variable, which is sub-exponential. Thus,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda\varepsilon(Z_1Z_2 - \langle u, v\rangle))] &= \mathbb{E}\left[\exp\left(\frac{\lambda^2}{2}(1 - \langle u, v\rangle^2) + \lambda\varepsilon\langle u, v\rangle\right)(Z_1^2 - 1)\right]\exp\left(\frac{\lambda^2}{2}(1 - \langle u, v\rangle^2)\right) \\
&\le \exp\left(2\left(\frac{\lambda^2}{2}(1 - \langle u, v\rangle^2) + \lambda\varepsilon\langle u, v\rangle\right)^2\right)\exp\left(\frac{\lambda^2}{2}(1 - \langle u, v\rangle^2)\right) \\
&\le \exp(3\lambda^2),
\end{aligned}$$

where we first used sub-exponential tail bound for Chi-square distribution, used $a + \frac{1-a^2}{2} < 1$ for $a < 1$, and the fact that we will use small enough $\lambda < 1$. Finally, with optimal choice of λ, we have

$$\mathbb{P}(E_2^c) \le \exp(-\frac{nt^2}{48||\boldsymbol{\beta}^*||^2}).$$

Combining all bounds for the probability, with probability at least $1-\delta$, we have $A \leq c_1\sqrt{(\sigma^2 + ||\boldsymbol{\beta}^*||^2)\frac{1}{n}\log(1/\delta)}$ for some universal constant c_1.

Bounding B: Standard results from random matrix theory imply that $||\hat{\Sigma_p} - I||_{op} \leq c_2\sqrt{\frac{d}{n}\log(1/\delta)}$ with high probability. We will consider events under this condition.

Since inverse operator is hard to handle, we modify it using Taylor's expansion

$$\begin{aligned}\hat{\Sigma}^{-1} &= (I - (I - \hat{\Sigma}))^{-1} \\ &= I + (I - \hat{\Sigma}) + (I - \hat{\Sigma})^2 + ...,\end{aligned}$$

which in turn yields $\mu^T(\hat{\Sigma}^{-1} - I)v = \mu^T(I - \hat{\Sigma})v + ||\mu||\tilde{O}(\frac{d}{n})$.

For simplicity, let us define $u = \frac{\mu}{||\mu||}$ and derive a bound for $u^T(I - \hat{\Sigma})v$. Now we are left with bounding $u^T(I - \hat{\Sigma})v = u^T v - \frac{1}{n}\sum_i (x_i^T u)(x_i^T v)$. Let two random variables $Z_1 = X^T u$, $Z_2 = X^T v$. Then Z_1, Z_2 are jointly Gaussian with zero-mean and covariance $\left[\begin{smallmatrix} 1 & \langle u,v\rangle \\ \langle u,v\rangle & 1\end{smallmatrix}\right]$. Then the probability we are to give a concentration bound is

$$\mathbb{P}(\frac{1}{n}\sum_i z_{1,i}z_{2,i} - \langle u,v\rangle \geq t) \leq \frac{\mathbb{E}[\exp(\lambda Z_1 Z_2)]^n}{e^{\lambda n t}e^{n\lambda\langle u,v\rangle}}.$$

Using the same procedure used to bound $\mathbb{P}(E_2^c)$ before, we have

$$\mathbb{P}(\frac{1}{n}\sum_i z_{1,i}z_{2,i} - \langle u,v\rangle \geq t) \leq \exp(-nt^2/12),$$

which gives with high probability, $u^T(I - \hat{\Sigma})v \leq \tilde{O}(\sqrt{\frac{1}{n}})$.

Finally, $||\mu|| \leq O(\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2})$ due to Lemma 2, we have $B \leq c_4\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2}(\sqrt{\frac{1}{n}\log(1/\delta)} + \frac{d}{n})$.

Bounding C: We have $||\hat{\mu} - \mu|| \leq c_5\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2}\sqrt{\frac{d}{n}\log(1/\delta)}$ from [4] with probability at least $1-\delta$, as well as $||\hat{\Sigma}^{-1} - I||_{op} \leq c_2\sqrt{\frac{d}{n}\log(1/\delta)}$ from random matrix theory. Therefore, we get

$$|(\hat{\mu} - \mu)^T(\hat{\Sigma}^{-1} - I)v| \leq ||\hat{\mu} - \mu||\ ||\hat{\Sigma}^{-1} - I||_{op}\ ||v|| \leq \tilde{O}(\sqrt{\sigma^2 + ||\boldsymbol{\beta}^*||^2}\frac{d}{n}).$$

This gives a bound for C.

Finally, combining the bounds on A, B and C with $v = \boldsymbol{\beta}^*$, we get the first part of the theorem. □

Corollary 3. *Suppose that the norm of the estimator $\|\boldsymbol{\beta}\|$ is larger than $\frac{\|\boldsymbol{\beta}^*\|}{10}$. Then, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ samples for one finite-sample based EM iteration, we have*

$$\cos\tilde{\theta}' \geq \kappa(1 - 10\epsilon)\cos\theta - O\left(\max\left(\frac{\epsilon}{\sqrt{d}}, \epsilon^2\right)\right), \tag{21}$$

$$\sin^2\tilde{\theta}' \leq \kappa'\sin^2\theta + O(\epsilon), \tag{22}$$

with $\kappa = \sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \geq 1$, and $\kappa' = (1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta)^{-1} < 1$.

Proof. We start from the end of the proof for Theorem 4. We now replace statistical errors in terms of ϵ using the sample complexity $n = \tilde{O}((1+\eta^{-2})d/\epsilon^2)$. Recall the way we compute cosine,

$$\begin{aligned}
\cos\tilde{\theta}' &= \frac{\langle \tilde{\boldsymbol{\beta}}', \boldsymbol{\beta}^* \rangle}{||\tilde{\boldsymbol{\beta}}'||\ ||\boldsymbol{\beta}^*||} \\
&= \frac{\langle \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle}{||\tilde{\boldsymbol{\beta}}'||\ ||\boldsymbol{\beta}^*||} + \frac{\langle \tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle}{||\tilde{\boldsymbol{\beta}}'||\ ||\boldsymbol{\beta}^*||} \\
&= \cos\theta' \frac{||\boldsymbol{\beta}'||}{||\tilde{\boldsymbol{\beta}}'||} + \frac{\langle \tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle}{||\tilde{\boldsymbol{\beta}}'||\ ||\boldsymbol{\beta}^*||} \\
&\geq \cos\theta' \big(1 - \frac{\epsilon}{||\boldsymbol{\beta}'||/||\boldsymbol{\beta}^*|| + \epsilon}\big) - \max\big(\frac{\epsilon}{\sqrt{d}}, \epsilon^2\big) \frac{||\boldsymbol{\beta}^*||}{||\tilde{\boldsymbol{\beta}}'||} \\
&\geq \cos\theta'(1 - 10\epsilon) - O\big(\max\big(\frac{\epsilon}{\sqrt{d}}, \epsilon^2\big)\big) \\
&\geq \kappa(1 - 10\epsilon)\cos\theta - O\big(\max\big(\frac{\epsilon}{\sqrt{d}}, \epsilon^2\big)\big),
\end{aligned}$$

where the last two inequalities follows from the Lemma 2 and equation (30) in the proof of Theorem 2.

Now for sine, we have that

$$\begin{aligned}
\sin^2\tilde{\theta}' &= 1 - \cos^2\tilde{\theta}' \\
&\leq 1 - \cos^2\theta' + O(\epsilon) \\
&\leq \sin^2\theta' + O(\epsilon) \\
&\leq \kappa' \sin^2\theta + O(\epsilon),
\end{aligned}$$

where the last inequality comes from Theorem 1. □

C.4 Proof of Theorem 6

Theorem 6. *Consider one iteration of Easy-EM algorithm. There exist absolute constants $c_1, c_2 > 0$, such that with probability at least $1 - \delta$,*

$$\begin{aligned}
\|\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}'\| &\leq c_1 \sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2} \sqrt{\frac{d}{n}\log(1/\delta)}, \\
|(\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}')^\top \boldsymbol{\beta}^*| &\leq c_2 \sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2} \sqrt{\frac{1}{n}\log(1/\delta)}.
\end{aligned}$$

Furthermore, suppose that the norm of current estimator $\|\boldsymbol{\beta}\|$ is larger than $\frac{\|\boldsymbol{\beta}^*\|}{10}$. Then, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ samples for one Easy-EM iteration, we have

$$\begin{aligned}
\cos\tilde{\theta}'' &\geq \kappa(1 - 10\epsilon)\cos\theta - O\big(\frac{\epsilon}{\sqrt{d}}\big), \\
\sin^2\tilde{\theta}'' &\leq \kappa' \sin^2\theta + O(\epsilon),
\end{aligned}$$

with $\kappa = \sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \geq 1$, *and* $\kappa' = \big(1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta\big)^{-1} < 1$.

Proof. Bound for $|(\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}')^\top \boldsymbol{\beta}^*|$ directly follows from bounding A in the proof of Theorem 4. For the norm, standard covering set argument tells we can take union bound over 1/2-covering set of unit sphere to bound $P(\sup_{v \in \mathbb{S}^d} |(\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}')^\top v| \geq t)$, from which we can conclude

$$\|\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}'\| \leq c_1 \sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2} \sqrt{\frac{d}{n}\log(1/\delta)},$$

with probability at least $1 - \delta$.

Finally, bound for cosine and sine can be derived by the exactly same procedure used in the proof of Corollary 3. □