Global Optimality of the EM Algorithm for Mixtures of Two Component Linear Regressions

Jeongyeol Kwon[♭*], Wei Qian[†*], Constantine Caramanis[♯],
Yudong Chen[♭], Damek Davis[†], Nhat Ho[♯]

[†]Cornell University
[♯]The University of Texas at Austin
[♭]University of Wisconsin-Madison

[*]Equal Contribution

Abstract

Recent results established that EM enjoys global convergence for Gaussian Mixture Models. For Mixed Linear Regression, however, only local convergence results have been established, and those only for the high signal-to-noise ratio (SNR) regime. In this work, we completely characterize the global optimality of EM: we show that starting from any randomly initialized point, the EM algorithm converges to the true parameter $\boldsymbol{\beta}^*$ at the minimax statistical rates under all SNR regimes. Toward this goal, we first show the global convergence of the EM algorithm at the population level. Then we provide a complete characterization of statistical and computational behaviors of EM under all SNR regimes with finite samples. In particular: (i) When the SNR is sufficiently large, the EM updates converge to the true parameter $\boldsymbol{\beta}^*$ at the standard parametric convergence rate $O((d/n)^{1/2})$ after $O(\log(n/d))$ iterations. (ii) In the regime where the SNR is above $O((d/n)^{1/4})$ and below some constant, the EM iterates converge to a $O(\mathrm{SNR}^{-1}(d/n)^{1/2})$ neighborhood of the true parameter, when the number of iterations is of the order $O(\mathrm{SNR}^{-2}\log(n/d))$. (iii) In the low SNR regime where the SNR is below $O((d/n)^{1/4})$, we show that EM converges to a $O((d/n)^{1/4})$ neighborhood of the true parameters, after $O((n/d)^{1/2})$ iterations. By providing tight convergence guarantees of the EM algorithm in middle-to-low SNR regimes, we reveal that in low SNR, EM changes rate, matching the $n^{-1/4}$ rate of the MLE, a behavior that previous work had been unable to show.

Key words: The EM Algorithm, Latent Variable Model, Mixture of Linear Regression, Global Convergence, Sample Complexity, Minimax Rates

1 Introduction

The expectation-maximization (EM) algorithm is a general-purpose technique for estimating the model parameters in problems with unobserved latent variables [12, 34]. In particular, EM computes successively tighter upper bounds of the negative log-likelihood function in the hope of finding a good minimizer. In general, optimizing the likelihood in the presence of missing data is an intractable problem due to the non-convexity of the negative log-likelihood function. Nevertheless, EM is still widely used in practice due to its simplicity and good empirical performance [16, 25, 7, 23, 3]. Relatively little is understood about the theoretical properties of EM.

Recent work has made progress in deriving theoretical guarantees for EM for several statistical problems. It has been demonstrated that when the Signal-to-Noise Ratio (SNR) is high and certain regularity assumptions hold, EM converges locally if initialized near the global optimum; see, e.g., the work in [39, 1, 17, 40, 41] and the references therein. For the special case of Gaussian Mixture Models (GMM) with two components, Xu et al. [36] and Daskalakis et al. [10] have shown that a two-phase version of EM converges from random initialization. As far as we know, no comparable global convergence result is known for the related problem of Mixed Linear Regression (MLR), despite the empirical success of EM in this setting [11, 15].

[*]Preliminary results from this work were presented in 2019 Conference on Learning Theory (COLT) [20] and 2021 Internal Conference on Artifical Intelligence and Statistics (AISTATS) [21]

The lack of global convergence guarantees for EM under MLR is not simply an oversight. Rather, as we show later, the structures of MLR differ significantly from GMM, even on the population (infinite sample) level; consequently, EM exhibits very different behaviors under these two models. Existing techniques used to analyze EM under GMM—often based on showing contraction in ℓ_2 distance—are fundamentally insufficient for establishing global convergence of EM under MLR. Furthermore, most prior work has studied instances with strong separation (high SNR) and established linear convergence of the EM algorithm with the standard parametric statistical rate $n^{-1/2}$. In contrast, the understanding of the EM algorithm in the weak separation (low SNR) settings, especially mixed linear regression, remains incomplete.

1.1 Basic Setup and the EM Algorithm

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

$$y_i = c_i \langle \boldsymbol{\beta}^*, \boldsymbol{x}_i \rangle + e_i, \qquad i = 1, ..., n, \tag{1}$$

where e_i is the noise term, and $c_i \in \{\pm 1\}$ is the hidden/latent variable denoting whether the i-th data point $(\boldsymbol{x}_i, y_i)$ is generated by $+\boldsymbol{\beta}^*$ or $-\boldsymbol{\beta}^*$. Finding the true parameter $\boldsymbol{\beta}^*$ is known to be NP-hard in general even without noise [40]. Accordingly, a common assumption in the literature stipulates that the covariates and noise terms, $\boldsymbol{x}_i$ and e_i, are sampled independently from Gaussian distributions; that is, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $e_i \sim \mathcal{N}(0, \sigma^2)$, where the noise variance σ^2 is known. We assume, moreover, that the hidden variables $\{c_i\}$ take values ± 1 with equal probability and are independent of each other and of everything else. We define SNR as $\eta := \|\boldsymbol{\beta}^*\|/\sigma$. We assume that η is bounded from above by some (large enough) constant $\rho = O(1)$.

EM is an iterative algorithm for optimizing the likelihood function of a latent variable model. At each iteration, EM performs two steps: the E-step that computes the expectation of the log-likelihood conditioned on the current estimate of $\boldsymbol{\beta}^*$, and the M-step that optimizes this conditional expectation. For MLR, when we plug in the likelihood of the assumed Gaussian distribution and replace the expectation with an empirical average over observed data $\{\boldsymbol{x}_i, y_i\}$, the M-step becomes the weighted least squared loss minimization problem. In this case, the finite-sample-based EM update, given the current estimator $\boldsymbol{\beta}$, has the following closed form expression:

$$\text{(finite-sample EM)} \qquad \widetilde{\boldsymbol{\beta}}' = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \tanh\left(\frac{\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle}{\sigma^2} y_i\right) y_i \boldsymbol{x}_i\right); \tag{2}$$

for a derivation see Balarishnan et al. [1] or Klusowski et al. [17].

The infinite-sample limit of the finite-sample EM, which we call the *population EM*, has the following expression:

$$\text{(population EM)} \qquad \boldsymbol{\beta}' = \mathbb{E}_{\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[\mathbb{E}_{Y | \boldsymbol{X} \sim \mathcal{N}(\langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle, \sigma^2)} \left[\tanh\left(\frac{\langle \boldsymbol{X}, \boldsymbol{\beta} \rangle}{\sigma^2} Y\right) Y\right] \cdot \boldsymbol{X}\right]. \tag{3}$$

The above expression follows from taking the limit $n \to \infty$ in the EM update formula (2) and simplifying the result using the symmetry of the distribution of Y given $\boldsymbol{X}$.

1.2 Main Contributions

In this work, we show that EM for MLR with two components converges *globally* from random initialization. We first establish this result in the infinite sample limit, i.e., for the population version of EM. Along the way, we provide a complete characterization of the landscape of the population likelihood function, by classifying its local maxima, local minima and saddle points. This geometric result implies *non-contraction* in ℓ_2 distance of the EM iterates—in sharp constrast to previous result to GMM—which therefore necessitates a new convergence analysis based on the *angle*.

We then provide a finite sample analysis, starting by coupling the finite-sample version of EM with the population EM. While the ideas remain the same for the middle-to-high SNR regimes, as we see below, finite-sample EM shows a very different behavior from population EM in the low SNR regime. We reveal this transition in statistical and computational behaviors from middle to low SNR regimes that previous analysis had missed. Collecting the results, we provide a complete picture of the EM algorithm under all SNR regimes: we show that EM converges to the true parameter starting from *any* randomly initialized point at known minimax rates [8] in all SNR regimes. We describe our contributions in more details as follows.

1. **Population Analysis for Global Convergence**: Previous work on analyzing the EM algorithm for MLR relies on demonstrating that the ℓ_2 distance between the current iterate and the true solution $\boldsymbol{\beta}^*$, contracts at every iteration provided that the initial distance is already small. Such a contraction, however, cannot hold globally, as the EM update initialized randomly may in fact result in a *larger distance* from $\boldsymbol{\beta}^*$. This phenomenon was pointed out in [17]. Nevertheless, we prove the global convergence from careful observations on the population landscape as described below:

   1.1 **Population Landscape**: We provide a geometric explanation in this paper by showing the existence of saddle points of the log-likelihood function in the direction orthogonal to $\boldsymbol{\beta}^*$. These saddle points prevent a global convergence in ℓ_2 distance of EM (which is equivalent to gradient ascent). On the other hand, we show that $\pm\boldsymbol{\beta}^*$ are the only local maxima, hence suggesting that global convergence can be proved by other means.

   1.2 **Global Convergence via Decreasing Angle**: Instead of proving a global convergence via the ℓ_2 distance, we show that the angle between the iterate and $\boldsymbol{\beta}^*$ is always decreasing (unless we start from an exactly orthogonal vector—a measure zero event). Consequently, EM quickly enters a local region where the current iterate is well aligned with the direction of $\boldsymbol{\beta}^*$. In this local region, we show that a contraction in distance indeed holds. We use this argument to demonstrate that EM converges to $\boldsymbol{\beta}^*$ from *any* randomly initialized point with high probability.

2. **Finite-Sample Analysis and Minimax Rates**: Using our population results, we provide the finite-sample analysis for the EM algorithm. However, unlike in the population case, we show that finite-sample EM shows very different behaviors in different SNR regimes as described below:

   2.1 **High-to-middle SNR regimes**: when $\eta \gtrsim (d/n)^{1/4}$ (up to some logarithmic factor), we show that finite-sample EM converges to $\boldsymbol{\beta}^*$ within a neighborhood of $O(\max\{1, \eta^{-1}\}(d/n)^{1/2})$ after $O(\max\{1, \eta^{-2}\}\log(n/d))$ number of iterations.

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

   For the finite-sample analysis, we focus primarily on two aspects of the EM algorithm: (i) statistical rate, and (ii) computational complexity. In the high SNR regime, we have linear convergence to true parameters within $\sqrt{d/n}$ rate as noted previously in the literature. In contrast, in the low SNR regime when $\eta \lesssim (d/n)^{1/4}$, the statistical rate is $(d/n)^{1/4}$. We explain this transition in statistical rate with a convergence property of the population EM in the middle-to-low SNR regimes. The upper bound on the statistical error given by EM matches the known lower bound for this problem in all SNR regimes [8]. For the computational complexity, the number of iterations increases quadratically in the inverse of SNR until SNR reaches $(d/n)^{1/4}$. One can also observe that the number of iterations is naturally interpolated at SNR $= (d/n)^{1/4}$ from $\eta^{-2}\log(n/d)$ to $\sqrt{n/d}$. This transition in computational complexity could also be of independent interest for other mixture models with small separations. We note that our results do not require sample-splitting (a technique using fresh samples every iteration) which is crucial for getting the sample optimality results in middle-to-low SNR regimes.

In summary, we obtain the following overall guarantee for the finite-sample EM with n samples:

Theorem 1. *Let $\widetilde{\boldsymbol{\beta}}_0$ be a random initial vector in $\mathbb{R}^d$ such that the direction of $\widetilde{\boldsymbol{\beta}}_0$ is randomly sampled from the uniform distribution on the unit sphere. The norm of initial vector can be any non-zero constant such that $\|\widetilde{\boldsymbol{\beta}}_0\| \geq c\sigma(d\log^2(n/\delta)/n)^{1/4}$ and $n > Cd^2$ for some universal constants $c, C > 0$. There exist universal constants $C_1, \ldots, C_5 > 0$ such that the following holds.*

(a) (Middle-to-high SNR regimes) When $\eta \geq C_1(d\log^2(n/\delta)/n)^{1/4}$, with probability at least $1-\delta$, we have

$$\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq C_2\sigma\max\{1, \eta^{-1}\}(d\log^2(n/\delta)/n)^{1/2},$$

after we run the standard EM algorithm (2) for $T = C_3\max\{1, \eta^{-2}\}\log(n/d)$ iterations.

(b) (Low SNR regime) When $\eta \leq C_1(d\log^2(n/\delta)/n)^{1/4}$, with probability at least $1-\delta$, we have

$$\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq C_4\sigma(d\log^2(n/\delta)/n)^{1/4},$$

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

1.3 Key Challenges and Comparison to Existing Approaches for GMM

As mentioned earlier, several recent works have consider the related problem of 2-component GMM and established global convergence of the EM algorithm [36, 10, 13]. Here we highlight the key challenges in our MLR setting as well as the differences between our analysis and those for GMM in prior works, deferring a more detailed discussion to subsequent sections. Additional discussion on related work is provided at the end of this subsection.

Population Analysis. A key difference between MLR and GMM is the presence of the covariates $\boldsymbol{X}$ in the regression setting. Therefore, each observation $(\boldsymbol{X}, y)$ in MLR only provides information along the $\boldsymbol{X}$ direction for the relative position of the current iterate $\boldsymbol{\beta}$ and the true $\boldsymbol{\beta}^*$. This difference has far-reaching consequences: the geometry of the negative log-likelihood function of MLR and the dynamics of the EM algorithm are significantly different from those in the GMM setting.

In particular, unlike GMM, in MLR there is a non-trivial region where the EM iterate does not contract in Euclidean distance to the true parameter. This difference is illustrated in Figure 1: note that for MLR the distance $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ first increases than decreases. Consequently, the approaches in [10, 17], which are based on distance contraction in GMM, do not work in our setting. For MLR, we need consider alternative measures (or potential functions) under which the EM iteration converges quickly. Specifically, we establish angle contraction results in Section 3, and our analysis is divided into 3 phases:

- Phase 1: We start with random initialization, and thus, we start with a small cosine value between the EM iterate $\boldsymbol{\beta}$ and the true $\boldsymbol{\beta}^*$. We show that the cosine value increases at a (constant) linear rate and thus EM escapes the small-angle region in $O(\log d)$ steps.
- Phase 2: Once the cosine value reaches $O(1)$, the sine value becomes a more appropriate potential function, which decays at a (constant) linear rate to 0. Note that the increase rate in the cosine value slows down when it is close to 1.
- Phase 3: Eventually we want to show that iterate $\boldsymbol{\beta}$ linearly converges to $\boldsymbol{\beta}^*$ in ℓ_2-distance, which happens after sufficient angle alignment.

Xu et al. [36] have used similar angle alignment arguments to show convergence of the EM algorithm for GMM. However, they only used the sine value as their potential function, restricting the analysis to the asymptotic regime (in their analysis, the convergence rates have not been explicitly specified). With random initialization, the sine values converges slowly during the first phase. We circumvent the issue by establishing the linear increase in the cosine values during phase 1, showing that EM escapes the initial phase after $O(\log d)$ iterations. The work by Daskalakis et al. [10] has provided a non-asymptotic convergence result for 2-GMM; however, they rely on global ℓ_2-distance convergence, which does not hold in 2-MLR.

Finite-Sample Analysis. Prior work has established the local convergence of EM for both 2-component GMM and 2-compnent MLR in the *high SNR* regime ($\eta = \|\beta^*\|/\sigma > 1$) [40, 10, 17]. To our best knowledge, no prior work has shown the minimax optimality of the EM algorithm in the *middle or low SNR* regimes. Our analysis for the low SNR regime is inspired by the technique developed in Dwivedi et al. [13]. However, they can only address the over-specified settings (i.e., the SNR is $\|\boldsymbol{\beta}^*\|/\sigma = 0$), whereas we extend the applicability of their techniques to show the minimax statistical rates in all SNR regimes. In particular, we explicitly show the transition of statistical rates from high to low SNR regimes $\sigma \max(1, \eta^{-1}) \cdot \sqrt{d/n}$ to $(d/n)^{1/4}$ through the careful analysis of angle concentration and localization, which has not been done in the context of analyzing EM algorithms.

1.3.1 Other Related Work

As mentioned, our knowledge of when EM converges to a true solution is still limited. In general, it is known that the EM algorithm may settle in a bad local optimum [34]. Classical results on convergence were infinitesimally local, and asymptotic [28, 37, 25]. Recent study on the theoretical understanding of EM has been initiated in Balakrishnan et al. [1], which proposed a novel framework to analyze the EM algorithm. Motivated by this work, there has been a line of work that provides local analysis of EM when it starts from a well initialized point [39, 40, 38, 19, 18].

More recent work has provided global analysis for the GMM problem. For the mixture of two Gaussians, Xu et al. [36] and Daskalakis et al. [10] establish guarantee convergence of EM for this specific problem from a random initialization. Extensions to other variants of GMM are considered in the work [27, 26, 2]. For GMM with more components, however, Jin et al. [14] proves that bad local optima exist and randomly initialized EM converges to such a local solution with high probability.

For MLR, only the local convergence of EM has been recently established: when there are two components, the EM algorithm converges to the global optimum if we start from a point sufficiently close to the true parameter in ℓ_2 distance; see, e.g., [40, 39, 41, 1] and the references therein. A better local contraction region was suggested in Klusowski et al. [17], where the convergence is guaranteed inside a region where the angle between the initial solution and the true parameter is small. Still, all known results remain inherently local for MLR, and in particular, are not satisfied by a random initialization, even when a norm bound on the true parameter is known.

Moreover, previous results on MLR are strictly restricted to high SNR regimes, *i.e.,* when $\|\boldsymbol{\beta}^*\|$ is sufficiently larger than σ. In a closely related problem of learning mixtures of two Gaussians, [13] recently studied the EM algorithm in an extreme case of the over-specified models, *i.e.,* there is no separation between two components. However, their analysis is restricted to strictly over-specified settings, and it has not been obvious to extend their result to low SNR models. In another recent work, [35] has studied the EM algorithm for learning a mixture of two weakly-separated location Gaussians, establishing a minimax rate of the EM algorithm after $O(\sqrt{n/d})$ iterations in middle-to-low SNR regimes. However, their result requires the initialization to be already within a small Euclidean ball of $(d/n)^{1/4}$-radius, which is restrictive. Our result does not suffer from small initialization issue as in [35]. Furthermore, our proof strategy can be applied to resolve the open issue with small initialization in [35].

MLR is an interesting problem by itself, for which many algorithms beyond EM have been proposed. The work in Chen et al. [8, 9] developed a lifted convex formulation approach that achieves tight minimax error rates. A good initialization strategy for EM based on Stein's second-order lemma was proposed in Yi et al. [40], though this seems to rely on the noiseless setting which they study. The above two papers have focused on MLR of two components case. Recent work has extended the focus to multiple components. The work in [42, 24] develops gradient descent based algorithms. In parallel, the work in [6, 41, 29] considers algorithms that are based on tensor decomposition of third order moments. EM is an attractive option among these algorithms due to its generality, simplicity and computational efficiency; moreover, EM is often applied to the output of other algorithms to obtain an improved estimate.

1.4 Notations

We establish the notation used throughout the remainder of the paper. We use $\angle(\boldsymbol{u}, \boldsymbol{v})$ to denote the angle between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. The ℓ_2 norm for a vector is denoted by $\|\cdot\|$, and the spectral norm (the largest singular value) of a matrix is denoted by $\|\cdot\|_{\text{op}}$. For two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}$ is the usual inner product between them.

We use $(\boldsymbol{X}, Y)$ as a generic random variable representing the covariate-response pair from the MLR model (1), and use $\{(\boldsymbol{x}_i, y_i)\}$ as independent copies of $(\boldsymbol{X}, Y)$. Due to a symmetry between the regressors $\pm\boldsymbol{\beta}^*$, we focus on the convergence to one of them, say $\boldsymbol{\beta}^*$. We use $\boldsymbol{\beta}_t$ to denote the estimate of $\boldsymbol{\beta}^*$ at the t^{th} iteration of the population EM, and use $\theta_t := \angle(\boldsymbol{\beta}_t, \boldsymbol{\beta}^*)$ to denote the angle formed by $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^*$. When we intend to understand a single iteration of the EM, we drop the subscript t, and use $\boldsymbol{\beta}$ in place of $\boldsymbol{\beta}_t$ for the current iterate and $\boldsymbol{\beta}'$ in place of $\boldsymbol{\beta}_{t+1}$ for the next iterate. Similarly, we use θ for θ_t and and θ' for θ_{t+1}. We assume without loss of generality that the initial angle θ_0 is in $[0, \pi/2)$, where $\pi/2$ is excluded as it has measure zero. An initial solution falling in the remainder of the circle has precisely the same behavior, but with a convergence to $-\boldsymbol{\beta}^*$ instead of $\boldsymbol{\beta}^*$.

For the iterates and angles in the finite-sample EM, we use $\widetilde{\cdot}$ to distinguish them from the population case. For instance, $\widetilde{\boldsymbol{\beta}}_t$ denotes the t^{th} iterate of the finite-sample EM and $\widetilde{\theta}_t$ denotes the angle between $\widetilde{\boldsymbol{\beta}}_t$ and $\boldsymbol{\beta}^*$. Similarly, for a single iteration of finite-sample EM with the current iterate $\boldsymbol{\beta}$, the notations $\widetilde{\boldsymbol{\beta}}'$ and $\widetilde{\theta}'$ denote the next iterate and its angle with $\boldsymbol{\beta}^*$, respectively.

Recall that σ is the known standard deviation of the noise $\{e_i\}$, and the SNR is defined as $\eta := \frac{\|\boldsymbol{\beta}^*\|}{\sigma}$, with the assumption that $\eta \le \rho = O(1)$.

1.5 Paper Organization

In Section 2, we demonstrate a few structural properties of the population EM update. The global convergence result of the population EM is provided in Section 3. The global convergence and minimax results of the finite-sample EM in the high and low SNR regimes are provided in Section 4. The proofs of our main results are provided in Sections 6, 7 and 8. The paper is concluded in Section 9 with a discussion of future directions.

2 Population EM and Likelihood Landscape

In this section, we derive several structural properties of the population EM update. By connecting the EM update with the log likelihood of MLR, we provide a characterization of the landscape of the likehood function. These results highlight the main challenges in the MLR problem and the reasons why they can be resolved, which serves as a starting point of our subsequent proof for global convergence.

2.1 Explicit Expression for the Population EM Update

Given the current iterate $\boldsymbol{\beta}$, we consider one iteration of the population EM update (3) which yields the next iterate $\boldsymbol{\beta}'$. Since the distribution of the covariate $\boldsymbol{X}$ is spherically symmetric, we may choose a convenient an orthonormal basis $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_d\}$ of $\mathbb{R}^d$ as follows. We let $\boldsymbol{v}_1$ be a unit vector in the direction of $\boldsymbol{\beta}$ and $\boldsymbol{v}_2$ be a unit vector that is in $\operatorname{span}\{\boldsymbol{\beta}, \boldsymbol{\beta}^*\}$ and orthogonal to $\boldsymbol{v}_1$. In this case, $\boldsymbol{X}$ can be written as $\boldsymbol{X} := \sum_{i=1}^d \alpha_i \boldsymbol{v}_i$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Introduce the shorthands $b_1 := \langle \boldsymbol{\beta}, \boldsymbol{v}_1 \rangle = \|\boldsymbol{\beta}\|$, $b_1^* := \langle \boldsymbol{\beta}^*, \boldsymbol{v}_1 \rangle, b_2^* := \langle \boldsymbol{\beta}^*, \boldsymbol{v}_2 \rangle$ and $\sigma_2^2 := \sigma^2 + b_2^{*2}$. We may write the next iterate $\boldsymbol{\beta}'$ as

$$\boldsymbol{\beta}' = \mathbb{E}_{\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[\mathbb{E}_{Y | \boldsymbol{\alpha} \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)} \left[\tanh\left(\frac{b_1 \alpha_1}{\sigma^2} Y \right) Y \right] \sum_{i=1}^d \alpha_i \boldsymbol{v}_i \right]. \tag{4}$$

Without loss of generality, we assume that $b_1, b_1^*, b_2^* \geq 0$. The following lemma provides an explicit expression of $\boldsymbol{\beta}'$ under the above orthonormal basis.

Lemma 1 (Explicit Update for Population EM)**.** *Let $\boldsymbol{\beta} \neq \boldsymbol{0}$ be the current iterate and $\boldsymbol{\beta}'$ be the next iterate defined in equation (4). Then $\boldsymbol{\beta}'$ is in span$(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ and can be written as $\boldsymbol{\beta}' = b_1' \boldsymbol{v}_1 + b_2' \boldsymbol{v}_2$ with*

$$b_1' = b_1^* S + R \quad \text{and} \quad b_2' = b_2^* S, \tag{5}$$

where S and R have the following expressions:

$$S := \mathbb{E}_{\alpha_1, z} \left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2} (\sigma_2 z + \alpha_1 b_1^*) \right) + \frac{\alpha_1 b_1}{\sigma^2} (\sigma_2 z + \alpha_1 b_1^*) \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2} (\sigma_2 z + \alpha_1 b_1^*) \right) \right], \tag{6a}$$

$$R := (\sigma^2 + \|\boldsymbol{\beta}^*\|^2) \mathbb{E}_{\alpha_1, z} \left[\frac{\alpha_1^2 b_1}{\sigma^2} \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2} (\sigma_2 z + \alpha_1 b_1^*) \right) \right]. \tag{6b}$$

The expectations above are taken over $\alpha_1 \sim \mathcal{N}(0,1)$ and $z \sim \mathcal{N}(0,1)$. Moreover, we have $S \geq 0$ and $R > 0$, where $S = 0$ if and only if $b_1^* = 0$.

Lemma 1 is proved in Section 6.1. Qualitatively, the lemma establishes that the next iterate $\boldsymbol{\beta}'$ remains in the linear subspace spanned by the current iterate $\boldsymbol{\beta}$ and the true parameter $\boldsymbol{\beta}^*$. Moreover, if $\boldsymbol{\beta}$ is orthogonal to $\boldsymbol{\beta}^*$, then $\boldsymbol{\beta}'$ remains in span$(\boldsymbol{\beta})$. Consequently, if we run the population update starting from some initial solution $\boldsymbol{\beta}_0$, then it holds that $\boldsymbol{\beta}_t \in \operatorname{span}(\boldsymbol{\beta}_0, \boldsymbol{\beta}^*)$ for all $t = 1, 2, \ldots$

The quantities b_1' and b_2' in Lemma 1 represent the projections of $\boldsymbol{\beta}'$ along $\boldsymbol{v}_1$ (direction of $\boldsymbol{\beta}$) and $\boldsymbol{v}_2$ (the orthogonal direction to $\boldsymbol{\beta}$), respectively. From the expressions of b_1' and b_2', we can further deduce the following quantitative properties of the population EM dynamics:

1. **Decreasing angle**: When $\angle(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \in \left(0, \frac{\pi}{2}\right)$, then $\angle(\boldsymbol{\beta}', \boldsymbol{\beta}^*) < \angle(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, that is, each iteration of population EM strictly decreases the angle between the iterate and the true parameter.
2. **Contraction along $\boldsymbol{\beta}$**: In the direction of $\boldsymbol{v}_1$ (equivalently, $\boldsymbol{\beta}$), $\boldsymbol{\beta}'$ moves towards a unique fixed point $E(\boldsymbol{v}_1)$; i.e., $|b_1' - E(\boldsymbol{v}_1)| \leq |b_1 - E(\boldsymbol{v}_1)|$ with equality holds if and only if $b_1 = E(\boldsymbol{v}_1)$.

The first property immediately follows from the expression of b_2'. In particular, note that $0 \leq \tan\angle(\boldsymbol{\beta}', \boldsymbol{\beta}) = \frac{b_2'}{b_1'} \leq \frac{b_2^*}{b_1^*} = \tan\angle(\boldsymbol{\beta}^*, \boldsymbol{\beta})$. When $\frac{b_2'}{b_1'} > 0$, the angle strictly decreases; when $\frac{b_2'}{b_1'} = 0$, the angle remains the same. In particular, the latter case $\frac{b_2'}{b_1'} = 0$ happens if and only if $b_2' = 0$, which means either $b_2^* = 0$ (i.e., $\boldsymbol{\beta} \in \operatorname{span}(\boldsymbol{\beta}^*)$) or $S = 0$ (i.e., $\boldsymbol{\beta} \perp \boldsymbol{\beta}^*$). The second property follows from the expression of b_1'; the derivation is given in Lemma 10.

2.2 Structural Properties of Population EM and Likelihood Function

The population *negative* log-likelihood function $\mathcal{L}$ of the MLR model (1) is given by

$$\mathcal{L}(\boldsymbol{\beta}) = -\mathbb{E}_{\boldsymbol{X}} \mathbb{E}_{Y | \boldsymbol{X}} \left[\log\left(\frac{1}{2\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y - \langle \boldsymbol{X}, \boldsymbol{\beta} \rangle)^2}{2\sigma^2} \right) + \frac{1}{2\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y + \langle \boldsymbol{X}, \boldsymbol{\beta} \rangle)^2}{2\sigma^2} \right) \right) \right], \tag{7}$$

where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $Y|\boldsymbol{X} \sim \mathcal{N}(\langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle, \sigma^2)$. Interestingly, it can be shown that the population EM update is equivalent to applying *gradient descent* to the population negative log-likelihood.

Lemma 2 (Connection Between EM and Gradient Descent)**.** *Given the current iterate $\boldsymbol{\beta}$, the next iterates produced by the population EM update (3) satisfies*

$$\boldsymbol{\beta}' = \boldsymbol{\beta} - \sigma^2 \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}).$$

Consequently, the set of fixed points of the population EM update is equal to the set of stationary points of the population negative log-likelihood $\mathcal{L}$.

Proof. Direct computation shows that the gradient of $\mathcal{L}$ given in (7) admits the expression:

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \left[\boldsymbol{\beta} - \mathbb{E}_{\boldsymbol{X}} \mathbb{E}_{Y|\boldsymbol{X}} \left[\tanh\left(\frac{\langle \boldsymbol{X}, \boldsymbol{\beta} \rangle Y}{\sigma^2} \right) Y \boldsymbol{X} \right] \right]. \tag{8}$$

Comparing this equation with the expression of the population EM update (5), we see that $\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}')$. The lemma follows. □

Using the two properties derived in the last subsection, we obtain the following complete characterization of the fixed points of the population EM as well as the stationary points of the population log likelihood.

Theorem 2 (Population EM and Log-likelihood)**.** *Let $\boldsymbol{v}$ be an arbitrary unit vector orthogonal to $\boldsymbol{\beta}^*$. In the subspace span$(\boldsymbol{v}, \boldsymbol{\beta}^*)$, the population negative log-likelihood function (7) has exactly five stationary points:*

$$\boldsymbol{\beta}^*, \ -\boldsymbol{\beta}^*, \ \boldsymbol{0}, \ E(\boldsymbol{v})\boldsymbol{v}, \ -E(\boldsymbol{v})\boldsymbol{v},$$

where $E(\boldsymbol{v}) > 0$. In particular, $\pm\boldsymbol{\beta}^*$ are global minima, $\boldsymbol{0}$ is a local maximum, and $\pm E(\boldsymbol{v})\boldsymbol{v}$ are saddle points whose Hessians have a strictly negative eigenvalue. Moreover, these five points are the only fixed points of the population EM (4) in span$(\boldsymbol{v}, \boldsymbol{\beta}^*)$.

Theorem 2 is proved in Section 6.2. In the left pane of Figure 1, we illustrate the landscape of the negative log-likelihood of MLR in dimension $d = 2$. Since $\pm\boldsymbol{\beta}^*$ are the only local minima, it can be expected that population EM (equivalent to gradient descent) converges to them from a random initialization—we establish this result rigorously in subsequent sections and provide non-asymptotic convergence rates. On the other hand, due to the existence of saddle points, the ℓ_2 distance of the EM iterates to $\boldsymbol{\beta}^*$ *cannot* contract globally. In particular, if the current iterate $\boldsymbol{\beta}$ is the near a saddle point and the maximum $\boldsymbol{0}$, the next iterate $\boldsymbol{\beta}'$ will first move toward the saddle point before making progress to $\boldsymbol{\beta}^*$, hence $\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| > \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$. This issue is only exacerbated in higher dimensions, where most $\boldsymbol{\beta}$'s are nearly orthogonal to $\boldsymbol{\beta}^*$ and hence likely to be near a saddle point. A similar non-contraction phenomenon for EM was pointed out in by Klusowski et al [17]; here we provide a geometric explanation in terms of the likelihood landscape.

We note that negative likelihood function of GMM does *not* have such non-zero saddle points, as illustrated in the right pane of Figure 1. Consequently, the ℓ_2 distance does decrease globally in this problem, as is established in previous global analysis of EM under GMM [10, 36]. This ℓ_2-distance-based analysis, however, is fundamentally insufficient for proving global convergence under MLR.

3 Convergence Analysis of the Population EM

In this section, we provide our main results on the global convergence of the population EM. As mentioned, a major challenge in the analysis is the non-contraction of the ℓ_2 distance of the EM iterates to the true parameter $\boldsymbol{\beta}^*$. To address this challenge, we adopt the new strategy of first proving a rapid decrease in angle and then proving a geometric decrease in ℓ_2 distance.

3.1 Convergence in Cosine

Recall that $\eta := \|\boldsymbol{\beta}^*\|/\sigma$ is the SNR, and θ_0, θ and θ' denote the angles that $\boldsymbol{\beta}^*$ forms with $\boldsymbol{\beta}_0$ (initial iterate), $\boldsymbol{\beta}$ (current iterate), and $\boldsymbol{\beta}'$ (next iterate), respectively. By symmetry we may assume without loss of generality that $\cos\theta_0$ is positive. For the early stage of the EM iterations, we focus on the cosine of the angle and show that it increases geometrically with a constant rate.

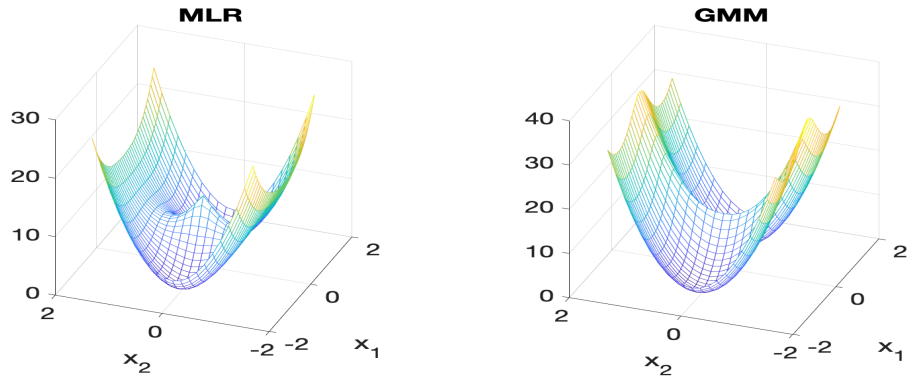


Figure 1: Negative log-likelihood functions of MLR (left) and GMM (right) with true parameter $\boldsymbol{\beta}^* = (1, 0)$. In both problems, $\pm\boldsymbol{\beta}^*$ are the only local minima. MLR has a local maximum at $\mathbf{0}$ and two non-zero saddle points along the x_2 axis that are orthogonal to $\boldsymbol{\beta}^*$. GMM has a saddle point at $\mathbf{0}$ and no other stationary points.

Theorem 3 (Cosine Convergence). *When $0 \le \theta < \frac{\pi}{2}$, the population EM iteration (4) satisfies*

$$\cos(\theta') \ge \kappa_1(\theta)\cos(\theta), \tag{9}$$

where $\kappa_1(\theta) = \sqrt{1 + \frac{\sin^2(\theta)}{\cos^2(\theta) + \frac{1}{2}(1+\eta^{-2})}}$. In particular, when $\theta \ge \frac{\pi}{3}$, we have $\kappa_1(\theta) \ge \sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}}$. Consequently, if $\cos(\theta_0) = \Theta(1/\sqrt{d})$, after $T = O\big(\max(1, \eta^{-2})\log d\big)$ iterations, we get $\theta_T < \pi/3$ or equivalently $\cos(\theta_T) \ge \frac{1}{2}$.

Theorem 3 is proved in Section 7.2. Note that using a random initialization, we have $\cos\theta_0 = \Theta(1/\sqrt{d})$ with high probability (see Lemma 16). Therefore, starting such an initial angle θ_0, Theorem 3 ensures that a logarithmic number of iterations of the population EM is sufficient to achieve $\cos\theta_t = O(1)$.

Theorem 3 provides explicit characterization of the linear convergence rate, where the ratio $\kappa_1(\theta)$ between $\cos\theta'$ and $\cos\theta$ is bounded away from 1 when θ is bounded away from 0. Therefore, this result is most useful in the early stage of EM. As θ goes to 0, the ratio $\kappa_1(\theta)$ approaches 1, in which case the cosine of the angle is no longer informative for establishing a linear convergence rate. In the following subsection, we establish a complementary result for the sine of the angle.

3.2 Convergence in Sine

Our next theorem shows that the sine of the angle converges geometrically to 0. This result is reminiscent of Theorem 3 in Xu et al. [36], where they considered GMM and used a similar argument to show *asymptotic* convergence. Here we provide an explicit rate of convergence by quantifying the amount of change in sine. This quantitative, non-asymptotic guarantee is critical when we port the population-level results to the finite sample setting.

Theorem 4 (Sine Convergence). *When $0 \le \theta < \frac{\pi}{2}$, the population EM iteration (4) satisfies*

$$\sin\theta' \le \kappa_2(\theta)\sin\theta, \tag{10}$$

where $\kappa_2(\theta) = \left(\sqrt{1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta}\right)^{-1} < 1$. In particular, when $\theta < \frac{\pi}{3}$, we have $\kappa_2(\theta) < \left(\sqrt{1 + \frac{\eta^2}{1+\eta^2}}\right)^{-1}$.

Theorem 4 is proved in Section 7.1. Note that the speed of convergence increases as the angle decreases. This result is most useful when the angle is bounded away from $\pi/2$—complementary to the case covered by Theorem 3. In particular, starting from an initial solution $\theta_0 < \pi/3$, Theorem 4 ensures that after $T = O\big(\max(1, \eta^{-2})\big)$ iterations, the population EM outputs a solution satisfying $\theta_T < \pi/8$.

We remark that in the high SNR regime ($\eta \gg 1$), the ratio $\kappa_2(\theta)$ can be much smaller than 1, despite depending on the initial angle. In the low SNR regime ($\eta \ll 1$), however, the ratio $\kappa_2(\theta)$ cannot be smaller than $1 - O(\eta^2)$, regardless of the initial angle.

3.3 Convergence in ℓ_2 Distance

Combining the above results on cosine and sine, we can conclude that eventually the population EM pushes any random initial solution into a region with a small angle around $\boldsymbol{\beta}^*$. At this point, EM safely transits to the stage that exhibits a contraction in ℓ_2 distance, which is the content of our next result.

Theorem 5 (ℓ_2 Contraction)**.** *Suppose we have that $\theta < \pi/8$. Recall the shorthands $b_1 := \|\boldsymbol{\beta}\|$, $b_1^* := \|\boldsymbol{\beta}^*\|\cos(\theta)$, $b_2^* := \|\boldsymbol{\beta}^*\|\sin(\theta)$ and $\sigma_2^2 := \sigma^2 + b_2^{*2}$. The following holds for the population EM iteration (4):*

- *If $b_2^* < \sigma$ or $\frac{\sigma_2^2}{\sigma^2} b_1 < b_1^*$, then*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le \kappa_3(\theta)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \kappa_3(\theta)(16\sin^3\theta)\|\boldsymbol{\beta}^*\|\frac{\eta^2}{1+\eta^2}, \tag{11a}$$

  where $\kappa_3(\theta) = \left(\sqrt{1 + \min\left(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*\right)^2/\sigma_2^2}\right)^{-1}$.

- *If $b_2^* \ge \sigma$ and $\frac{\sigma_2^2}{\sigma^2} b_1 > b_1^*$, we have*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le 0.6\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{11b}$$

Theorem 5 is proved in Section 7.3. Note that the bound (11a) has an additional term that depends on the angle and SNR. When b_1 is close to b_1^* and σ is small, we get a better contraction bound in (11b).

Equipped with the above per-iteration contraction result, we can bound the ℓ_2 error after t iterations of population EM and conclude that it converges to $\boldsymbol{\beta}^*$.

Corollary 1 (ℓ_2 Convergence)**.** *Suppose that the initial solution satisfies $\theta_0 < \pi/8$. There exists a constant $\kappa < 1$ such that after T iterations of the population EM, we have the error bound*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| < \kappa^T\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + T\kappa^T\|\boldsymbol{\beta}^*\|\frac{\eta^2}{1+\eta^2}. \tag{12}$$

In particular, the constant κ can be taken to be the maximum among

$$0.6,\ \sqrt{\left(1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}\right)^{-1}},\ \sqrt{1 - \frac{0.8\eta^2}{1+\eta^2}}. \tag{13}$$

Corollary 1 is proved in Section 7.4. We shall see in the proof that the value of κ depends on $\max(\kappa_2^3(\theta_0), \kappa_3(\theta_0))$, which is upper bounded by $\max(\kappa_2^3(\pi/8), \kappa_3(\pi/8))$ when $\theta_0 < \pi/8$. Therefore, the convergence rate κ depends on the SNR η as well as the norm $\|\boldsymbol{\beta}_0\|$ of the initial solution. For different values of the SNR η, the rate is either a constant or $1 - O(\eta^2)$, as was in the case of bounding the sine. Therefore, $T = O\big(\max(1, \eta^{-2})\log(1/\epsilon)\big)$ iterations is sufficient to achieve a solution ϵ-close to $\boldsymbol{\beta}^*$.

Combining the above results on the cosine, sine and ℓ_2 distance, we conclude that starting from a random initial solution, the population EM converges to $\boldsymbol{\beta}^*$ and achieves an ϵ error in ℓ_2 distance in $O\big(\max(1, \eta^{-2})\log(d/\epsilon)\big)$ iterations.

4 Finite Sample Analysis

We now turn to proving the convergence of the finite-sample EM update given in equation (2). Throughout this section, we assume that the number of samples n satisfies $n \ge Cd$ for some sufficiently large constant $C > 0$. Our analysis is divided into two cases: the middle-high SNR regime and the low SNR regime. For high and middle SNR, *i.e.,* $\eta \gtrsim (d/n)^{1/4}$, we relate the finite EM update with the angle convergence argument we used for the population EM. In contrast, for a low SNR, *i.e.,* $\eta \lesssim (d/n)^{1/4}$, we do not require any angle convergence argument since we only need to show that the norm of the iterate shrinks until it enters in the ball of radius $(d/n)^{1/4}$. Thus, we handle the low-SNR regime in Section 4.3 separately.

For the bulk of this section, we assume the following:

$$\text{Middle-to-High SNR regime:} \quad \eta \ge C(d\log^2(n/\delta)/n)^{1/4}, \tag{14}$$

for some universal constant $C > 0$. In this regime, we show that at each iteration, the finite-sample update is close to its population counterpart up to a "statistical fluctuation" term ϵ_f, defined as:

$$\epsilon_f := c\sqrt{d\ln^2(n/\delta)/n}, \tag{15}$$

for some absolute constant $c > 0$.

In this section, we use $\boldsymbol{\beta}$ to denote our current iterate, $\boldsymbol{\beta}'$ for the output from one step of the *population* EM, and $\widetilde{\boldsymbol{\beta}}'$ for the output from one step of the *finite-sample* EM. Accordingly, $\widetilde{\theta}'$ denotes the angle between $\widetilde{\boldsymbol{\beta}}'$

and $\boldsymbol{\beta}^*$. When we consider the sequence of iterates generated by the finite-sample EM, we use $\widetilde{\boldsymbol{\beta}}_t$ for the t^{th} iterate and $\widetilde{\theta}_t$ for its angle with $\boldsymbol{\beta}^*$.

Our results, summarized below, establish that the finite-sample EM converges in four phases in the middle-to-high SNR regime:

1. **Possible initialization from Spectral Method:** Start from a randomly initialized vector $\boldsymbol{\beta}_0$. With high probability, the vector $\boldsymbol{\beta}_0$ satisfies $\cos(\widetilde{\theta}_0) = \Theta(1/\sqrt{d})$. We compare the statistical fluctuation ϵ_f to the threshold $\min(1, \eta^2)/\sqrt{d}$, which amounts to the increase in cosine values. If $\epsilon_f > \min(1, \eta^2)/\sqrt{d}$ (equivalently, if $n < \max(1, \eta^{-4}) \cdot d^2 \ln^2(n/\delta)$), then we first use the standard spectral method to get an initial vector $\widetilde{\boldsymbol{\beta}}_0$ such that $\cos(\widetilde{\theta}_0) = \Omega\big(\max(1, \eta^{-2}) \cdot \epsilon_f\big)$. Otherwise, we set $\widetilde{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0$ and directly go to Phase 2.
2. **Decreasing Angle:** Starting from $\widetilde{\boldsymbol{\beta}}_0$ obtained from Phase 1, which satsifies $\cos(\widetilde{\theta}_0) \geq \Omega(\epsilon_f)$, run the finite-sample EM for $T_1 = O\big(\log(1/\epsilon_f) \cdot \max(1, \eta^{-2})\big)$ iterations to get an iterate $\widetilde{\boldsymbol{\beta}}_{T_1}$ satisfying $\sin(\widetilde{\theta}_{T_1}) \geq \sin(\pi/25)$.
3. **Convergence in ℓ_2:** Starting from $\widetilde{\boldsymbol{\beta}}_{T_1}$ obtained from Phase 2, run the finite-sample EM for $T_2 = O\big(\max(1, \eta^{-2}) \log(n/d)\big)$ iterations to get an iterate $\widetilde{\boldsymbol{\beta}}_{T_2}$ satisfying $\|\widetilde{\boldsymbol{\beta}}_{T_2} - \boldsymbol{\beta}^*\| \leq O\big(\max(1, \eta^{-1})\sqrt{d/n}\big)$. This matches the known minimax rates in the middle-to-high SNR regime [8].

Remark 1 (Initialization)**.** *In this paper we do not resort to the sample-splitting scheme, in which one draws a new batch of samples at every iteration. In doing so, the challenge is to establish the right uniform bound on the statistical deviation over the parameter domain of interest. In the conference version of our paper [20], we show that the deviation in cosine value is* $\max(\epsilon_f/\sqrt{d}, \epsilon_f^2)$ *when a sample-splitting scheme is used. This allows us to analyze the EM algorithm as it is, instead of using a spectral method for the initialization. It seems that there is hard trade-off in the analysis between removing the sample-splitting scheme and avoiding the need for spectral initialization. As our focus is on the minimax-optimality of last iterates of the EM algorithm, we compromise some generality in our analysis by assuming spectral initialization when* n *is small.*

4.1 Global Convergence in Angle

We now provide the details for Phase 2 outlined above. As discussed in the introduction, our approach is based on coupling the finite sample EM iterate with the population EM iterate. The work in Balakrishnan et al. [1] establishes a concentration bound on the ℓ_2 distance between the population and finite-sample iterates in the form of

$$\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| = O\Big(\sqrt{\|\boldsymbol{\beta}^*\|^2 + \sigma^2}\sqrt{d/n}\Big).$$

This type of bound implies local contraction in distance. However, it is not sufficient for us, as we need to control the angle when the iterate is outside of the local region for ℓ_2 contraction.

We establish a more refined bound, which shows that the statistical error is (at most) proportional to the norm of the current iterate:

Lemma 3. *For any given* $r > 0$*, there exists a universal constant* $c > 0$ *such that we have*

$$\mathbb{P}\Bigg(\sup_{\|\widetilde{\boldsymbol{\beta}}'\| \leq r} \|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}\| \leq cr\sqrt{d\log^2(n/\delta)/n}\Bigg) \geq 1 - \delta. \tag{16}$$

Lemma 3 is proved in Appendix A.2. Note that the bound is holds uniformly over the parameter space, which is the crucial property that allows us to remove sample-splitting in the analysis. Using equation (16), we prove the following angle concentration bound.

Lemma 4. *With probability at least* $1 - \delta$*, the following holds for all* $\boldsymbol{\beta}$ *satisfying* $\|\boldsymbol{\beta}\| \leq C\sqrt{\|\boldsymbol{\beta}^*\|^2 + \sigma^2}$ *for some universal constant* $C > 0$*:*

$$\cos(\widetilde{\theta}') \geq \kappa_1(\theta)(1 - 10\epsilon_f)\cos(\theta) - \epsilon_f, \tag{17}$$

$$\sin^2(\widetilde{\theta}') \leq \kappa_2^2(\theta)\sin^2(\theta) + O(\epsilon_f), \tag{18}$$

where $\kappa_1(\theta) = \sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \geq 1$*, and* $\kappa_2(\theta) = \Big(1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta\Big)^{-1} < 1$.

Lemma 4, proved in Section 8.1, allows us to show that at each iteration, the finite-sample EM decreases the angle between the iterate and the true parameter, up to a quantity that depends on the statistical fluctuation $\epsilon_f \propto \sqrt{d/n}$ (and hence on the sample size). The key idea in the proof of the lemma is that when we bound the

statistical error of the cosine value, we need to control the error in the fixed direction $u = \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|$ instead of all directions in $\mathbb{R}^d$.

A consequence of Lemma 4 is that when the statistical fluctuation ϵ_f is small relative to the SNR η, the finite-sample EM iterates have strictly decreasing angles in Phase 2 (and the angles remain small in Phase 3). This result is formalized in the following corollary, whose proof is given in Section 8.2.

Corollary 2. *If $\epsilon_f \le c_1 \min(1, \eta^2)$ for a sufficiently small constant $c_1 > 0$, then with probability $1 - \delta$, we have $\widetilde{\theta}' < \theta$ in each iteration of Phases 2, and $\widetilde{\theta}' \le \frac{\pi}{25}$ in Phase 3.*

We now combine the above arguments to establish multi-step convergence of the angle. This is done in the theorem below, whose proof is given in Section 8.3.

Theorem 6 (Cosine Convergence, Finite-Sample)**.** *Suppose that $\widetilde{\boldsymbol{\beta}}^{(0)}$ is an iterate obtained from Phase 1. We run the finite-sample EM with $n = \max(1, \eta^{-2}) d/\epsilon_f^2$ samples. As long as $\widetilde{\theta}^{(t)} > \pi/25$ for all $t < T$, there exists an universal constant $c_1 > 0$ such that with probability $1 - \delta$,*

$$\cos(\widetilde{\theta}^{(t)}) \ge \big(1 + c_1 \cdot \min(1, \eta^2)\big) \cdot \cos(\widetilde{\theta}^{(t-1)}). \tag{19}$$

In particular, if $\cos(\widetilde{\theta}^{(0)}) = \Theta(1)$, then we have $\cos(\widetilde{\theta}^{(T)}) \ge 0.95$ after $T = O\big(\max(1, \eta^{-2}) \log d\big)$ iterations.

4.2 Local Convergence after Initialization: Minimax Rates

Now that we have reached an angle below $\pi/25$, the following theorem provides a convergence guarantee in ℓ_2 distance. One subtle issue is that with only the angle argument above, we have not yet said anything about the norm of the iterate. If the norm of the iterate is too small, then the EM iteration might get stuck around 0, which is a suboptimal stationary point (Theorem 2). To avoid over-complicating the analysis, we assume for now that the norm of the iterate is also well-initialized such that $\|\boldsymbol{\beta}_0\| \ge 0.9\|\boldsymbol{\beta}^*\|$. We later remove this assumption by supplying a norm initialization lemma after the angle alignment in Section 8.4.

Theorem 7 (ℓ_2 Convergence, Finite-Sample in Middle-to-High SNR Regimes)**.** *Suppose that $\widetilde{\boldsymbol{\beta}}_0$ is an iterate obtained from Phase 2 whose angle with $\boldsymbol{\beta}^*$ satisfies $\widetilde{\theta}_0 < \frac{\pi}{25}$. Furthermore, suppose that $\|\widetilde{\boldsymbol{\beta}}_0\| \ge 0.9\|\boldsymbol{\beta}^*\|$. Then, for any $\delta > 0$, there exist universal constants $C_1, C_2 > 0$ such that with probability at least $1 - \delta$,*

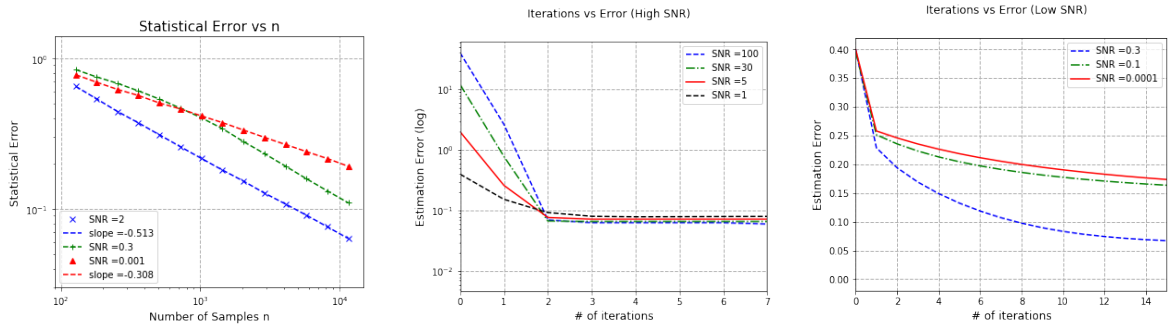
$$\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \le C_1 \sigma \max\{1, \eta^{-1}\} \big(d \log^2(n\eta/\delta)/n\big)^{1/2}$$

after $T \ge C_2 \max\{1, \eta^{-2}\} \log(n\eta/d)$ iterations.

In the high SNR regime with $\eta \gtrsim 1$, our result matches the minimax rate and in particular guarantees exact recovery when the noise variance σ goes to zero. Our proof of this bound uses an approach different from what is typically used in the literature. In particular, instead of coupling $\widetilde{\boldsymbol{\beta}}'$ and $\boldsymbol{\beta}'$ directly, we use the sample covariance matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ to our advantage, which allows us to decompose the error $\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^*$ in a way that correctly captures the behavior of the finite-sample iterate $\widetilde{\boldsymbol{\beta}}'$ *near* $\boldsymbol{\beta}^*$. In this local region, the finite-sample EM in fact behaves similarly to the standard least-squares estimator applied to two separate linear regression problems, in which case the statistical error does not depend on the regressors $\pm\boldsymbol{\beta}^*$. We conjecture that a more careful analysis can also resolve even the logarithmic dependency on η, and leave it as future work.

Another interesting point arises in the middle SNR regime where $(d/n)^{1/4} \lesssim \eta < 1$. Our statistical rate scales as $\eta^{-1}\sqrt{d/n}$, which matches the known lower bound in the middle SNR regime [8]. Note that this bound holds only when $\eta \ge (d/n)^{1/4}$; if η becomes smaller, the problem transits to the low SNR regime, which we investigate in detail in the next subsection. The main challenge in the middle SNR regime is to guarantee the progress toward $\boldsymbol{\beta}^*$ despite the slow convergence rate $(1 - \eta^2)$. Since the statistical fluctuation ϵ_f per iteration is uniformly $\sqrt{d/n}$, a naive approach based on the concentration of the EM operator would require $n \ge \eta^{-6}$ so that not only the EM iteration moves forward but the accumulation of statistical errors is also controlled in all iterations. To avoid the excessive sample requirement above, we adopt the localization argument used in the recent works [13], which established the convergence behaviors of the EM algorithm under over-specified Gaussian mixtures. Specifically, the localized bound in Lemma 3 is the key for obtaining the minimax rate in the middle SNR regime as well as for the removal of sample-splitting.

With Lemma 3, the core of our analysis consists of two main steps: (i) refinement of the convergence rate of the population EM operator, namely, the contraction coefficient of population EM is shown to be $1 - O(\max\{\|\boldsymbol{\beta}\|^2 - \eta^2, \eta^2\})$, (ii) multi-level application of uniform concentration bound for the EM operators, which shows that the statistical deviation is proportional to $\|\boldsymbol{\beta}\|\sqrt{d/n}$. The EM update is shown to make progress until $\eta^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| < \|\boldsymbol{\beta}\|\sqrt{d/N}$, at which point EM achieves the desired minimax statistical error $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \approx \sigma\eta^{-1}\sqrt{d/N}$. For the complete proof, see Section 8.5.



(a) (b) (c)

Figure 2: Convergence behavior of the EM algorithm when $d = 5$: (a) statistical rates of EM iterates (i.e., $\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\|$ at the last iteration) for different SNRs; (b) linear convergence in high SNR regime; (c) slow convergence in low SNR regime.

4.3 Finite Sample Analysis: Low SNR Regime

In this subsection, we turn our focus to the low SNR regime, where

$$\text{Low SNR regime:} \quad \eta \leq C(d\log^2(n/\delta)/n)^{1/4}, \tag{20}$$

for some universal constant $C > 0$. In this regime, instead of bounding the distance between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, *i.e.,* $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$, we aim to obtain a bound simply for $\|\boldsymbol{\beta}\|$. The triangle inequality then gives $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \|\boldsymbol{\beta}\| + \|\boldsymbol{\beta}^*\|$. Therefore, if we can show that

$$\|\boldsymbol{\beta}_T\| \lesssim \sigma(d/n)^{1/4},$$

after some T iterations, then with the low SNR condition $\|\boldsymbol{\beta}^*\| \lesssim \sigma(d/n)^{1/4}$, we obtain the desired bound $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \lesssim (d/n)^{1/4}$. Therefore, we do not need the angle convergence argument in this regime; proving convergence of the norm suffices. Intuitively, in the low SNR regime, the EM algorithm essentially cannot distinguish between $\boldsymbol{\beta}^* = 0$ and $\boldsymbol{\beta}^* \neq 0$. In fact, this is true for any algorithm in view of the known lower bound in the low SNR regime [8].

Finite sample analysis in the low SNR regime starts with the following Taylor-like approximation on the norm of the population EM iterate:

Lemma 5. *There exists some universal constants $c_u > 0$ such that,*

$$\|\boldsymbol{\beta}\|(1 - 4(\|\boldsymbol{\beta}\|/\sigma)^2 - c_u\eta^2) \leq \|\boldsymbol{\beta}'\| \leq \|\boldsymbol{\beta}\|(1 - (\|\boldsymbol{\beta}\|/\sigma)^2 + c_u\eta^2).$$

Lemma 5 implies that the population EM iterates moves toward 0 until $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}^*\|$, afer which the iterate stays in the ball of radius $\tilde{O}(\sigma\eta)$. To prove this result, we apply the localization argument, which is valid until $\boldsymbol{\beta}$ reaches $\sigma\eta \approx \sigma(d\log^2(n/\delta)/n)^{1/4}$. The final product of our analysis is the following finite-sample convergence theorem for the low SNR regime.

Theorem 8 (ℓ_2 Convergence, Finite-Sample in Low SNR Regime)**.** *Suppose $\eta \leq C(d\log^2(n/\delta)/n)^{1/4}$ and $\|\widetilde{\boldsymbol{\beta}}_0\| = O(\sigma)$. Then there exist universal constants $C_1, C_2 > 0$ such that with probability at least $1 - \delta$, we have*

$$\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq C_1\sigma(d\log^2(n/\delta)/n)^{1/4}$$

after $T \geq C_2\log(\log(n/d))\sqrt{n/(d\log^2(n/\delta))}$ iterations of finite-sample EM.

The initialization condition $\|\widetilde{\boldsymbol{\beta}}_0\| = O(\sigma)$ is mild: even if we start from an iterate with a much larger norm, one step of EM would bring the norm down to $O(\sigma)$. The proof of Theorem 8 is given in Appendix 8.6.

5 Experiments

In this section, we corroborate our theoretical results via numerical examples. In Figure 2, we present the statistical rate and convergence behavior of EM algorithm under different SNR regimes. We set $d = 5$ and initialize the EM iteration in a neighborhood of the true parameters such that $\boldsymbol{\beta}_0 = \boldsymbol{\beta}^* + ru$, where

$r = \max\{1, \|\boldsymbol{\beta}^*\|\} \cdot 0.1$ and u is a random unit vector. To evaluate the statistical rate, we run the EM algorithm with different sample size $n \in \{128, 180, 256, ...\}$ (i.e., n increases by a factor of $\sqrt{2}$ each time). The final error $\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\|$ is averaged over 5,000 independent runs. The stopping criterion is $\|\widetilde{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T-1}\| \le 0.0001$. In Figure 2(a), we observe the standard $n^{-1/2}$ rate in the high SNR regime, and an approximately $n^{-1/4}$ rate in the low SNR regime. Interestingly, with an intermediate SNR $= 0.3$, the statistical rate transitions from $n^{-1/4}$ to $n^{-1/2}$ as n increases. This is consistent with the definition of low SNR $\|\boldsymbol{\beta}^*\| \lesssim (d/n)^{1/4}$, which is relative to the sample size n rather than being an absolute value.

We next investigate the convergence behavior of EM. We run the EM algorithm with a fixed sample size $n = 32768$. The estimation error $\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\|$ in each iteration t is averaged over 5,000 independent runs. Figure 2(b) shows the high SNR regime. Note that the y-axis is in log-scale and we can see the linear convergence (up to the statistical error). In contrast, in the low SNR regime showed in Figure 2(c), we can observe that the convergence of the EM algorithm is no longer linear and becomes significantly slower.

6 Proofs for Section 2

In this section, we prove the technical results in Section 2. In particular, Lemma 1 is proved in Section 6.1 and Theorem 2 is proved in Section 6.2.

6.1 Proof of Lemma 1

We restate the lemma below for readers' convenience.

Lemma 1 (Explicit Update for Population EM)**.** *Let $\boldsymbol{\beta} \neq \mathbf{0}$ be the current iterate and $\boldsymbol{\beta}'$ be the next iterate defined in equation (4). Then $\boldsymbol{\beta}'$ is in span$(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ and can be written as $\boldsymbol{\beta}' = b_1'\boldsymbol{v}_1 + b_2'\boldsymbol{v}_2$ with*

$$b_1' = b_1^* S + R \quad \text{and} \quad b_2' = b_2^* S, \tag{5}$$

where S and R have the following expressions:

$$S := \mathbb{E}_{\alpha_1, z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right], \tag{6a}$$

$$R := (\sigma^2 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}_{\alpha_1, z}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right]. \tag{6b}$$

The expectations above are taken over $\alpha_1 \sim \mathcal{N}(0,1)$ and $z \sim \mathcal{N}(0,1)$. Moreover, we have $S \ge 0$ and $R > 0$, where $S = 0$ if and only if $b_1^* = 0$.

Proof. Recall that we have dereived a representation of the EM update $\boldsymbol{\beta}'$ in equation (3) after choosing an appropriate orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^d$ of $\mathbb{R}^d$ in which $\boldsymbol{v}_1 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$ is the unit vector in the direction of the current estimator, and $\boldsymbol{v}_2$ is the unit vector in span$\{\boldsymbol{\beta}, \boldsymbol{\beta}^*\}$ that is orthogonal to $\boldsymbol{v}_1$. We restate equation (3) below:

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

where the expectation is taken over $\alpha_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and $Y \mid \alpha_1, \ldots, \alpha_d \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)$,where $b_1 := \langle \boldsymbol{\beta}, \boldsymbol{v}_1\rangle = \|\boldsymbol{\beta}\| > 0$, $b_1^* =: \langle \boldsymbol{\beta}^*, \boldsymbol{v}_1\rangle$, and $b_2^* := \langle \boldsymbol{\beta}^*, \boldsymbol{v}_2\rangle$. Since the inner expectation does not depend on α_j for $j \ge 3$, we have

$$\mathbb{E}_{\alpha_1, \ldots, \alpha_d}\left[\mathbb{E}_{Y|\alpha_1, \ldots, \alpha_d}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}Y\right)Y\right]\alpha_j\right] = \mathbb{E}_{\alpha_1, \alpha_2}\mathbb{E}_{Y|\alpha_1, \alpha_2}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}Y\right)Y\right] \cdot \mathbb{E}_{\alpha_j}[\alpha_j] = 0.$$

This implies that $\boldsymbol{\beta}'$ is in the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and the expression (3) can be rewritten as $\boldsymbol{\beta}' = b_1'\boldsymbol{v}_1 + b_2'\boldsymbol{v}_2$, where b_1' and b_2'

$$b_1' = \mathbb{E}_{\alpha_1, \alpha_2}\left[\mathbb{E}_{Y|\alpha_1, \alpha_2}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}Y\right)Y\right]\alpha_1\right], \tag{21a}$$

$$b_2' = \mathbb{E}_{\alpha_1, \alpha_2}\left[\mathbb{E}_{Y|\alpha_1, \alpha_2}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}Y\right)Y\right]\alpha_2\right]. \tag{21b}$$

Note that we can write $Y \overset{\mathrm{d}}{=} \alpha_1 b_1^* + \alpha_2 b_2^* + \sigma z$ ($\overset{\mathrm{d}}{=}$ means equality in distribution) for some $z \sim \mathcal{N}(0,1)$ that is independent of α_1 and α_2. We call it the first representation of Y. In addition, since α_2 and z are independent and hence $\alpha_2 b_2^* + \sigma z$ is Gaussian with mean 0 and variance σ_2^2, we can also write $Y \overset{\mathrm{d}}{=} \alpha_1 b_1^* + \sigma_2 z$ for some $z \sim \mathcal{N}(0,1)$ that is independent of α_1, . We call it the second representation of Y. We next prove that b_1' and b_2' have the explicit expressions claimed in Lemma 1. The key tool is the Stein's lemma for the Gaussian distribution.

We start with the second coordinate b_2'. Continuing from equation (21b), we have

$$
\begin{aligned}
b_2' &\overset{\text{(i)}}{=} \mathbb{E}_{\alpha_1,\alpha_2,z}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\right)(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\alpha_2\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_{\alpha_1,\alpha_2,z}\frac{\partial}{\partial \alpha_2}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\right)(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\right], \\
&= b_2^* \cdot \mathbb{E}_{\alpha_1,\alpha_2,z}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma z + \alpha_1 b_1^* + \alpha_2 b_2^*)\right)\right] \\
&\overset{\text{(iii)}}{=} b_2^* \cdot \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right],
\end{aligned}
$$

where in step (i), we use the first representation of Y; in step (ii), we apply Stein's lemma with respect to α_2; and in step (iii), we use the second representation of Y. This shows that $b_2' = b_2^* S$ as desired.

For the first coordinate b_1', we use a similar strategy but apply Stein's lemma in a different way. Using the second representation for Y, we rewrite equation (21a) as

$$
\begin{aligned}
b_1' &= \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)(\sigma_2 z + \alpha_1 b_1^*)\alpha_1\right] && (22) \\
&= b_1^* \cdot \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\alpha_1^2\right] + \sigma_2 \cdot \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) z\alpha_1\right]. && (23)
\end{aligned}
$$

Applying Stein's lemma to the first term in equation (23) with respect to α_1 yields

$$
\begin{aligned}
& b_1^* \cdot \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\alpha_1^2\right] \\
=& b_1^* \cdot \mathbb{E}_{\alpha_1,z}\frac{\partial}{\partial \alpha_1}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\alpha_1\right] \\
=& b_1^* \cdot \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) + \alpha_1\left(\frac{2b_1^* b_1 \alpha_1}{\sigma^2} + \frac{b_1\sigma_2}{\sigma^2}z\right)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right] \\
=& b_1^* \cdot \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right] \\
& + b_1^{*2} \cdot \mathbb{E}_{\alpha_1,z}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right]. && (24)
\end{aligned}
$$

On the other hand, applying Stein's lemma to the second term in equation (23) with respect to z yields

$$
\sigma_2 \mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\alpha_1 z\right] = \sigma_2^2 \mathbb{E}_{\alpha_1,z}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right]. \tag{25}
$$

Plugging the above identities (24) and (25) into equation (23), and using the relation that $b_1^* + \sigma_2^2 = \|\boldsymbol{\beta}^*\|^2 + \sigma^2$, we obtain that $b_1' = b_1^* S + R$ as desired.

Finally, we have $R > 0$ since it is the expectation of a random variable that is positive almost surely. For the quantity S, we prove the following bounds in Section 6.1.1.

Lemma 6 (Lower and Upper Bounds for S)**.** *Let S, b_1, b_1^* and σ_2 be as in Lemma 1. We have*

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

The lemma implies that $S \ge 0$; moreover, $S = 0$ if and only $b_1 = 0$ or $b_1^* = 0$. Since $b_1 := \|\boldsymbol{\beta}\| \neq 0$ by assumption, the proof of Lemma 1 is complete. □

6.1.1 Proof of Lemma 6

Proof. Recall the expression for S:

$$\begin{aligned} S =& \mathbb{E}_{\alpha_1}\mathbb{E}_z\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)+\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)\right] \\ =& 2\mathbb{E}_{\alpha_1:\alpha_1\geq 0}\mathbb{E}_z\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)+\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)\right], \end{aligned}$$

where the second equation holds due to the symmetry of the distribution for z. We make use of two elementary properties of the Gaussian distribution and the tanh function:

Lemma 7 (Lemma 1 [10]). *Let $u,\theta\geq 0$ and $X\sim\mathcal{N}(u,\sigma^2)$, then $\mathbb{E}_X[\tanh'(\theta X/\sigma^2)\theta X]\geq 0$.*

Lemma 8 (Lemma 2 [10]). *Let $u,\theta\geq 0$ and $X\sim\mathcal{N}(u,\sigma^2)$, then $\mathbb{E}_X[\tanh(\theta X/\sigma^2)]\geq 1-\exp\left(-\frac{\min(u,\theta)\cdot u}{2\sigma^2}\right)$.*

We apply Lemmas 7 and 8 with $u=\alpha_1 b_1^*$ and $\theta=\alpha_1\frac{\sigma_2^2}{\sigma^2}b_1$ to obtain the following lower bounds on the two terms inside the inner expectation of S:

$$\begin{aligned} &\mathbb{E}_z\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)\right]\geq 1-\exp\left[-\frac{\alpha_1^2 b_1^*\min\left(b_1^*,\frac{\sigma_2^2}{\sigma^2}b_1\right)}{2\sigma_2^2}\right] \\ &\mathbb{E}_z\left[\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)\right]\geq 0 \end{aligned}$$

Combining these two lower bounds, we obtain that

$$\begin{aligned} S \geq& 2\mathbb{E}_{\alpha_1:\alpha_1\geq 0}\left[1-\exp\left[-\frac{\alpha_1^2 b_1^*\min\left(b_1^*,\frac{\sigma_2^2}{\sigma^2}b_1\right)}{2\sigma_2^2}\right]\right] \\ =& \mathbb{E}_{\alpha_1}\left[1-\exp\left[-\frac{\alpha_1^2 b_1^*\min\left(b_1^*,\frac{\sigma_2^2}{\sigma^2}b_1\right)}{2\sigma_2^2}\right]\right]=1-\left(\sqrt{1+\frac{\min\left(\frac{\sigma_2^2}{\sigma^2}b_1,b_1^*\right)b_1^*}{\sigma_2^2}}\right)^{-1}. \end{aligned}$$

This proves the lower bound on S in Lemma 6. For the upper bound, we use the expression in equation (24) from proof of Lemma 1 to obtain that

$$\begin{aligned} S =& \mathbb{E}_{\alpha_1,z}\left[\alpha_1^2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)-\frac{b_1 b_1^*}{\sigma^2}\alpha_1^2\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)\right] \\ \leq& \mathbb{E}_{\alpha_1,z}\left[\alpha_1^2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z+\alpha_1 b_1^*)\right)\right]\leq\mathbb{E}_{\alpha_1}[\alpha_1^2]=1, \end{aligned}$$

where the two inequalities above hold since $\tanh'(x)\geq 0$ and $\tanh(x)\leq 1$ for any x. □

6.2 Proof of Theorem 2

Recall in Lemma 2 we show that the fixed points of population EM are the same as the stationary points of the negative log-likelihood function. To prove Theorem 2, we first establish several technical lemmas.

We begin with an elementary lemma on smooth concave functions.

Lemma 9. *Let $f:\mathbb{R}^+\to\mathbb{R}$ be a smooth and concave function, with a strictly decreasing derivative. Suppose that f satisfies $f(0)=0$, $f'(0)>0$, and $\lim_{x\to\infty}f(x)=-\infty$. Then there exists a unique $t>0$ such that $f(t)=0$ and $f'(t)<0$. Moreover, $f(x)>0$ if $x\in(0,t)$ and $f(x)<0$ if $x\in(t,\infty)$.*

Proof. Since f has a continuous gradient at 0 with $f'(0)>0$, there exists $t_1>0$ such that $f'(x)>0$ for all $x\leq t_1$. We thus conclude that $f(x)>0$ for all $x\in(0,t_1]$ by the Fundamental theorem of Calculus. By the continuity of f and the condition that $\lim_{x\to\infty}f(x)=-\infty$, there exists $t_2>0$ such that $f(t_2)<0$. Rolle's theorem ensures that there exists $t\in(t_1,t_2)$ such that $f(t)=0$. Since $f(0)=0$, the mean value theorem ensures that there exists $t_3\in(0,t)$ such that $f'(t_3)=0$. Using the assumption that f has a strictly decreasing derivative, we have $f'(x)\leq 0$ for all $x\geq t_3$ and $f'(x)>0$ for all $x\in(0,t_3)$. In particular, $f'(t)<0$ as $t>t_3$. Moreover, it follows that $f(x)$ is strictly increasing on $(0,t_3)$ and it is strictly decreasing on (t_3,∞), therefore, $f(x)>0$ when $x\in(0,t)$ and $f(x)<0$ when $x>t$. □

Using the above lemma, we can characterize the dynamic of the population EM iteration along the direction of the current iterate $\boldsymbol{\beta}$.

Lemma 10 (Dynamics Along $\boldsymbol{\beta}$)**.** *Suppose that $\langle \boldsymbol{\beta}, \boldsymbol{\beta}^* \rangle \geq 0$ and $\boldsymbol{\beta} \neq \mathbf{0}$. Let $\boldsymbol{v}_1$ be the unit vector of $\boldsymbol{\beta}$, and b_1' be the notation used in Lemma 1, which denotes the the projection of the next EM iterate $\boldsymbol{\beta}'$ onto* $\mathrm{span}(\boldsymbol{v}_1)$*. There exists a unique positive number $E(\boldsymbol{v}_1)$ satisfying*

$$\begin{cases} \|\boldsymbol{\beta}\| < b_1' < E(\boldsymbol{v}_1) & \text{if } \|\boldsymbol{\beta}\| < E(\boldsymbol{v}_1), \\ E(\boldsymbol{v}_1) < b_1' < \|\boldsymbol{\beta}\| & \text{if } \|\boldsymbol{\beta}\| > E(\boldsymbol{v}_1), \\ b_1' = E(\boldsymbol{v}_1) & \text{if } \|\boldsymbol{\beta}\| = E(\boldsymbol{v}_1). \end{cases}$$

Proof. We use the same notations as in the proof of Lemma 1. When $\boldsymbol{v}_1$ is fixed, b_1' only depends on $b_1 = \|\boldsymbol{\beta}\|$ from the expression (21a). Accordingly, we write

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

to emphasize b_1' is a function of b_1. Let us check a few properties of f:

1. f is smooth since the tanh funcion is smooth.
2. f is strictly increasing and concave, since its derivative

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

   is positive and is strictly decreasing with respect to b_1.
3. $f(0) = 0$ and $f'(0) > 1$, since

$$f'(0) = \mathbb{E}_{\alpha_1,\alpha_2} \mathbb{E}_{Y|\alpha_1,\alpha_2} \left[\frac{(Y\alpha_1)^2}{\sigma^2} \right] = \frac{3b_1^{*2} + b_2^{*2} + \sigma^2}{\sigma^2} > 1.$$

Let us define the shifted function $g(b_1) := f(b_1) - b_1$. The function g is a strictly concave and smooth function from Property 2 above. Moreover, we have $g(0) = 0$ and $g'(0) > 0$ from Property 3, and $\lim_{b_1 \to \infty} g(b_1) = -\infty$ from Property 4. With these properties of g, we deduce from Lemma 9 that there exists a unique $E(\boldsymbol{v}_1) > 0$ for g such that $g(E(\boldsymbol{v}_1)) = 0$. Moreover, we have $g(b_1) > 0$ when $b_1 < E(\boldsymbol{v}_1)$, $g(b_1) < 0$ when $b_1 > E(\boldsymbol{v}_1)$. Equivalently, we have

$$\begin{cases} \|\boldsymbol{\beta}\| < b_1' < E(\boldsymbol{v}_1) & \text{if } 0 < \|\boldsymbol{\beta}\| < E(\boldsymbol{v}_1), \\ \|\boldsymbol{\beta}\| > b_1' > E(\boldsymbol{v}_1) & \text{if } \|\boldsymbol{\beta}\| > E(\boldsymbol{v}_1), \\ b_1' = E(\boldsymbol{v}_1) & \text{if } \|\boldsymbol{\beta}\| = E(\boldsymbol{v}_1). \end{cases}$$

This completes the proof of Lemma 10. □

With Lemma 10, we can characterize the fixed points of population EM in a two-dimensional subspace $\mathrm{span}(\boldsymbol{v}, \boldsymbol{\beta}^*)$.

Lemma 11 (Five Fixed Points in $\mathrm{span}(\boldsymbol{v}, \boldsymbol{\beta}^*)$)**.** *Let $\boldsymbol{v}$ be an arbitrary unit vector satisfying $\boldsymbol{v} \perp \boldsymbol{\beta}^*$. In span$(\boldsymbol{\beta}^*, \boldsymbol{v})$, the population EM update has exactly five fixed points: 0, $\boldsymbol{\beta}^*$, $-\boldsymbol{\beta}^*$, $E(\boldsymbol{v})\boldsymbol{v}$ and $-E(\boldsymbol{v})\boldsymbol{v}$, where the number $E(\boldsymbol{v}) > 0$ is given in the proof of Lemma 10.*

Proof. Recall our notation that $\boldsymbol{\beta}$ is the current iterate of population EM and $\boldsymbol{\beta}'$ is the corresponding be next iterate. When $\boldsymbol{\beta} = \mathbf{0}$, we have $\boldsymbol{\beta}' = 0$ and thus $\mathbf{0}$ is a fixed point. It remains to consider non-zero fixed points.

We deduce from Lemma 1 that $\boldsymbol{\beta}$ is a fixed point if and only if $b_2 = b_2^* S = 0$, which means either $b_2^* = 0$ or $S = 0$. Note that $b_2^* = 0$ if and only if $\boldsymbol{\beta}$ is in the same direction as $\boldsymbol{\beta}^*$. Also note that $S = 0$ if and only if $b_1^* = 0$ (as we consider $b_1 \neq 0$), or equivalently $\boldsymbol{\beta}$ is in the direction of $\boldsymbol{v}$. We conclude that any non-zero fixed point must be either in $\mathrm{span}(\boldsymbol{\beta}^*)$ or in $\mathrm{span}(\boldsymbol{v})$.

Finally, recall Lemma 10, which states that there is a unique non-zero contraction point along the positive direction of $\boldsymbol{\beta}$. Therefore, in $\mathrm{span}(\boldsymbol{\beta}^*)$, $\boldsymbol{\beta}^*$ and $-\boldsymbol{\beta}^*$ are the only two fixed points. In $\mathrm{span}(\boldsymbol{v})$, $E(\boldsymbol{v})\boldsymbol{v}$ and $-E(\boldsymbol{v})\boldsymbol{v}$ are the only two fixed points. □

We are now ready to prove Theorem 2, which is restated below for readers' convenience.

Theorem 2 (Population EM and Log-likelihood). *Let $\boldsymbol{v}$ be an arbitrary unit vector orthogonal to $\boldsymbol{\beta}^*$. In the subspace span$(\boldsymbol{v}, \boldsymbol{\beta}^*)$, the population negative log-likelihood function (7) has exactly five stationary points:*

$$\boldsymbol{\beta}^*,\ -\boldsymbol{\beta}^*,\ \mathbf{0},\ E(\boldsymbol{v})\boldsymbol{v},\ -E(\boldsymbol{v})\boldsymbol{v},$$

where $E(\boldsymbol{v}) > 0$. In particular, $\pm\boldsymbol{\beta}^*$ are global minima, $\mathbf{0}$ is a local maximum, and $\pm E(\boldsymbol{v})\boldsymbol{v}$ are saddle points whose Hessians have a strictly negative eigenvalue. Moreover, these five points are the only fixed points of the population EM (4) in span$(\boldsymbol{v}, \boldsymbol{\beta}^*)$.

Proof. In the subspace span$(\boldsymbol{v}, \boldsymbol{\beta}^*)$, Lemma 11 shows that population EM has exactly five fixed points $\pm\boldsymbol{\beta}^*$, $\mathbf{0}$ and $\pm E(\boldsymbol{v})\boldsymbol{v}$, which by Lemma 2 are the only stationary points of the negative log-likelihood $\mathcal{L}$. Since $\mathcal{L}(\boldsymbol{\beta})$ equals KL divergence between the MLR model with parameter $\boldsymbol{\beta}$ and the true model with parameter $\boldsymbol{\beta}^*$, we see that $\pm\boldsymbol{\beta}^*$ minimizes $\mathcal{L}$ (with value 0) and is hence the global maxima.

It remains to classify the other three stationary points. We do so by characterizing their Hessian, making use the following proposition.

Proposition 1 (Hessian of Negative Log-Likelihood). *The population negative log-likelihood $\mathcal{L}$ defined in (7) has the Hessian matrix*

$$\mathcal{H}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\left(\boldsymbol{I} - \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{Y|\boldsymbol{X}}\left[\frac{1}{\sigma^2}Y^2\boldsymbol{X}\boldsymbol{X}^\top \tanh'\left(\frac{Y\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle}{\sigma^2}\right)\right]\right).$$

Moreover, if $\boldsymbol{\beta}$ is a stationary point orthogonal to $\boldsymbol{\beta}^*$, then

$$\langle \boldsymbol{\beta}^*, \mathcal{H}(\boldsymbol{\beta})\boldsymbol{\beta}^*\rangle \le -\frac{\|\boldsymbol{\beta}^*\|^4}{\sigma^2(\sigma^2 + \|\boldsymbol{\beta}^*\|^2)}.$$

The proof of the proposition is postponed to Section 6.2.1. Using the proposition, we find the Hessian of $\mathcal{L}$ at $\mathbf{0}$ is negative definite:

$$\mathcal{H}(\mathbf{0}) = \frac{1}{\sigma^2}\left(\boldsymbol{I} - \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{Y|\boldsymbol{X}}\left[\frac{1}{\sigma^2}Y^2\boldsymbol{X}\boldsymbol{X}^\top\right]\right) = -\frac{1}{\sigma^4}\mathbb{E}_{\boldsymbol{X}}\left[\langle \boldsymbol{\beta}^*, \boldsymbol{X}\rangle^2\boldsymbol{X}\boldsymbol{X}^\top\right] \preceq 0,$$

thereby proving that $\mathbf{0}$ is a local maxima.

Finally, we consider the stationary point $E(\boldsymbol{v})\boldsymbol{v}$ (the proof for $-E(\boldsymbol{v})\boldsymbol{v}$ is similar). We claim that $E(\boldsymbol{v})\boldsymbol{v}$ is a local minimum of $\mathcal{L}$ restricted to the direction $\boldsymbol{v}$. The claim follows from the following three observations: (i) the population EM update does not increase the value of $\mathcal{L}$, a general property of the EM algorithm. (ii) in Section 2.1 we showed that if population EM is initialized in the subspace span$(\boldsymbol{v})$ with $\boldsymbol{v}$ orthogonal to $\boldsymbol{\beta}^*$, then the iterates remain in span$(\boldsymbol{v})$ (see the discussion after Lemma 1); (iii) Lemma 11 implies that in span$(\boldsymbol{v})$, population EM contracts to the point $E(\boldsymbol{v})\boldsymbol{v}$. On the other hand, we find that $E(\boldsymbol{v})\boldsymbol{v}$ is a local maximum of $\mathcal{L}$ restricted to the direction of $\boldsymbol{\beta}^*$, as Proposition 1 ensures that $\langle \boldsymbol{\beta}^*, \mathcal{H}\big(E(\boldsymbol{v})\boldsymbol{v}\big)\boldsymbol{\beta}^*\rangle$ is strictly negative. Combining pieces, we conclude that $E(\boldsymbol{v})\boldsymbol{v}$ is a saddle point, thereby completing the proof of Theorem 2. □

6.2.1 Proof of Proposition 1

Proof. Recall that Lemma 2 relates the gradient of the log-likelihood to the population EM update:

$$\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}').$$

Plugging in the expression for next iterate $\boldsymbol{\beta}'$ of the population EM update (3) and differentiating with respect to $\boldsymbol{\beta}$, we find that the Hessian matrix is

$$\begin{aligned}\mathcal{H}(\boldsymbol{\beta}) =& \frac{1}{\sigma^2}(\boldsymbol{I} - \nabla_{\boldsymbol{\beta}}\boldsymbol{\beta}') \\ =& \frac{1}{\sigma^2}\left(\boldsymbol{I} - \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{Y|\boldsymbol{X}}\left[\frac{1}{\sigma^2}Y^2\boldsymbol{X}\boldsymbol{X}^\top \tanh'\left(\frac{Y\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle}{\sigma^2}\right)\right]\right).\end{aligned}$$

Let $\boldsymbol{\beta}$ a stationary point orthogonal to $\boldsymbol{\beta}^*$. As before, we use the orthonormal basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d\}$ satisfying $\boldsymbol{v}_1 = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$ and $\boldsymbol{v}_2 = \widehat{\boldsymbol{\beta}}^*$, and write $\boldsymbol{X} = \sum_i \alpha_i \boldsymbol{v}_i$, with $\alpha_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i = 1, \ldots, d$. Also recall that $b_1 = \langle \boldsymbol{\beta}, \boldsymbol{v}_1\rangle$ and $b_1' = \langle \boldsymbol{\beta}', \boldsymbol{v}_1\rangle$ are respectively the projections of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ onto the direction $\boldsymbol{v}_1$ (see Lemma 1). The stationary point $\boldsymbol{\beta}$ is a fixed point of population EM, which means that

$$b_1 = b_1' = \mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\alpha_1 Y \tanh\left(\frac{b_1\alpha_1}{\sigma^2}Y\right). \tag{26}$$

Let $\widehat{\boldsymbol{\beta}}^* = \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|$ be the unit vector of $\boldsymbol{\beta}^*$. We compute $\sigma^2\langle\widehat{\boldsymbol{\beta}}^*, \mathcal{H}(\boldsymbol{\beta})\widehat{\boldsymbol{\beta}}^*\rangle$ as follows:

$$\begin{aligned}
\sigma^2\langle\widehat{\boldsymbol{\beta}}^*, \mathcal{H}(\boldsymbol{\beta})\widehat{\boldsymbol{\beta}}^*\rangle =& 1 - \frac{1}{\sigma^2}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[Y^2\alpha_2^2\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right] \\
\overset{\text{(i)}}{=} & 1 - \frac{1}{b_1}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\frac{\partial}{\partial\alpha_1}\left[Y\alpha_2^2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right] \\
=& 1 - \frac{1}{b_1}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1 Y\alpha_2^2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right] \\
\overset{\text{(ii)}}{=} & 1 - \frac{1}{b_1}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\frac{\partial}{\partial\alpha_2}\left[\alpha_1 Y\alpha_2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right],
\end{aligned}$$

where in steps (i) and (ii), we apply Stein's Lemma with respect to α_1 and α_2, respectively. We decompose the last right hand side into three terms:

$$\begin{aligned}
\sigma^2\langle\widehat{\boldsymbol{\beta}}^*, \mathcal{H}(\boldsymbol{\beta})\widehat{\boldsymbol{\beta}}^*\rangle =& \underbrace{1 - \frac{1}{b_1}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1 Y\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right]}_{A} \\
& \underbrace{- \frac{b_2^*}{b_1}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1\alpha_2\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right]}_{B} \\
& \underbrace{- \frac{b_2^*}{\sigma^2}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1^2\alpha_2 Y\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right]}_{C}. \qquad (27)
\end{aligned}$$

The term A equals 0 thanks to the fixed point condition (26). It remains to control the terms B and C.

For term B, we apply Stein's Lemma with respect to α_2 to obtain:

$$\begin{aligned}
B =& \frac{b_2^*}{b_1}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\frac{\partial}{\partial\alpha_2}\left[\alpha_1\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right] \\
=& \frac{b_2^{*2}}{\sigma^2}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1^2\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right].
\end{aligned}$$

Note that Y admits the representation $Y \overset{\text{d}}{=} b_1^*\alpha_1 + \sigma_2 z = \sigma_2 z$ with $\sigma_2 = \sqrt{\|\boldsymbol{\beta}^*\|^2 + \sigma^2}$ and $z \sim \mathcal{N}(0,1)$ is independent of α_1; moreover, we have $b_1^* = 0$ since $\boldsymbol{\beta}$ is orthogonal to $\boldsymbol{\beta}^*$. It follows that

$$\begin{aligned}
B \overset{\text{(i)}}{=}& \frac{b_2^{*2}}{\sigma^2}\mathbb{E}_{\alpha_1,z}\left[\alpha_1^2\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}\sigma_2 z\right)\right] \\
=& \frac{b_2^{*2}}{b_1\sigma_2}\mathbb{E}_{\alpha_1,z}\frac{\partial}{\partial z}\left[\alpha_1\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}\sigma_2 z\right)\right] \\
\overset{\text{(ii)}}{=}& \frac{b_2^{*2}}{b_1\sigma_2^2}\mathbb{E}_{\alpha_1,z}\left[\sigma_2 z\alpha_1\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}\sigma_2 z\right)\right] \\
\overset{\text{(iii)}}{=}& \frac{b_2^{*2}}{b_1\sigma_2^2}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1 Y\tanh\left(\frac{b_1\alpha_1}{\sigma^2}Y\right)\right] \overset{\text{(iv)}}{=} \frac{\|\boldsymbol{\beta}^*\|^2}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}, \qquad (28)
\end{aligned}$$

where steps (i) and (iii) follows from the aforementioned representation of Y, step (ii) holds by applying Stein's Lemma with respect to z, and step (iv) follows from the fixed point condition (26).

We turn to the term C. Using symmetry of the distribution for z as well as the even property of the function $\tanh'$, we may take the expectation conditioning on the event that $\alpha_1 \geq 0, \alpha_2 \geq 0$. Doing so gives

$$\begin{aligned}
C :=& \frac{b_2^*}{\sigma^2}\mathbb{E}_{\alpha_1,\alpha_2}\mathbb{E}_{Y|\alpha_1,\alpha_2}\left[\alpha_1^2\alpha_2 Y\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}Y\right)\right] \\
=& \frac{b_2^*}{\sigma^2}\mathbb{E}_{\alpha_1,\alpha_2,z}\left[\alpha_1^2\alpha_2(b_2^*\alpha_2 + \sigma z)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(b_2^*\alpha_2 + \sigma z)\right)\right] \\
=& \frac{4b_2^*}{\sigma^2}\mathbb{E}_{\alpha_1,\alpha_2:\alpha_1\geq 0,\alpha_2\geq 0}\alpha_1^2\alpha_2\left[\mathbb{E}_z(b_2^*\alpha_2 + \sigma z)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(b_2^*\alpha_2 + \sigma z)\right)\right] \geq 0, \qquad (29)
\end{aligned}$$

where the last step follows from Lemma 8 in Section 6.1.1.

Plugging equations (28) and (29) into equation (27), we obtain that

$$\sigma^2\langle\widehat{\boldsymbol{\beta}}^*, \mathcal{H}(\boldsymbol{\beta})\widehat{\boldsymbol{\beta}}^*\rangle = A - B - C \le -\frac{\|\boldsymbol{\beta}^*\|^2}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}.$$

Multiplying both sides by $\|\boldsymbol{\beta}^*\|^2/\sigma^2$ proves Proposition 1. □

7 Proofs for Section 3

In this section, we prove the technical results in Section 3. In particular, Theorems 4 and 3 on angle convergence are proved in Sections 7.1 and 7.2, respectively. Theorem 5 and Corollary 1 on ℓ_2 distance contraction are proved in Sections 7.3 and 7.4, respectively.

7.1 Proof of Theorem 4

We restate the theorem below for readers' convenience.

Theorem 4 (Sine Convergence). *When $0 \le \theta < \frac{\pi}{2}$, the population EM iteration (4) satisfies*

$$\sin\theta' \le \kappa_2(\theta)\sin\theta, \tag{10}$$

where $\kappa_2(\theta) = \left(\sqrt{1 + \frac{2\eta^2}{1+\eta^2}\cos^2\theta}\right)^{-1} < 1$. In particular, when $\theta < \frac{\pi}{3}$, we have $\kappa_2(\theta) < \left(\sqrt{1 + \frac{\eta^2}{1+\eta^2}}\right)^{-1}$.

Proof. Using the explicit expression (5) of the population EM update given in Lemma 1, we compute the sine of the angle θ' between $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}^*$:

$$\begin{aligned}
\sin\theta' &= \frac{Rb_2^*}{\|\boldsymbol{\beta}^*\|\sqrt{R^2 + S^2\|\boldsymbol{\beta}^*\|^2 + 2SRb_1^*}} \\
&= \sin\theta\frac{1}{\sqrt{1 + (S/R)^2\|\boldsymbol{\beta}^*\|^2 + 2(S/R)b_1^*}} \\
&\le \sin\theta\frac{1}{\sqrt{1 + 2(S/R)b_1^*}}.
\end{aligned} \tag{30}$$

Recall that we have defined the quantities $b_1^* = \|\boldsymbol{\beta}^*\|\cos(\theta)$ and $b_2^* = \|\boldsymbol{\beta}^*\|\sin(\theta)$. Since $R > 0$ by Lemma 1, it suffices to prove the lower bound $S \ge \frac{b_1^*}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}R$, which gives us the claimed result by plugging it into (30).

To establish the lower bound on S, we first observe from the expression for S and R in equation (6) that

$$\begin{aligned}
S =& \underbrace{\mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}\sigma_2 z\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right]}_{A} \\
&+ b_1^*\mathbb{E}_{\alpha_1,z}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right] = A + \frac{b_1^*}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}R.
\end{aligned}$$

We claim that $A \ge 0$. Indeed, applying Stein's lemma with respect to z yields

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

We further rewrite the last right hand side as

$$\begin{aligned}
&\mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)z^2\right] \\
=& \frac{1}{2}\mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)z^2\right] + \frac{1}{2}\mathbb{E}_{\alpha_1,z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(-\sigma_2 z + \alpha_1 b_1^*)\right)z^2\right] \\
=& \frac{1}{2}\mathbb{E}\left[\left(\tanh\left(-\frac{\alpha_1 b_1}{\sigma^2}\sigma_2 z + \frac{\alpha_1^2 b_1^* b_1}{\sigma^2}\right) + \tanh\left(\frac{\alpha_1 b_1}{\sigma^2}\sigma_2 z + \frac{\alpha_1^2 b_1^* b_1}{\sigma^2}\right)\right)y^2\right] \ge 0,
\end{aligned}$$

where the last step follows from the numerical inequality that $\tanh(c + x) + \tanh(-c + x) \ge 0$ for all $x \ge 0$ and any real number c. Combining pieces, we obtain that $A \ge 0$ and hence $S \ge \frac{b_1^*}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}R$ as desired.

Finally, note that $\kappa_2(\theta)$ is increasing with respect to θ. It follows that $\kappa_2(\theta) < \kappa_2(\pi/2) = 1$ for all $\theta \in [0, \pi/2)$, and that $\kappa_2(\theta) \le \kappa_2(\frac{\pi}{3}) = \left(\sqrt{1 + \frac{\eta^2}{1+\eta^2}}\right)^{-1}$ when $\theta \le \frac{\pi}{3}$. This proves the last part of Theorem 4. □

7.2 Proof of Theorem 3

We restate the theorem below for readers' convenience.

Theorem 3 (Cosine Convergence). *When $0 \le \theta < \frac{\pi}{2}$, the population EM iteration (4) satisfies*

$$\cos(\theta') \ge \kappa_1(\theta)\cos(\theta), \tag{9}$$

where $\kappa_1(\theta) = \sqrt{1 + \frac{\sin^2(\theta)}{\cos^2(\theta) + \frac{1}{2}(1+\eta^{-2})}}$. In particular, when $\theta \ge \frac{\pi}{3}$, we have $\kappa_1(\theta) \ge \sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}}$. Consequently, if $\cos(\theta_0) = \Theta(1/\sqrt{d})$, after $T = O\big(\max(1, \eta^{-2})\log d\big)$ iterations, we get $\theta_T < \pi/3$ or equivalently $\cos(\theta_T) \ge \frac{1}{2}$.

Proof. Theorem 4 establishes that $\sin\theta' \le \kappa_2(\theta)\sin(\theta)$ for all $\theta \in [0, \frac{\pi}{2})$, with $\kappa_2(\theta) = \left(\sqrt{1 + \frac{2\eta^2}{1+\eta^2}\cos^2(\theta)}\right)^{-1}$. It follows that

$$\begin{aligned}
\cos(\theta') =& \sqrt{1 - \sin^2(\theta')} \\
\ge& \sqrt{1 - \kappa_2(\theta)^2 \sin^2(\theta)} \\
=& \sqrt{\cos^2(\theta) + (1 - \kappa_2(\theta)^2)\sin^2(\theta)} \\
=& \cos(\theta)\sqrt{1 + \frac{1 - \kappa_2(\theta)^2}{\cos^2(\theta)}\sin^2(\theta)} \\
=& \cos(\theta)\sqrt{1 + \frac{\sin^2(\theta)}{\frac{1}{2}(1+\eta^{-2}) + \cos^2(\theta)}} = \cos(\theta)\kappa_1(\theta).
\end{aligned} \tag{31}$$

Since $\kappa_1(\theta)$ is increasing with respect to θ, it follows that $\kappa_1(\theta) \ge \kappa_1(\frac{\pi}{3})$ when $\theta \in [\frac{\pi}{3}, \frac{\pi}{2})$. □

7.3 Proof of Theorem 5

We first state a lemma that is essential for the proof of Theorem 5. Recall the notations defined in Section 2. Also recall the explicit expression (5) for the population EM update, which involves the quantities S and R:

$$\begin{aligned}
S :=& \mathbb{E}_{\alpha_1, z}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right] \\
R :=& (\sigma^2 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}_{\alpha_1, z}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(\sigma_2 z + \alpha_1 b_1^*)\right)\right].
\end{aligned}$$

Lemma 12 (Property of b_1'). *b_1' is increasing in b_1. Consequently, b_1' is upper bounded by the limit value*

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

Proof. We first show that b_1' is increasing in b_1 by making use of the expression (21a) for b_1' previously derived. Differentiating b_1' with respect to b_1 gives

$$\frac{\mathrm{d}b_1'}{\mathrm{d}b_1} = \mathbb{E}_{\alpha_1, \alpha_2}\mathbb{E}_{Y|\alpha_1, \alpha_2}\left[\tanh'\left(\frac{b_1\alpha_1}{\sigma^2}Y\right)Y^2\alpha_1^2\right] \ge 0. \tag{33}$$

Next, we show the limit value of b_1'. Recall that $b_1' = b_1^* S + R$ by equation (5) in Lemma 1. Applying Stein's lemma with respect to z, we may rewrite the term R as

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

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

$$\begin{aligned}
\lim_{b_1 \to \infty} R =& \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}{\sigma_2}\mathbb{E}_{\alpha_1, z}[\mathrm{sign}(\alpha_1(\sigma_2 z + \alpha_1 b_1^*))z\alpha_1] \\
=& \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}{\sigma_2}\left[\frac{1}{\pi}\int_0^\infty 2\alpha_1 e^{-\frac{\alpha_1^2}{2}}\left(\int_{\frac{\alpha_1 b_1^*}{\sigma_2}}^\infty z e^{-\frac{z^2}{2}}\mathrm{d}z\right)\mathrm{d}\alpha_1\right]
\end{aligned}$$

$$= \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}{\sigma_2} \left[\frac{2}{\pi} \int_0^\infty \alpha_1 e^{-\frac{\alpha_1^2 b_1^{*2}}{2\sigma_2^2}} e^{-\frac{\alpha_1^2}{2}} \mathrm{d}\alpha_1 \right]$$
$$= \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}{\sigma_2} \frac{2}{\pi} \frac{\sigma_2^2}{b_1^{*2} + \sigma_2^2} = \frac{2\sigma_2}{\pi}. \tag{34}$$

Turning to the term S, we observe that $\lim_{c\to\infty} cx \tanh'(cx) = 0$ for all x. It follows that

$$\lim_{b_1\to\infty} S = \mathbb{E}_{\alpha_1, z}\left[\operatorname{sign}(\alpha_1(\sigma_2 z + \alpha_1 b_1^*))\right]$$
$$= \frac{1}{\pi} \int_0^\infty \left(\int_{-\frac{\alpha_1 b_1^*}{\sigma_2}}^{\frac{\alpha_1 b_1^*}{\sigma_2}} e^{-\frac{z^2}{2}} dz \right) e^{-\frac{\alpha_1^2}{2}} \mathrm{d}\alpha_1$$
$$= \frac{2}{\pi} \int_0^\infty \left(\int_0^{\frac{\alpha_1 b_1^*}{\sigma_2}} e^{-\frac{z^2}{2}} dz \right) e^{-\frac{\alpha_1^2}{2}} \mathrm{d}\alpha_1 = \frac{2}{\pi} \tan^{-1}(b_1^*/\sigma_2). \tag{35}$$

Plugging equations (34) and (35) into $\lim_{b_1\to\infty} b_1' = b_1^* \lim_{b_1\to\infty} S + \lim_{b_1\to\infty} R$, we obtain the limit value of b_1', thereby completing the proof of the lemma. □

We are now ready to prove Theorem 5, which is restated below.

Theorem 5 (ℓ_2 Contraction)**.** *Suppose we have that $\theta < \pi/8$. Recall the shorthands $b_1 := \|\boldsymbol{\beta}\|$, $b_1^* := \|\boldsymbol{\beta}^*\| \cos(\theta)$, $b_2^* := \|\boldsymbol{\beta}^*\| \sin(\theta)$ and $\sigma_2^2 := \sigma^2 + b_2^{*2}$. The following holds for the population EM iteration (4):*

- *If $b_2^* < \sigma$ or $\frac{\sigma_2^2}{\sigma^2} b_1 < b_1^*$, then*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le \kappa_3(\theta)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \kappa_3(\theta)(16\sin^3\theta)\|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}, \tag{11a}$$

  where $\kappa_3(\theta) = \left(\sqrt{1 + \min\left(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*\right)^2 / \sigma_2^2} \right)^{-1}$.

- *If $b_2^* \ge \sigma$ and $\frac{\sigma_2^2}{\sigma^2} b_1 > b_1^*$, we have*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le 0.6 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{11b}$$

Proof. Using the basis system introduced in Section 2.1, we can write $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 = |b_1' - b_1^*|^2 + |b_2' - b_2^*|^2$. The second term $|b_2' - b_2^*|$ can be as

$$0 \le (b_2^* - b_2') = (1 - S) b_2^* \le \left(\sqrt{1 + \min\left(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*\right) b_1^* / \sigma_2^2} \right)^{-1} b_2^* \le \kappa_3(\theta) b_2^*, \tag{36}$$

where we use the lower bound of S from Lemma 6.

It remains to upper bound $|b_1' - b_1^*|$. We shall make use of the following consistency property of the population EM update: When $b_1 = \frac{\sigma^2}{\sigma_2^2} b_1^*$, the expression (22) for b_1' gives that

$$b_1' = \mathbb{E}_{\alpha_1} \alpha_1 \left[\mathbb{E}_{Y|\alpha_1 \sim \mathcal{N}(\alpha_1 b_1^*, \sigma_2^2)} \tanh\left(\frac{\alpha_1 b_1^*}{\sigma_2^2} Y \right) Y \right] = \mathbb{E}_{\alpha_1}[\alpha_1^2 b_1^*] = b_1^*. \tag{37}$$

We separate the analysis into three cases.

Case I. $b_1 \le \frac{\sigma^2}{\sigma_2^2} b_1^*$: In this case, we have

$$b_1' - \frac{\sigma_2^2}{\sigma^2} b_1 \overset{(i)}{=} \mathbb{E}_{\alpha_1} \left[\alpha_1 \mathbb{E}_{\substack{Y|\alpha_1 \\ \sim\mathcal{N}(\alpha_1 b_1^*, \sigma_2^2)}} \left[\tanh\left(\frac{\alpha_1 (\frac{\sigma_2^2}{\sigma^2} b_1)}{\sigma_2^2} Y \right) Y \right] - \alpha_1 \mathbb{E}_{\substack{Y|\alpha_1 \\ \sim\mathcal{N}(\alpha_1 \frac{\sigma_2^2}{\sigma^2} b_1, \sigma_2^2)}} \left[\tanh\left(\frac{\alpha_1 (\frac{\sigma_2^2}{\sigma^2} b_1)}{\sigma_2^2} Y \right) Y \right] \right]$$
$$\overset{(ii)}{\ge} \left(b_1^* - \frac{\sigma_2^2}{\sigma^2} b_1 \right) \mathbb{E}_{\alpha_1} \left[\alpha_1^2 \min_{\mu \in (\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*)} \frac{\partial}{\partial \mu} \left(\mathbb{E}_{z\sim\mathcal{N}(0,1)} \left[\tanh\left(\frac{\alpha_1 (\frac{\sigma_2^2}{\sigma^2} b_1)}{\sigma_2^2} (z + \mu) \right) (z + \mu) \right] \right) \right]$$

$$\overset{(iii)}{\geq} \left(b_1^* - \frac{\sigma_2^2}{\sigma^2} b_1\right) \mathbb{E}_{\alpha_1}\left[\alpha_1^2 \left(1 - \exp\left(-\frac{\alpha_1^2 \min\left(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*\right)^2}{2\sigma_2^2}\right)\right)\right], \tag{38}$$

where in step (i) we use the consistency property in equation (37), in step (ii) we use mean-value theorem along with the case assumption, and in step (iii) we apply Lemmas 7 and 8. Consequently, after some algebra we obtain that

$$0 \overset{(i)}{\leq} b_1^* - b_1' \leq \kappa_3^3(\theta)\left(b_1^* - \frac{\sigma_2^2}{\sigma^2} b_1\right) \leq \kappa_3^3(\theta)(b_1^* - b_1) \leq \kappa_3(\theta)(b_1^* - b_1), \tag{39}$$

where the inequality (i) holds thanks to Lemma 12, which states that b_1' is increasing in b_1. Combining the bounds (36) and (39), we obtain that

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}'\| \leq \kappa_3(\theta)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|.$$

Case II. $b_1 > \frac{\sigma^2}{\sigma_2^2} b_1^*$, $\sigma > b_2^*$: Following a similar procedure as above in equation (38), we have

$$0 \leq b_1' - b_1^* \leq \kappa_3^3(\theta)\left(\frac{\sigma_2^2}{\sigma^2} b_1 - b_1^*\right) = \kappa_3^3(\theta)(b_1 - b_1^*) + \kappa_3^3(\theta)\frac{b_2^{*2}}{\sigma^2} b_1. \tag{40}$$

By the case condition, we $\kappa_3(\theta) = \left(\sqrt{1 + \frac{b_1^{*2}}{\sigma_2^2}}\right)^{-1} = \sqrt{\frac{\sigma^2 + b_2^{*2}}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}}$. We further divide the analysis into two subcases:

Case II(a): Suppose that $b_1 > 2b_1^*$ or equivalently, $b_1 < 2(b_1 - b_1^*)$. Then we have

$$\begin{aligned} b_1' - b_1^* &\leq \kappa_3^3(\theta)(b_1 - b_1^*)\left(1 + 2\frac{b_2^{*2}}{\sigma^2}\right) \\ &= \kappa_3(\theta)(b_1 - b_1^*)\left(\frac{\sigma^2 + b_2^{*2}}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\right)\left(1 + \frac{2b_2^{*2}}{\sigma^2}\right) \\ &= \kappa_3(\theta)\underbrace{\left(\frac{\sigma^2 + b_2^{*2}}{\sigma^2 + b_1^{*2} + b_2^{*2}} \frac{\sigma^2 + 2b_2^{*2}}{\sigma^2}\right)}_{A}(b_1 - b_1^*). \end{aligned}$$

Note that the term A is less than 1 since the nominator is no bigger than the denominator. Indeed, we have

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

where step (i) holds because $b_2^* < \sigma$ and step (ii) holds because $\frac{b_2^*}{b_1^*} = \tanh(\theta) = \tan\frac{\pi}{8} < 1/2$. It follows that

$$0 \leq b_1' - b_1^* \leq \kappa_3(\theta)(b_1 - b_1^*),$$

in which case we have $\|\boldsymbol{\beta'} - \boldsymbol{\beta}^*\| \leq \kappa_3(\theta)\|\boldsymbol{\beta} - \boldsymbol{\beta^*}\|$ as desired.

Case II(b): Suppose that $b_1 < 2b_1^*$. Note that we can assume that $b_1 \frac{b_2^{*2}}{\sigma^2} \geq (\frac{1}{\kappa_3^2(\theta)} - 1)(b_1 - b_1^*)$. Otherwise, we can easily get $0 \leq b_1' - b_1^* \leq \kappa_3(b_1 - b_1^*)$ similarly by plugging the condition $b_1 \frac{b_2^{*2}}{\sigma^2} \leq (\frac{1}{\kappa_3^2(\theta)} - 1)(b_1 - b_1^*)$ into the inequality (40). Squaring both sides of the inequality (40), we obtain that

$$\begin{aligned} (b_1' - b_1^*)^2 &\leq \kappa_3^6(\theta)(b_1 - b_1^*)^2 + \kappa_3^6(\theta)\left(2\left(\frac{b_2^*}{\sigma}\right)^2 b_1(b_1 - b_1^*) + \left(\frac{b_2^*}{\sigma}\right)^4 b_1^2\right) \\ &\leq \kappa_3^6(\theta)(b_1 - b_1^*)^2 + \kappa_3^6(\theta)\left(\frac{b_2^*}{\sigma}\right)^4 b_1^2\left(\frac{2\kappa_3^2(\theta)}{1 - \kappa_3^2(\theta)} + 1\right) \\ &= \kappa_3^6(\theta)(b_1 - b_1^*)^2 + \underbrace{\kappa_3^6(\theta)\left(\frac{b_2^*}{\sigma}\right)^4 b_1^2\left(\frac{2\sigma^2 + 2b_2^{*2} + b_1^{*2}}{b_1^{*2}}\right)}_{B}. \end{aligned}$$

We bound the term B as follows:

$$\begin{aligned}
B &= \kappa_3^6(\theta)\left(\frac{b_2^*}{\sigma}\right)^4 b_1^2\left(\frac{2\sigma^2+2b_2^{*2}+b_1^{*2}}{b_1^{*2}}\right)\\
&= \kappa_3^2(\theta)\left(\frac{b_2^*}{\sigma}\right)^4 b_1^2\left(\frac{2\sigma^2+2b_2^{*2}+b_1^{*2}}{b_1^{*2}}\right)\left(\frac{\sigma^2+b_2^{*2}}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}\right)^2\\
&= \kappa_3^2(\theta)b_2^{*4}\left(\frac{b_1^2}{b_1^{*2}}\right)\left(\frac{2\sigma^2+2b_2^{*2}+b_1^{*2}}{\sigma^2+b_2^{*2}+b_1^{*2}}\right)\left(\frac{(\sigma^2+b_2^{*2})^2}{\sigma^4}\right)\frac{1}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}\\
&\overset{(i)}{\leq} \kappa_3^2(\theta)b_2^{*4}\cdot 4\cdot 2\cdot 4\cdot\left(\frac{1}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}\right) = \kappa_3^2(\theta)\frac{32b_2^{*2}}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}b_2^{*2},
\end{aligned}$$

where the inequality (i) follows from the assumption that $b_1 < 2b_1^*$ and $b_2^* < \sigma$. Therefore, we get $(b_1'-b_1^*)^2 \leq \kappa_3^2(\theta)(b_1-b_1^*)^2 + \kappa_3^2(\theta)\frac{32b_2^{*2}}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}b_2^{*2}$. Combining this bound with $(b_2'-b_2^*)^2 \leq \kappa_3^2(\theta)(b_2-b_2^*)^2$, we obtain

$$\|\boldsymbol{\beta}'-\boldsymbol{\beta}^*\|^2 \leq \kappa_3^2(\theta)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2 + \kappa_3^2(\theta)\frac{32b_2^{*2}}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}b_2^{*2}.$$

We further upper bound the last right hand side using the inequality $\sqrt{a^2+b^2} \leq a+\frac{b^2}{2a}$. Doing so and recalling the definition of the SNR $\eta := \frac{\|\boldsymbol{\beta}^*\|}{\sigma}$ gives

$$\begin{aligned}
\|\boldsymbol{\beta}'-\boldsymbol{\beta}^*\| &\leq \kappa_3(\theta)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\| + \kappa_3(\theta)\frac{16b_2^{*2}}{\sigma^2+\|\boldsymbol{\beta}^*\|^2}\frac{b_2^*}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|}b_2^*\\
&\leq \kappa_3(\theta)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\| + \kappa_3(\theta)(16\sin^3\theta)\|\boldsymbol{\beta}^*\|\frac{\eta^2}{1+\eta^2},
\end{aligned}$$

where we use the fact that $b_2^* = \|\boldsymbol{\beta}^*\|\sin(\theta)$ and $\frac{b_2^*}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|} = \frac{b_2^*}{\sqrt{(b_1-b_1^*)^2+(b_2^*)^2}} \leq 1$.

Case III. $b_1 > \frac{\sigma^2}{\sigma_2^2}b_1^*$, $\sigma < b_2^*$*:* In this case, we are able to establish a constant rate of contraction in local region with high SNR.

First note that $b_1' \geq b_1^*$ and the difference $(b_1'-b_1^*)$ is increasing in b_1. Therefore, invoking Lemma 12 yields

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

where we use the fact that $\sigma_2^2 = \sigma^2+b_2^{*2} \leq 2b_2^{*2}$, $\tan^{-1}\left(\frac{b_1^*}{b_2^*}\right) = \frac{\pi}{2}-\theta$, and $b_1^* = b_2^*\cot\theta$. One can verify that $\theta\cot\theta$ is decreasing in $[0,\frac{\pi}{2}]$. It follows that

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

On the other hand, we have

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

Combining the above two bounds, we obtain that

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

thereby completing the proof of Theorem 5. □

7.4 Proof of Corollary 1

We restate the corollary below for readers' convenience.

Corollary 1 (ℓ_2 Convergence)**.** *Suppose that the initial solution satisfies $\theta_0 < \pi/8$. There exists a constant $\kappa < 1$ such that after T iterations of the population EM, we have the error bound*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| < \kappa^T \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + T\kappa^T \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}. \tag{12}$$

In particular, the constant κ can be taken to be the maximum among

$$0.6,\ \sqrt{\left(1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}\right)^{-1}},\ \sqrt{1 - \frac{0.8\eta^2}{1+\eta^2}}. \tag{13}$$

Proof. Recall that θ is the angle between the current iterate and $\boldsymbol{\beta}^*$, and that θ' is the angle between the next iterate and $\boldsymbol{\beta}^*$. The sine convergence result in Theorem 4 ensures that $\theta' < \theta$. Recall Theorem 5, which establishes contraction of the ℓ_2 distance in each iteration. We shall first show that the contraction ratio decreases as angle gets smaller, that is, $\kappa_3(\theta') \le \kappa_3(\theta)$ or equivalently

$$\min\left(\frac{\sigma_2'^2}{\sigma^2} b_1', b_1'^*\right)^2 / \sigma_2'^2 \ge \min\left(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*\right)^2 / \sigma_2^2. \tag{41}$$

Since $\theta' < \theta$, we have $b_1'^* > b_1^*$, $b_2'^* < b_2^*$ and $\sigma_2' < \sigma_2$. The analysis is divided into two cases.

- If $\frac{\sigma_2^2}{\sigma^2} b_1 \ge b_1^*$, then the right hand side of equation (41) is b_1^{*2}/σ_2^2. From equation (40), we have $b_1' \ge b_1^*$. Thus the left hand side of equation (41) satisfies

$$\min\left(\frac{\sigma_2'^2}{\sigma^2} b_1', b_1'^*\right)^2 / \sigma_2'^2 \ge \min\left(\frac{\sigma_2'^2}{\sigma^2} b_1^*, b_1^*\right)^2 / \sigma_2^2 \ge b_1^{*2}/\sigma_2^2.$$

- If $\frac{\sigma_2^2}{\sigma^2} b_1 < b_1^*$, then the right hand side of equation (41) is $\frac{\sigma_2^2}{\sigma^2} b_1/\sigma^2$. From equation (38), we have $b_1' > \frac{\sigma_2^2}{\sigma^2} b_1$. Thus the left hand side of equation (41) satisfies

$$\min\left(\frac{\sigma_2'^2}{\sigma^2} b_1', b_1'^*\right)^2 / \sigma_2'^2 \ge \min\left(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*\right)^2 / \sigma_2^2 = \frac{\sigma_2^2}{\sigma^2} b_1/\sigma^2.$$

Combining the above two cases, we have shown that $\kappa_3(\theta') \le \kappa_3(\theta)$. This result implies that the ℓ_2 contraction ratio $\kappa_3(\theta_t)$ for the t-th iteration can be uniformly upper bounded by $\kappa_3(\theta_0)$. We also recall that $\kappa_2(\theta_0)$ is the corresponding contraction ratio for the sine convergence. We claim that their maximum, $\max(\kappa_3(\theta_0), \kappa_2^3(\theta_0))$, is upper bounded by the quantity κ defined in the statement of Corollary 1. Indeed, we have the bound

$$\begin{aligned}
&\max(\kappa_3(\theta_0), \kappa_2^3(\theta_0)) \\
&\overset{(i)}{\le} \max\left(0.6, \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2} b_1, b_1^*)^2}{\sigma_2^2}}\right)^{-1}, \left(\sqrt{1 + \frac{2b_1^{*2}}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}}\right)^{-3}\right) \\
&\le \max\left(0.6, \left(\sqrt{1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}}\right)^{-1}, \left(\sqrt{1 + \frac{\eta^2 \cos^2\theta_0}{1 + \eta^2 \sin^2\theta_0}}\right)^{-1}, \left(\sqrt{1 + \frac{2\eta^2 \cos^2\theta_0}{1+\eta^2}}\right)^{-3}\right) \\
&\overset{(ii)}{\le} \max\left(0.6, \left(\sqrt{1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}}\right)^{-1}, \sqrt{1 - \frac{0.8\eta^2}{1+\eta^2}}\right) := \kappa.
\end{aligned}$$

Here step (i) holds because the first two quantities correspond to the two possible contraction rates in Theorem 5, and the third quantity corresponds to $\kappa_2(\theta_0)^3$; step (ii) holds since $\theta_0 < \pi/8$.

With the above bound on the contraction ratios, we can then apply Theorem 5 to the t-th iteration of population EM to obtain

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\| \le \kappa \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\| + \kappa(16 \sin^3\theta_t) \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}$$

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

where the recursion step above holds because Theorem 4 ensures that $\sin^3(\theta_t) \le \kappa_2^3(\theta_{t-1}) \sin(\theta_{t-1}) \le \kappa_2^3(\theta_0) \sin(\theta_{t-1}) \le \kappa \sin^3(\theta_{t-1})$ for all $t \ge 1$, and the last inequality above holds because $\theta_0 < \pi/8$. This completes the proof of Corollary 1. □

8 Proofs for Finite-Sample EM in Middle-High SNR Regimes

In this section, we prove the technical results in Section 4 on finite-sample EM. We need the following norm conditions for the estimator can be naturally met as the EM iteration proceeds:

Lemma 13 (Norm Bounds)**.** *If $\|\widetilde{\boldsymbol{\beta}}\| \le \|\boldsymbol{\beta}^*\|/10$, then*

$$
\|\widetilde{\boldsymbol{\beta}}'\| \ge \|\widetilde{\boldsymbol{\beta}}\| (1 + d_1 \cdot \min\{1, (\|\widetilde{\boldsymbol{\beta}}\|/\sigma)^2\}).
$$

Otherwise, if $\|\widetilde{\boldsymbol{\beta}}\| \ge \|\boldsymbol{\beta}^*\|/10$, then we have

$$
\|\widetilde{\boldsymbol{\beta}}'\| \ge \frac{\|\boldsymbol{\beta}^*\|}{10} (1 + d_2 \cdot \min\{1, \eta^2\}).
$$

for some universal constants $d_1, d_2 > 0$. Furthermore, for every $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^d$, we have $\|\widetilde{\boldsymbol{\beta}}'\| \le 3\sqrt{\|\boldsymbol{\beta}^*\|^2 + \sigma^2}$.

Lemma 13 states that if we start from $\|\boldsymbol{\beta}^*\|/10$, then we stably remain above $\|\boldsymbol{\beta}^*\|/10$. On the other hand, if we start from small initialization, then we can wait for initial $O(\min(1, \|\widetilde{\boldsymbol{\beta}}\|/\sigma)^{-2})$ iterations for the starting estimator to become larger than $\|\boldsymbol{\beta}^*\|/10$. We defer the proofs to Appendix A.4.

8.1 Proof of Lemma 4

We restate the lemma below for readers' convenience.

Lemma 4. *With probability at least $1 - \delta$, the following holds for all $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}\| \le C\sqrt{\|\boldsymbol{\beta}^*\|^2 + \sigma^2}$ for some universal constant $C > 0$:*

$$
\cos(\widetilde{\theta}') \ge \kappa_1(\theta)(1 - 10\epsilon_f) \cos(\theta) - \epsilon_f, \tag{17}
$$

$$
\sin^2(\widetilde{\theta}') \le \kappa_2^2(\theta) \sin^2(\theta) + O(\epsilon_f), \tag{18}
$$

where $\kappa_1(\theta) = \sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \ge 1$, and $\kappa_2(\theta) = \left(1 + \frac{2\eta^2}{1+\eta^2} \cos^2\theta\right)^{-1} < 1$.

Proof. Since $n \ge d \log^2(n/\delta)/\epsilon_f^2$ with $\epsilon_f := c\sqrt{d \log^2(n/\delta)/n}$, with probability at least $1 - \delta$, from Lemma 3 it follows that

$$
\left| \langle \widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle \right| \le \|\boldsymbol{\beta}^*\| \|\boldsymbol{\beta}\| \cdot O(\epsilon_f), \tag{42}
$$

$$
\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \le \|\boldsymbol{\beta}\| \cdot O(\epsilon_f). \tag{43}
$$

The cosine of the angle between $\widetilde{\boldsymbol{\beta}}'$ and $\boldsymbol{\beta}^*$ can be bounded as follows,

$$
\begin{aligned}
\cos\widetilde{\theta}' = \frac{\langle \widetilde{\boldsymbol{\beta}}', \boldsymbol{\beta}^* \rangle}{\|\widetilde{\boldsymbol{\beta}}'\| \, \|\boldsymbol{\beta}^*\|} &= \frac{\langle \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle}{\|\widetilde{\boldsymbol{\beta}}'\| \, \|\boldsymbol{\beta}^*\|} + \frac{\langle \widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle}{\|\widetilde{\boldsymbol{\beta}}'\| \, \|\boldsymbol{\beta}^*\|} \\
&= \cos\theta' \frac{\|\boldsymbol{\beta}'\|}{\|\widetilde{\boldsymbol{\beta}}'\|} + \frac{\langle \widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle}{\|\widetilde{\boldsymbol{\beta}}'\| \, \|\boldsymbol{\beta}^*\|} \\
&\ge \cos\theta' \left(1 - \frac{\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\|}{\|\widetilde{\boldsymbol{\beta}}'\|}\right) - \frac{|\langle \widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle| / \|\boldsymbol{\beta}^*\|}{\|\widetilde{\boldsymbol{\beta}}'\|}
\end{aligned}
$$

where the last step from the triangle inequality. To proceed, we apply the concentration bounds (42) and (43); we also make use of Lemma 13 which ensures $\|\widetilde{\boldsymbol{\beta}}'\| = O(\|\boldsymbol{\beta}\|)$. It follows that

$$\begin{aligned}\cos\widetilde{\theta}' &\geq \cos\theta'(1 - O(\epsilon_f)) - O(\epsilon_f) \\ &\geq \kappa(\theta)(1 - O(\epsilon_f))\cos\theta - O(\epsilon_f), \end{aligned} \tag{44}$$

where the last step follows from Theorem 3 on the cosine convergence of population EM.

Turning to the sine of the angle, we have that

$$\begin{aligned}\sin^2\widetilde{\theta}' &= 1 - \cos^2\widetilde{\theta}' \\ &\leq 1 - \cos^2\theta' + \cos^2\theta' - \cos^2\widetilde{\theta}' \\ &\overset{\text{(i)}}{=} \sin^2\theta' + O(\epsilon_f) \\ &\overset{\text{(ii)}}{\leq} \kappa'(\theta)\sin^2\theta + O(\epsilon_f), \end{aligned} \tag{45}$$

where in step (i) we use a similar approach as before to obtain the bound $|\cos^2\theta' - \cos^2\widetilde{\theta}'| \leq 2|\cos(\theta') - \cos(\widetilde{\theta}')| = O(\epsilon_f)$, and in step (ii) we use Theorem 4 on the sine convergence of population EM. □

8.2 Proof of Corollary 2

We restate the corollary below for readers' convenience.

Corollary 2. *If $\epsilon_f \leq c_1 \min(1, \eta^2)$ for a sufficiently small constant $c_1 > 0$, then with probability $1 - \delta$, we have $\widetilde{\theta}' < \theta$ in each iteration of Phases 2, and $\widetilde{\theta}' \leq \frac{\pi}{25}$ in Phase 3.*

Proof. We consider three cases for θ. When $\theta \geq \frac{\pi}{3}$, by inequality (17) we have

$$\begin{aligned}\cos(\widetilde{\theta}') &\geq \kappa_1(\theta)(1 - 10\epsilon_f)\cos(\theta) - O(\epsilon_f) \\ &\overset{\text{(i)}}{\geq} \kappa_1(\theta)(1 - 10\epsilon_f)\cos(\theta) - \cos(\theta)O(\epsilon_f) \\ &\overset{\text{(ii)}}{\geq} \cos(\theta)\left(\kappa_1\left(\frac{\pi}{3}\right)(1 - 10\epsilon_f) - O(\epsilon_f)\right) \end{aligned}$$

where step (i) holds since the output of Phase 1 satisfies $\cos(\theta) = \Omega\left(\max(1, \eta^{-2}) \cdot \epsilon_f\right)$, and step (ii) holds since $\kappa_1(\theta)$ is increasing in θ. Since $\epsilon_f \leq c_1 \min(1, \eta^2)$ for a sufficiently small c_1 by assumption, we have $\kappa_1(\frac{\pi}{3})(1 - 10\epsilon_f) - O(\epsilon_f) > 1$ and hence $\widetilde{\theta}' < \theta$ in Phase 2 as desired.

When $\frac{\pi}{25} \leq \theta < \frac{\pi}{3}$, againby inequality (18) we have

$$\begin{aligned}\sin^2(\widetilde{\theta}') &\leq \kappa_2(\theta)\sin^2(\theta) + O(\epsilon_f) \\ &\leq \kappa_2\left(\frac{\pi}{3}\right)\sin^2(\theta) + O(\epsilon_f), \end{aligned}$$

where the last step holds because $\kappa_2(\theta)$ is increasing in θ). Under our assumption on ϵ_f and the case assumption on θ, we have $\kappa_2\left(\frac{\pi}{3}\right)\sin^2(\theta) + O(\epsilon_f) < \sin^2(\theta)$ and hence $\widetilde{\theta}' < \theta$ in Phase 3 as desired.

Finally, when $\theta \leq \frac{\pi}{25}$, again by inequality (18) we have

$$\begin{aligned}\sin^2(\widetilde{\theta}') &\leq \kappa_2(\theta)\sin^2(\theta) + O(\epsilon_f) \\ &\leq \kappa_2\left(\frac{\pi}{25}\right)\sin^2\left(\frac{\pi}{25}\right) + O(\epsilon_f), \end{aligned}$$

where the last step holds by the increasing property of $\kappa_2(\cdot)$ and the case assumption on θ. Under our assumption on ϵ_f, we have $\kappa_2\left(\frac{\pi}{25}\right)\sin^2\left(\frac{\pi}{25}\right) + O(\epsilon_f) \leq \sin^2(\frac{\pi}{25})$ and hence $\widetilde{\theta}' \leq \frac{\pi}{25}$ in Phase 3 as desired. □

8.3 Proofs of Theorem 6

We first prove Theorem 6, which is restated below for readers' convenience.

Theorem 6 (Cosine Convergence, Finite-Sample)**.** *Suppose that $\widetilde{\boldsymbol{\beta}}^{(0)}$ is an iterate obtained from Phase 1. We run the finite-sample EM with $n = \max(1, \eta^{-2})d/\epsilon_f^2$ samples. As long as $\widetilde{\theta}^{(t)} > \pi/25$ for all $t < T$, there exists an universal constant $c_1 > 0$ such that with probability $1 - \delta$,*

$$\cos(\widetilde{\theta}^{(t)}) \geq \left(1 + c_1 \cdot \min(1, \eta^2)\right) \cdot \cos(\widetilde{\theta}^{(t-1)}). \tag{19}$$

In particular, if $\cos(\widetilde{\theta}^{(0)}) = \Theta(1)$, then we have $\cos(\widetilde{\theta}^{(T)}) \geq 0.95$ after $T = O\big(\max(1, \eta^{-2})\log d\big)$ iterations.

Proof. The key idea in the above theorem is that when we bound the statistical error of cosine value, we need to bound an error in one fixed direction $u := \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|$ instead of all directions in $\mathbb{R}^d$ to bound ℓ_2 norm. More specifically, we first express the cosine value after one-step iteration:

$$\begin{aligned}
\cos\widetilde{\theta}' &= \frac{(\boldsymbol{\beta}^*)^\top \widetilde{\boldsymbol{\beta}}'}{\|\widetilde{\boldsymbol{\beta}}'\|\|\boldsymbol{\beta}^*\|} \\
&= \frac{u^\top(\boldsymbol{\beta}' - \tilde{\boldsymbol{\beta}}')}{\|\tilde{\boldsymbol{\beta}}'\|} + \frac{u^\top \boldsymbol{\beta}'}{\|\boldsymbol{\beta}'\|}\frac{\|\boldsymbol{\beta}'\|}{\|\widetilde{\boldsymbol{\beta}}'\|}, \\
&\geq -\max\left(\frac{\epsilon_f}{\sqrt{d}}, \epsilon_f^2\right) \cdot \frac{r}{\|\widetilde{\boldsymbol{\beta}}'\|} + \frac{u^\top \boldsymbol{\beta}'}{\|\boldsymbol{\beta}'\|}\frac{\|\boldsymbol{\beta}'\|}{\|\boldsymbol{\beta}'\| + r\epsilon_f} \\
&\geq \kappa_t \cos\alpha_t \left(1 - \frac{r\epsilon_f}{\|\boldsymbol{\beta}'\|}\right) - \max\left(\frac{\epsilon_f}{\sqrt{d}}, \epsilon_f^2\right)\frac{r}{\|\boldsymbol{\beta}'\| - r\epsilon_f},
\end{aligned}$$

where the last inequality comes from Theorem 3 for the population EM.

We need to show that we have $r/\|\boldsymbol{\beta}'\| = O(1)$. If this is true, we can set ϵ_f as some sufficiently small absolute constant (that does not depend on η). To show this, we apply Lemma 3 for several values of $r = C_0, C_0 2^{-1}, ..., C_0 2^{-l+1}, C_0 2^{-l}$ where $C_0 = 3C$ and $l = O(\log(n/d))$. We can replace δ by $\delta/\log(n/d)$ for union bound, which does not change the order of statistical error. Pick k such that $C_0 2^{-k} \leq \|\boldsymbol{\beta}\| \leq C_0 2^{-k+1} = r$.

When $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}^*\|/10$, we can apply the Lemma 13 to see

$$r/\|\boldsymbol{\beta}'\| \leq C_0 2^{-k+1}/(C_0 2^{-k}) = 2,$$

where we used $r = 2^{-k+1}$. Therefore, $r/\|\boldsymbol{\beta}'\| = O(1)$. On the other hand, if $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$, then we divide the cases when $\|\boldsymbol{\beta}^*\| \geq 1/\max(3, c_2)$ where $c_2 > 0$ satisfies the lower bound given in Lemma 5:

$$\|\boldsymbol{\beta}'\| \geq \|\boldsymbol{\beta}\|(1 - 4\|\boldsymbol{\beta}\|^2) - c_2\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2.$$

When $\|\boldsymbol{\beta}^*\| \geq 1/\max(3, c_2)$ and $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$, by Lemma 13 we have $r/\|\boldsymbol{\beta}'\| \leq C_0 \max(3, c_2) = O(1)$ since all parameters here are universal constants. On the other hand, if $\|\boldsymbol{\beta}^*\| \leq 1/\max(3, c_2)$ and $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$, then from Lemma 13 we have

$$\|\boldsymbol{\beta}'\| \geq \|\boldsymbol{\beta}\|(1 - 3\|\boldsymbol{\beta}\|^2) - c_2\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2 \geq \|\boldsymbol{\beta}\|/2.$$

Therefore, $r/\|\boldsymbol{\beta}'\| \leq C_0 2^{-k+1}/(C_0 2^{-k-1}) = 4 = O(1)$.

From the above case study, we have that

$$\cos\widetilde{\theta}_{t+1} \geq \kappa_t \cos\widetilde{\theta}_t(1 - c_1\epsilon_f) - c_2 \max\left(\frac{\epsilon_f}{\sqrt{d}}, \epsilon_f^2\right),$$

for some absolute constants $c_1, c_2 > 0$. Now observe that as long as $\sin\theta_t > c_\theta$, $\kappa_t = 1 + c_3 \min\{1, \eta^2\}$ for some sufficiently small constant $c_\theta, c_3 > 0$. Also, recall that we are considering the middle-to-high SNR regime when $\eta^2 \geq c_\eta\sqrt{d\log^2(n/\delta)/n}$ for some sufficiently large constant $c_\eta > 0$, whereas $\epsilon_f \leq c\sqrt{d\log^2(n/\delta)/n}$ for another fixed constant $c > 0$. Therefore, there exists a universal constant $c_4 > 0$ such that for all $\cos\widetilde{\theta} \geq \epsilon_f$, we have

$$\cos\widetilde{\theta}_{t+1} \geq (1 + c_4 \min(1, \eta^2))\cos\widetilde{\theta}_t.$$

After $t = O(\eta^{-2}\log(d))$ iterations starting from $\cos\widetilde{\theta}_0 = 1/\sqrt{d}$, we have $\cos\widetilde{\theta}_t \geq 0.95$ or $\sin\widetilde{\theta}_t \leq 0.1$. □

8.4 Stability and Convergence after Alignment

In this subsection, we see how the alignment is stabilized and the norm increases in case we start from small initialization.

Sine stays below some threshold. Once $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ are well-aligned, using $\sin^2\theta = 1 - \cos^2\theta$, similar arguments can be applied for sin values:

$$\begin{aligned}
&\sin^2\widetilde{\theta}' \leq (1 - c_1\min(1, \eta^2))\sin^2\widetilde{\theta}, && \text{if} \quad \sin^2\widetilde{\theta} \geq c_2 \\
&\sin^2\widetilde{\theta}' \leq c_2, && \text{else} \quad \sin^2\widetilde{\theta} \leq c_2,
\end{aligned}$$

for some absolute constants $c_1 > 0$ and sufficiently small $0 < c_2 < 0.01$ given that $\cos\widetilde{\theta} > 0.95$.

Initialization from small estimators after alignment. After the angle is aligned such that $\sin\theta \le c_2$. We see how fast $\|\boldsymbol{\beta}\|$ enters the desired initialization region that Theorem 7 requires, when $\|\boldsymbol{\beta}\| \le 0.9\|\boldsymbol{\beta}^*\|$.

Let us first consider the case $0.1\|\boldsymbol{\beta}^*\| \le \|\boldsymbol{\beta}\| \le 0.9\|\boldsymbol{\beta}^*\|$. We recall Theorem 5 such that

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

where $\kappa < 1 - c_3\eta^2$ for some absolute constant c_3. By appropriately setting c_2 and c_3, we have

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}'\| \le (1 - c_4\min(1,\eta^2))\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|,$$

for some constant $c_4 > 0$. Since we are in the regime $\eta^2 \ge c_\eta\sqrt{d\log^2(n/\delta)/n}$ for sufficiently large c_η, by appropriately setting the constants we have $\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^*\| \le (1 - c_5\min(1,\eta^2))\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$ for some absolute constant $c_5 > 0$, as long as we are in the region $0.1\|\boldsymbol{\beta}^*\| \le \|\boldsymbol{\beta}\| \le 0.9\|\boldsymbol{\beta}^*\|$. Hence after $O(\max(1,\eta^{-2}))$ iterations, we reach to the desired initialization region.

Now we consider the case $\|\boldsymbol{\beta}\| \le 0.1\|\boldsymbol{\beta}^*\|$. In this case, by Lemma 13, we can show that

$$\|\boldsymbol{\beta}'\| \ge \|\boldsymbol{\beta}\|(1 + c_6\min\{1, \|\boldsymbol{\beta}\|^2, \|\boldsymbol{\beta}^*\|^2\}),$$

for some universal constant $c_6 > 0$. After $O(\max\{\|\boldsymbol{\beta}\|^{-2}, \|\boldsymbol{\beta}^*\|^{-2}\})$ iterations, we enter $\|\boldsymbol{\beta}\| \ge \|\boldsymbol{\beta}^*\|/10$. Note that when we start with $\|\widetilde{\boldsymbol{\beta}}_0\| = \Omega(1)$, $\|\widetilde{\boldsymbol{\beta}}_t\|$ will stay above $\min\{\Omega(1), \|\theta^*\|/10\}$ throughout all iterations due to Lemma 13 and Lemma 5.

8.5 Proof of Theorem 7

We restate the theorem below for readers' convenience.

Theorem 7 (ℓ_2 Convergence, Finite-Sample in Middle-to-High SNR Regimes)**.** *Suppose that $\widetilde{\boldsymbol{\beta}}_0$ is an iterate obtained from Phase 2 whose angle with $\boldsymbol{\beta}^*$ satisfies $\widetilde{\theta}_0 < \frac{\pi}{25}$. Furthermore, suppose that $\|\widetilde{\boldsymbol{\beta}}_0\| \ge 0.9\|\boldsymbol{\beta}^*\|$. Then, for any $\delta > 0$, there exist universal constants $C_1, C_2 > 0$ such that with probability at least $1 - \delta$,*

$$\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \le C_1\sigma\max\{1,\eta^{-1}\}\big(d\log^2(n\eta/\delta)/n\big)^{1/2}$$

after $T \ge C_2\max\{1,\eta^{-2}\}\log(n\eta/d)$ iterations.

To prove the theorem, we consider two cases when $\eta \ge 1$ and $\eta \le 1$.

Case (i) $1 \le \eta = O(1)$**:** Given the initialization conditions in Theorem 7, we can get the following corollary of Theorem 5.

Corollary 3. *When $\eta \ge 1$ and $\sin\theta < 0.1$, we have*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| < 0.9\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|.$$

Furthermore, from the uniform concentration Lemma 3 for all $\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le O(\|\boldsymbol{\beta}^*\|)$, we have

$$\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \le C\|\boldsymbol{\beta}^*\|\sqrt{d\log^2(n/\delta)/n},$$

with probability $1 - \delta$ for some universal constant $c > 0$. From here, with $\eta = O(1)$, we can check that

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \lesssim 0.9^t\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + O\left(\sigma\sqrt{d\log^2(n/\delta)/n}\right).$$

Case (ii) $C(d\log^2(n/\delta)/n)^{1/4} \le \eta \le 1$**:** In this case, the result of Theorem 5 shows that:

Corollary 4. *When $\eta \le 1$ and $\sin\theta < 0.1$, we have*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le \left(1 - \frac{1}{8}\eta^2\right)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{46}$$

In order to analyze the convergence of finite-sample EM operator, we first divide the iterations into several epochs. Let $\bar{C}_0 = \|\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\|$. We consider that in each l^{th} epoch, $\boldsymbol{\beta}$ satisfies $\bar{C}_0 2^{-l-1} \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \bar{C}_0 2^{-l}$. Note that such consideration of dividing into several epochs is only conceptual, and does not affect the implementation of the EM algorithm.

Consider we are in l^{th} epoch such that $\bar{C}_0 2^{-l-1} \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \bar{C}_0 2^{-l}$. The key idea is that in each epoch, EM makes a progress toward the ground truth as long as the improvement in population operator overcomes the statistical error, *i.e.,*

$$\frac{1}{8}\eta^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \geq 2cr\sqrt{d\log^2(n/\delta)/n},$$

where c is a constant in Lemma 3. Here, since $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}^*\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$, we can set $r = \|\boldsymbol{\beta}^*\| + \bar{C}_0 2^{-l}$. This in turn implies that in l^{th} epoch, if the following is true:

$$\frac{1}{8}\eta^2 \bar{C}_0 2^{-l-1} \geq 2cr\sqrt{d\log^2(n/\delta)/n} \geq 4c(\|\boldsymbol{\beta}^*\| + \bar{C}_0 2^{-l})\sqrt{d\log^2(n/\delta)/n},$$

then we have

$$\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^*\| \leq \left(1 - \frac{1}{16}\eta^2\right)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|,$$

due to the concentration of finite-sample EM operators. Arranging the terms, we require that

$$\bar{C}_0 2^{-l}\left(\eta^2 - c_1\sqrt{d\log^2(n/\delta)/n}\right) \geq c_2\|\boldsymbol{\beta}^*\|\sqrt{d\log^2(n/\delta)/n},$$

for some universal constants $c_1, c_2 > 0$. Recall that we are in middle SNR regime where (with appropriately set constants)

$$\eta^2 \geq (c_1 + 1)\sqrt{d\log^2(n/\delta)/n}.$$

Therefore, $\boldsymbol{\beta}$ is guaranteed to move closer to $\boldsymbol{\beta}^*$ as long as

$$\bar{C}_0 2^{-l} \leq c_2\|\boldsymbol{\beta}^*\|\eta^{-2} \cdot \sqrt{d\log^2(n/\delta)/n} \leq c_2\eta^{-1} \cdot \sigma\sqrt{d\log^2(n/\delta)/n}.$$

Note that each epoch takes $O(\eta^{-2})$ iterations to enter the next epoch. We can conclude that after $l = O(\log(n/d))$ epochs, we enter the region where $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq c_2\eta^{-1}\sigma\sqrt{d\log^2(n/\delta)/n}$ for some absolute constant $c_2 > 0$.

For δ probability bound, we can replace δ with $\delta/\log(n/d)$ and take a union bound of the uniform deviation of finite-sample EM operators given in Lemma 3 for all epochs. This does not change the complexity in the final statistical error.

Finally, the required number of iterations in each epoch is $O(\eta^{-2})$ to make $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$ a half. Since the total number of epoch we require is $O(\log(n/d))$, the total number of iterations is at most $O(\eta^{-2}\log(n/d))$, concluding the proof in middle-high SNR regime.

Remark 2. *When $\|\boldsymbol{\beta}^*\| \gg \sigma$ so that $\eta = \omega(1)$, we have to show that the final statistical error is only proportional to σ. For this case, we are not aware of how to give a good uniform concentration bound on the finite-sample based EM operator. Furthermore, the analysis have to take a completely different path using a event-wise concentration (e.g., [19]) to tighten the statistical fluctuation of EM operators. See our conference version [21] for more details on how we handle this case in the high SNR regime.*

8.6 Proof of Theorem 8

We restate the theorem below for readers' convenience.

Theorem 8 (ℓ_2 Convergence, Finite-Sample in Low SNR Regime)**.** *Suppose $\eta \leq C(d\log^2(n/\delta)/n)^{1/4}$ and $\|\widetilde{\boldsymbol{\beta}}_0\| = O(\sigma)$. Then there exist universal constants $C_1, C_2 > 0$ such that with probability at least $1 - \delta$, we have*

$$\|\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq C_1\sigma(d\log^2(n/\delta)/n)^{1/4}$$

after $T \geq C_2\log(\log(n/d))\sqrt{n/(d\log^2(n/\delta))}$ iterations of finite-sample EM.

We divide the phases into two when $\|\widetilde{\boldsymbol{\beta}}_0\|$ is greater than 0.2σ, and then study when we start from the norm smaller than 0.2σ. Note that we start from $\|\widetilde{\boldsymbol{\beta}}_0\| = O(\sigma)$.

8.6.1 Proof for the Case $\|\widetilde{\boldsymbol{\beta}}_0\| \geq 0.2\sigma$

First, suppose $\|\boldsymbol{\beta}\| \geq 2/3\sigma$. Then,

$$\begin{aligned}\|\boldsymbol{\beta}'\| &\leq \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}\left[(\boldsymbol{X}^\top\boldsymbol{\beta}^*)(\boldsymbol{X}^\top u)\tanh\left(\frac{YX^\top\boldsymbol{\beta}}{\sigma^2}\right)\right] + \mathbb{E}\left[z(\boldsymbol{X}^\top u)\tanh\left(\frac{YX^\top\boldsymbol{\beta}}{\sigma^2}\right)\right]\\ &\leq \sup_{u\in\mathbb{S}^{d-1}} \sqrt{\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2]\mathbb{E}[(X^\top u)^2]} + \mathbb{E}[|z(\boldsymbol{X}^\top u)|],\\ &\leq \|\boldsymbol{\beta}^*\| + \mathbb{E}[|z(\boldsymbol{X}^\top u)|] \leq \|\boldsymbol{\beta}^*\| + 2\sigma/\pi.\end{aligned}$$

where $z\sim\mathcal{N}(0,\sigma^2)$ such that $Y = \boldsymbol{X}^\top\boldsymbol{\beta}^* + z$. Since the uniform deviation of finite-sample EM is given by Lemma 3 as $2\|\boldsymbol{\beta}\|\sqrt{d\log^2(n/\delta)/n}$, we can conclude that

$$\begin{aligned}\|\widetilde{\boldsymbol{\beta}}'\| &\leq \|\boldsymbol{\beta}'\| + O\left(\sigma\sqrt{d\log^2(n/\delta)/n}\right)\\ &\leq \|\boldsymbol{\beta}^*\| + 2/\pi + O\left(\sigma\sqrt{d\log^2(n/\delta)/n}\right) \leq 2/3\sigma.\end{aligned}$$

Next, suppose $0.2\sigma \leq \|\boldsymbol{\beta}\| \leq 2/3\sigma$. Let $\boldsymbol{v}_1 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$, and $\boldsymbol{v}_2$ is orthogonal to $\boldsymbol{v}_1$ such that $\mathrm{span}(\boldsymbol{v}_1,\boldsymbol{v}_2) = \mathrm{span}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$. We can start from

$$\boldsymbol{\beta}' = \mathbb{E}\left[Y\alpha_1\tanh\left(\frac{Y\alpha_1\|\boldsymbol{\beta}\|}{\sigma^2}\right)\right]\boldsymbol{v}_1 + \mathbb{E}\left[Y\alpha_2\tanh\left(\frac{Y\alpha_1\|\boldsymbol{\beta}\|}{\sigma^2}\right)\right]\boldsymbol{v}_2,$$

where $\alpha_1 = \boldsymbol{X}^\top\boldsymbol{v}_1$ and $\alpha_2 = \boldsymbol{X}^\top\boldsymbol{v}_2$. We will see in Appendix A.1 that $\langle\boldsymbol{\beta}',\boldsymbol{v}_2\rangle \leq \frac{1}{2}\|\boldsymbol{\beta}\|\eta^2 \leq c_0\sigma\sqrt{d\log^2(n/\delta)/n}$ for some absolute constant $c_0 > 0$. Therefore, we focus on bounding the first term.

Let $a = 4$, and define event $\mathcal{E} := \{\alpha_1^2 + (z/\sigma)^2 \leq a\}$. We expand $\boldsymbol{\beta}'$ as follows:

$$\begin{aligned}\boldsymbol{\beta}'^\top\boldsymbol{v}_1 &\leq \frac{\|\boldsymbol{\beta}\|}{\sigma^2}\mathbb{E}[y^2\alpha_1^2 1_{\mathcal{E}}] + \mathbb{E}[|Y\alpha_1|1_{\mathcal{E}^c}]\\ &\leq \frac{\|\boldsymbol{\beta}\|}{\sigma^2}\mathbb{E}[z^2\alpha_1^2 1_{\mathcal{E}}] + \mathbb{E}[|z\alpha_1|1_{\mathcal{E}^c}] + O(\|\boldsymbol{\beta}^*\|).\end{aligned}$$

By converting the above expression to Rayleigh distribution with $\alpha_1 = r\cos w, (z/\sigma) = r\sin w$, we can more explicitly find the values of the expectations in the above equation. That is,

$$\mathbb{E}[(z/\sigma)^2\alpha_1^2 1_{\mathcal{E}}] = \frac{1}{2\pi}\int_0^{2\pi}\cos^2 w\sin^2 w dw\int_0^4 r^5\exp(-r^2/2)dr \approx 1 - 0.013,$$

and

$$\mathbb{E}[|(z/\sigma)\alpha_1|1_{\mathcal{E}^c}] = \frac{1}{2\pi}\int_0^{2\pi}|\cos w\sin w|dw\int_4^\infty r^3\exp(-r^2/2)dr \leq 0.002,$$

Therefore, using $\|\boldsymbol{\beta}\| > 0.2\sigma$,

$$\langle\boldsymbol{\beta}',\boldsymbol{v}_1\rangle \leq \|\boldsymbol{\beta}\|(1-0.013) + O(\sigma + \|\boldsymbol{\beta}^*\|) \leq \gamma\|\boldsymbol{\beta}\|,$$

where $\gamma = 0.997 < 1$. Since the deviation of finite-sample EM operator is in order $\sigma\sqrt{d\log^2(n/\delta)/n}$, we can conclude that

$$\|\boldsymbol{\beta}'\| \leq \gamma\|\boldsymbol{\beta}\| + O\left(\sigma\sqrt{d\log^2(n/\delta)/n}\right).$$

Hence we can conclude that after $T = O(1)$ iterations, $\|\widetilde{\boldsymbol{\beta}}_T\| \leq 0.2\sigma$.

8.6.2 Convergence after $\|\boldsymbol{\beta}_0\| \leq 0.2\sigma$

As mentioned in the main text, the core idea of the low SNR regime is that EM essentially cannot distinguish the cases between $\boldsymbol{\beta}^* = 0$ and $\boldsymbol{\beta}^* \neq 0$. Therefore, instead of studying the contraction of population EM operator to $\boldsymbol{\beta}^*$, we study its contraction to 0.

From Lemma 3, we immediately have that

$$\sup_{\|\boldsymbol{\beta}\|\leq r}\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \leq cr\sqrt{d\log^2(n/\delta)/n},$$

for some universal constant $c > 0$. Given the contraction of population EM operator and the deviation bound between the sample and population EM operators, we are ready to study the convergence behaviors of EM algorithm under the low SNR regime. Our proof argument follows the localization argument used in Case (ii) of middle SNR regime. In particular, let the target error be $\epsilon_n := c\sqrt{d\log^2(n/\delta)/n}$ with some absolute constant $c > 0$. We assume that we start from the initialization region where $\|\boldsymbol{\beta}\|/\sigma \leq \epsilon_n^{\alpha_0}$ for some $\alpha_0 \in [0, 1/2)$.

The localization argument proceeds as the following: suppose that $\epsilon_n^{\alpha_{l+1}} \leq \|\boldsymbol{\beta}\|/\sigma \leq \epsilon_n^{\alpha_l}$ at the l^{th} epoch for $l \geq 0$. We let $c > 0$ sufficiently large such that

$$\epsilon_n \geq 4c_u\eta^2 + 4\sup_{\boldsymbol{\beta}\in\mathbb{B}(\boldsymbol{\beta}^*, r_l)} \|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\|/r_l,$$

with $r_l = \epsilon_n^{\alpha_l}$. During this period, from Lemma 5 on contraction of population EM, and Lemma 3 concentration of finite sample EM, we can check that

$$\begin{aligned}\|\widetilde{\boldsymbol{\beta}}'\| &\leq \|\boldsymbol{\beta}\| - 0.5\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}\|/\sigma)^2 + c_u\|\boldsymbol{\beta}\|\eta^2 + \sup_{\boldsymbol{\beta}\in\mathbb{B}(\boldsymbol{\beta}^*, r)} \|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \\ &\leq \|\boldsymbol{\beta}\| - \frac{\sigma}{2}\epsilon_n^{3\alpha_{l+1}} + \frac{\sigma}{4}\epsilon_n^{\alpha_l+1}.\end{aligned}$$

Note that this inequality is valid as long as $\epsilon_n^{\alpha_{l+1}} \leq \|\boldsymbol{\beta}\| \leq \epsilon_n^{\alpha_l}$. Now we define a sequence α_l using the following recursion:

$$\alpha_{l+1} = \frac{1}{3}(\alpha_l + 1). \tag{47}$$

The limit point of this recursion is $1/2$, which will give $\epsilon_n^{\alpha_\infty} \approx (d/n)^{1/4}$ as argued in the main text. Hence during the l^{th} epoch, we have

$$\|\widetilde{\boldsymbol{\beta}}'\| \leq \|\boldsymbol{\beta}\| - \frac{\sigma}{4}\epsilon_n^{\alpha_l+1}.$$

Furthermore, the number of iterations required in l^{th} epoch is

$$t_l := (\epsilon_n^{\alpha_l} - \epsilon_n^{\alpha_{l+1}})/\epsilon_n^{\alpha_l+1} \leq \epsilon_n^{-1}.$$

After getting out of l^{th} epoch, it gets into $(l+1)^{th}$ epoch which can be analyzed in the same way. From this, we can conclude that after going through l epochs in total, we have $\|\boldsymbol{\beta}\| \leq \epsilon_n^{\alpha_{l+1}}$. Note that the number of EM iterations taken up to this point is $l\epsilon_n^{-1}$.

It is easy to check $\alpha_l = (1/3)^l(\alpha_0 - 1/2) + 1/2$ from (47). We can set $l = C\log(1/\beta)$ for some universal constant C such that α_l is $1/2 - \beta$ for arbitrarily small $\beta > 0$. In conclusion,

$$\|\widetilde{\boldsymbol{\beta}}_t\|/\sigma \leq \epsilon_n^{1/2-\beta} \leq c \cdot (d\ln^2(n/\delta)/n)^{1/4-\beta/2},$$

with high probability as long as $t \geq \epsilon_n^{-1} l \gtrsim \sqrt{d/n}\log(1/\beta)$ where c is some universal constant. Hence we can set $\beta = C/\log(d/n)$ to get a desired result $\|\widetilde{\boldsymbol{\beta}}_t\| \leq c\sigma \cdot (d\ln^2(n/\delta)/n)^{1/4}$. Since $\|\boldsymbol{\beta}^*\| \leq C_0\sigma(d\ln^2(n/\delta)/n)^{1/4}$, it implies $\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\| \leq c_1\sigma(d\ln^2(n/\delta)/n)^{1/4}$ where c_1 is some universal constant.

Note that we need the union bound of the concentration of sample EM operators for all $l = 1, ..., C\log(1/\beta)$, such that the argument holds for all epochs. For this purpose, we can replace δ by $\delta/\log(1/\beta)$. This does not change the order of ϵ_n, hence the proof is complete.

9 Conclusion

In this paper, we studied the EM algorithm for a mixture of two linear regression models. In the large sample limit, we showed that EM converges to true parameters globally without any specialized initialization. In finite sample case, we showed that EM enjoys the same convergences behavior, with the optimal statistical rates in all SNR regimes of interest, matching the lower bounds provided in [8].

We believe that this work builds a ground for the analysis of the EM algorithm, as well as the landscape of MLE problems for a mixture of two Gaussian-style distributions. One potential direction is to analyze the EM algorithm in more general mixture models. This includes models with unequal mixing weights; in the case of mixture of two Gaussians, this has been done in [33], a follow-up of this paper. Considering mixture models with more general noise covariance and more than two mixture components, is also of interest and would require additional techniques due to the existence of sub-optimal local minima [14]. Another interesting regime is the high-dimensional case with sparse parameters [32, 39, 5], aiming to achieve the minimax rates in all SNR regimes, which may exhibit additional challenges due to statistical-computational trade-off [4]. We leave them as interesting future work.

Acknowledgement

J. Kwon and C. Caramanis are partially supported by NSF EECS-1609279, CCF-1302435, and CNS-1704778. W. Qian and Y. Chen are partially supported by NSF grants CCF-1657420, CCF-1704828 and CCF-2233152. Nhat Ho was partially supported by the NSF IFML 2019844 award and research gifts by UT Austin ML grant.

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, Feb. 2017. ISSN 0090-5364. doi: 10.1214/16-AOS1435. URL http://projecteuclid.org/euclid.aos/1487667618.

[2] B. Barazandeh and M. Razaviyayn. On the behavior of the expectation-maximization algorithm for mixture models. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 61–65. IEEE, 2018.

[3] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press, 2019.

[4] M. Brennan and G. Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

[5] T. T. Cai, J. Ma, and L. Zhang. CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

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

[7] J. Chen and P. Li. Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37(5A):2523–2542, 2009.

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

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

[10] C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Conference on Learning Theory*, pages 704–710, 2017.

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

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[13] R. Dwivedi, K. Khamaru, M. Wainwright, M. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *The annals of statistics*, 48(6), 2020.

[14] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

[15] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[16] M. I. Jordan and L. Xu. Convergence results for the em approach to mixtures of experts architectures. *Neural networks*, 8(9):1409–1431, 1995.

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

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

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

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

[21] J. Kwon, N. Ho, and C. Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

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

[23] P. Li, J. Chen, and P. Marriott. Non-finite fisher information and homogeneity: an em approach. *Biometrika*, 96(2):411–426, 2009.

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

[25] J. Ma, L. Xu, and M. I. Jordan. Asymptotic convergence rate of the em algorithm for gaussian mixtures. *Neural Computation*, 12(12):2881–2907, 2000.

[26] S. G. Nagarajan and I. Panageas. On the analysis of EM for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, pages 634–659. PMLR, 2020.

[27] W. Qian, Y. Zhang, and Y. Chen. Global convergence of least squares EM for demixing two log-concave densities. In *Neural Information Processing Systems (NeurIPS)*, 2019.

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

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

[30] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2017. URL http://www-personal.umich.edu/~romanv/papers/HDP-book/HDP-book.pdf.

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

[32] Z. Wang, Q. Gu, Y. Ning, and H. Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[33] N. Weinberger and G. Bresler. The em algorithm is adaptively-optimal for unbalanced symmetric gaussian mixtures. *Journal of Machine Learning Research*, 23(103):1–79, 2022.

[34] C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

[35] Y. Wu and H. H. Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4(3):143–220, 2022.

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

[37] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.

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

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

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

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

[42] K. Zhong, P. Jain, and I. S. Dhillon. Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems*, pages 2190–2198, 2016.

Appendices

A Technical Lemmas for Finite-Sample EM

In this Appendix, we give several deferred proofs for the main theorem. For the simplicity of the presentation, we assume here that $\sigma = 1$, but original results hold with proper scaling.

A.1 Proof of Lemma 5

Let $\alpha_1 = \boldsymbol{X}^\top \boldsymbol{v}_1$ and $\alpha_2 = \boldsymbol{X}^\top \boldsymbol{v}_2$, where $\boldsymbol{v}_1 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$ and $\text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \text{span}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$.

Upper Bound: We first bound the first coordinate of the population operator from equation (3):

$$\langle \boldsymbol{\beta}', v_1 \rangle = \mathbb{E}_{\alpha_1,\alpha_2,Y} \left[\tanh(Y\alpha_1\|\boldsymbol{\beta}\|)\alpha_1 Y \right],$$

We will expand the above equation using Taylor series bound of $x\tanh(x)$:

$$x^2 - \frac{x^4}{3} \leq x\tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15}. \tag{48}$$

Now we unfold the equation above, we have

$$\begin{aligned}
\langle \boldsymbol{\beta}', v_1 \rangle &= \frac{1}{\|\boldsymbol{\beta}\|}\mathbb{E}_{\alpha_1,\alpha_2,Y} \left[\tanh(Y\alpha_1\|\boldsymbol{\beta}\|)Y\alpha_1\|\boldsymbol{\beta}\| \right] \\
&\leq \frac{1}{\|\boldsymbol{\beta}\|}\mathbb{E}_{\alpha_1,\alpha_2,Y} \left[(Y\alpha_1\|\boldsymbol{\beta}\|)^2 - \frac{(Y\alpha_1\|\boldsymbol{\beta}\|)^4}{3} + \frac{2(Y\alpha_1\|\boldsymbol{\beta}\|)^6}{15} \right] \\
&\leq \frac{1}{\|\boldsymbol{\beta}\|}\mathbb{E}_{\alpha_1,z} \Bigg[(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^* + \alpha_2 b_2^*))^2 - \frac{(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^* + \alpha_2 b_2^*))^4}{3} \\
&\qquad + \frac{2(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^* + \alpha_2 b_2^*))^6}{15} \Bigg],
\end{aligned}$$

where $z \sim \mathcal{N}(0,1)$ and we used $Y = \alpha_1 b_1^* + \alpha_2 b_2^* + z$ with $b_1^* = \langle \boldsymbol{\beta}^*, \boldsymbol{v}_1 \rangle$ and $b_2^* = \langle \boldsymbol{\beta}^*, \boldsymbol{v}_2 \rangle$. Note here that, any (constantly) higher order terms of Gaussian distribution is constant. Hence instead of computing all coefficients explicitly for all monomials, we can simplify the argument as

$$\begin{aligned}
\langle \boldsymbol{\beta}', v_1 \rangle &\leq \frac{1}{\|\boldsymbol{\beta}\|}\mathbb{E}_{\alpha_1,z} \left[(\alpha_1\|\boldsymbol{\beta}\|z)^2 - \frac{(\alpha_1\|\boldsymbol{\beta}\|z)^4}{3} + \frac{2(\alpha_1\|\boldsymbol{\beta}\|z)^6}{15} \right] + c_1\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2, \\
&= \|\boldsymbol{\beta}\|(1 - 3\|\boldsymbol{\beta}\|^2 + 30\|\boldsymbol{\beta}\|^4) + c_1\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2,
\end{aligned} \tag{49}$$

for some universal constant $c_1 > 0$. Since we assumed $\|\boldsymbol{\beta}\| < 0.2$, we have $3\|\boldsymbol{\beta}\|^2 - 30\|\boldsymbol{\beta}\|^4 \geq \|\boldsymbol{\beta}\|^2$. We conclude that

$$\langle \boldsymbol{\beta}', \boldsymbol{v}_1 \rangle \leq \|\boldsymbol{\beta}\|(1 - \|\boldsymbol{\beta}\|^2 + c_1\|\boldsymbol{\beta}^*\|^2).$$

Then we bound the value in the second coordinate of the population operator:

$$\langle \boldsymbol{\beta}', \boldsymbol{v}_2 \rangle = \mathbb{E}_{\alpha_1,\alpha_2,Y} \left[\tanh(Y\alpha_1\|\boldsymbol{\beta}\|)Y\alpha_2 \right],$$

where $Y|(\alpha_1, \alpha_2) \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, 1)$. In order to derive an upper bound for the above equation, we rely on the following equation which we defer the proof to the end of this section:

$$\mathbb{E}\left[\tanh(Y\alpha_1\|\boldsymbol{\beta}\|)Y\alpha_2 \right] = b_2^* \, \mathbb{E}\left[\alpha_1^2 \tanh(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^*)) - \|\boldsymbol{\beta}\| b_1^* \alpha_1^2 \tanh'(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^*)) \right], \tag{50}$$

where $z \sim \mathcal{N}(0, 1 + b_2^{*2})$ with subsuming α_2 from the equation. From (50), we can check that

$$\begin{aligned}
\mathbb{E}\left[\tanh(Y\alpha_1\|\boldsymbol{\beta}\|)Y\alpha_2\right] &\leq b_2^* \, \mathbb{E}\left[\alpha_1^2 \tanh(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^*))\right] \\
&= \frac{b_2^*}{2}\mathbb{E}\left[\alpha_1^2 \tanh(\alpha_1\|\boldsymbol{\beta}\|(z + \alpha_1 b_1^*)) + \alpha_1^2 \tanh(\alpha_1\|\boldsymbol{\beta}\|(-z + \alpha_1 b_1^*))\right] \\
&\leq b_2^* \mathbb{E}\left[\alpha_1^2 \tanh(\alpha_1^2\|\boldsymbol{\beta}\| b_1^*)\right], \\
&\leq \|\boldsymbol{\beta}\| b_1^* b_2^* \mathbb{E}\left[\alpha_1^4\right] \leq \frac{1}{2}\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2,
\end{aligned}$$

where we used $\tanh(a + x) + \tanh(a - x) \leq 2\tanh(a)$ for any $a > 0$ and $x \in \mathbb{R}$.

From the above results, we have shown that

$$\|\boldsymbol{\beta}'\| \leq |\langle \boldsymbol{\beta}', \boldsymbol{v}_1\rangle| + |\langle \boldsymbol{\beta}', \boldsymbol{v}_2\rangle| \leq \|\boldsymbol{\beta}\|\left(1 - \|\boldsymbol{\beta}\|^2 + c\|\boldsymbol{\beta}^*\|^2\right), \tag{51}$$

for some universal constant $c > 0$.

Lower Bound: To prove the lower bound of the population EM operator, we again expand the equation using Taylor series (48):

$$\|\boldsymbol{\beta}'\| \geq |\langle \boldsymbol{\beta}', \boldsymbol{v}_1\rangle| \geq \|\boldsymbol{\beta}\|(1 - 3\|\boldsymbol{\beta}\|^2) - c_2\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2. \tag{52}$$

The result follows immediately with some absolute constant $c_2 > 0$.

Proof of equation (50): For the left hand side, we apply the Stein's lemma with respect to α_2. It gives that

$$\begin{aligned}
\mathbb{E}[\tanh(\|\boldsymbol{\beta}\|\alpha_1 Y)Y\alpha_2] &= \mathbb{E}\left[\frac{d}{d\alpha_2}\tanh(\|\boldsymbol{\beta}\|\alpha_1 Y)Y\right] \\
&= \mathbb{E}\left[\frac{d}{d\alpha_2}\tanh(\|\boldsymbol{\beta}\|\alpha_1(\bar{z} + \alpha_1 b_1^* + \alpha_2 b_2^*))(\bar{z} + \alpha_1 b_1^* + \alpha_2 b_2^*)\right] \\
&= \mathbb{E}[b_2^* \tanh(\|\boldsymbol{\beta}\|\alpha_1(\bar{z} + \alpha_1 b_1^* + \alpha_2 b_2^*)) \\
&\quad + (\|\boldsymbol{\beta}\|\alpha_1 b_2^*)(\bar{z} + \alpha_1 b_1^* + \alpha_2 b_2^*)\tanh'(\|\boldsymbol{\beta}\|\alpha_1(\bar{z} + \alpha_1 b_1^* + \alpha_2 b_2^*)] \\
&= b_2^* \, \mathbb{E}[\tanh(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)) + \|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)\tanh'(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)))]
\end{aligned}$$

where $\bar{z} \sim \mathcal{N}(0, 1)$ and $z \sim \mathcal{N}(0, 1 + b_2^{*2})$. For the right hand side, we apply the Stein's lemma with respect to α_1. First, we check the first term in the right hand side that

$$\begin{aligned}
&\mathbb{E}[\alpha_1^2 \tanh(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*))] \\
&\quad = \mathbb{E}\left[\frac{d}{d\alpha_1}(\alpha_1 \tanh(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)))\right] \\
&\quad = \mathbb{E}\left[\tanh(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)) + \alpha_1 \frac{d}{d\alpha_1}\tanh(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)\right] \\
&\quad = \mathbb{E}\left[\tanh(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)) + \|\boldsymbol{\beta}\|\alpha_1(z + 2\alpha_1 b_1^*)\tanh'(\|\boldsymbol{\beta}\|\alpha_1(z + \alpha_1 b_1^*)\right].
\end{aligned}$$

Plugging this into (50) and subtracting the remaining term gives the result that matches to the left hand side.

A.2 Proof of Lemma 3

For this result, we need the following lemma:

Lemma 14. *Suppose $\|\boldsymbol{\beta}^*\| \leq \rho$ for some universal constant $\rho > 0$. Then for any given $r > 0$, with probability at least $1 - \delta$, we have*

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|\leq r}\left\|\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \mathbb{E}\left[Y\boldsymbol{X}\tanh(Y\boldsymbol{X}^\top\boldsymbol{\beta})\right]\right\| \leq cr\sqrt{\frac{d\ln^2(n/\delta)}{n}}, \tag{53}$$

for some universal constant $c > 0$.

Proof of Lemma 3. Let us assume that $n \geq Cd$ for sufficiently large constant $C > 0$. To simplify the notation, we use $\hat{\Sigma}_n = \frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$. Observe that

$$\|\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \leq \|\hat{\Sigma}_n^{-1}\|_{\mathrm{op}} \|\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i \tanh(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}) - \boldsymbol{\beta}'\| + \|\hat{\Sigma}_n^{-1} - I\|_{\mathrm{op}} \|\boldsymbol{\beta}'\|.$$

The first term can be bounded by $c_1 r\sqrt{d\log^2(n/\delta)/n}$ with some absolute constant $c_1 > 0$ using the results of Lemma 14.

For the second term, since $\boldsymbol{X} \sim \mathcal{N}(0, I)$, from a standard concentration of measure we directly get

$$\|\hat{\Sigma}_n^{-1} - I\|_{\mathrm{op}} = \|\hat{\Sigma}_n^{-1}\|_{\mathrm{op}} \|\hat{\Sigma}_n - I\|_{\mathrm{op}} \leq c_2\sqrt{d\ln(1/\delta)/n},$$

for some universal constant $c_2 > 0$. If we can show that $\|\boldsymbol{\beta}'\| \leq O(r)$, then we are done. To see this, first we check that

$$\|\boldsymbol{\beta}'\| = \|\mathbb{E}[Y\boldsymbol{X}\tanh(Y\boldsymbol{X}^\top\boldsymbol{\beta})]\| \leq \|\boldsymbol{\beta}\| \|\mathbb{E}[Y^2\boldsymbol{X}\boldsymbol{X}^\top]\|_{\mathrm{op}}.$$

It is easy to check that $\mathbb{E}[Y^2\boldsymbol{X}\boldsymbol{X}^\top] = I + 2\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}$, hence $\|\mathbb{E}[Y^2 X X^\top]\|_{\mathrm{op}} = 1 + 2\|\boldsymbol{\beta}^*\|^2 \leq 1 + 2C^2 = O(1)$. Therefore, $\|\boldsymbol{\beta}'\| \leq c_3\|\boldsymbol{\beta}\| \leq c_3 r$ with a constant $c_3 = (1 + 2C^2)$. This completes the proof of Lemma 3. □

A.3 Proof of Lemma 14

Proof. We start with the standard discretization argument for bounding the concentration of measures in l_2 norm. Let $Z(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \boldsymbol{\beta}'$. The standard symmetrization argument gives that [31].

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \|Z(\boldsymbol{\beta})\| \geq t\right) \leq 2\mathbb{P}\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i y_i \boldsymbol{x}_i \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\| \geq t/2\right), \tag{54}$$

where ε_i are independent Rademacher random variables. We define a good event $\mathcal{E} := \{\forall i \in [n], |y_i| \leq \tau, |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \leq C\tau\}$ as before, where $\tau = \Theta\left(\sqrt{\log(n/\delta)}\right)$. Then the probability defined in (54) can be decomposed as

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i y_i \boldsymbol{x}_i \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\| \geq t/2 \middle| \mathcal{E}\right) + P(\mathcal{E}^c).$$

We are interested in bounding the following quantity for Chernoff bound:

$$\mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \frac{\lambda}{n}\left\|\sum_{i=1}^n \varepsilon_i y_i \boldsymbol{x}_i \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\|\right) \middle| \mathcal{E}\right],$$

where we used Chernoff-Bound with some $\lambda > 0$ for the last inequality. We first go some steps before we can apply the Ledoux-Talagrand contraction arguments [22], with $f_i(\boldsymbol{\beta}) := \tanh\left(|y_i| \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$. First, we use discretization argument for removing l_2 norm inside the expectation.

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \frac{\lambda}{n}\left\|\sum_{i=1}^n \varepsilon_i y_i \boldsymbol{x}_i \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\|\right) \middle| \mathcal{E}\right] \\
&\leq \mathbb{E}\left[\exp\left(\sup_{u\in\mathbb{S}^d}\sup_{\|\boldsymbol{\beta}\|\leq r} \frac{\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i (\boldsymbol{x}_i^\top u) \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right) \middle| \mathcal{E}\right] \\
&\leq \mathbb{E}\left[\exp\left(\sup_{j\in[M]}\sup_{\|\boldsymbol{\beta}\|\leq r} \frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i (\boldsymbol{x}_i^\top u_j) \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right) \middle| \mathcal{E}\right] \\
&\leq \sum_{j=1}^M \mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i (\boldsymbol{x}_i^\top u_j) \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right) \middle| \mathcal{E}\right],
\end{aligned}$$

where M is $1/2$-covering number of the unit sphere $\mathbb{S}$ and $\{u_1, ..., u_M\}$ is the corresponding covering set. Now for each u_j, we can apply the Ledoux-Talagrand contraction lemma since $|f_i(\boldsymbol{\beta}_1) - f_i(\boldsymbol{\beta}_2)| \leq |y_i||\boldsymbol{x}_i^\top \boldsymbol{\beta}_1 - \boldsymbol{x}_i^\top \boldsymbol{\beta}_2|$ for $\boldsymbol{\beta} \in \mathbb{B}(0, r)$:

$$\mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\leq r} \frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i \boldsymbol{x}_i^\top u_j \tanh\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right) \middle| \mathcal{E}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\le r}\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i|y_i|\boldsymbol{x}_i^\top u_j\tanh\left(|y_i|\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right)\middle|\mathcal{E}\right]\\
&\le \mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\le r}\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i y_i^2(\boldsymbol{x}_i^\top\boldsymbol{\beta})(\boldsymbol{x}_i^\top u_j)\right)\middle|\mathcal{E}\right]\\
&\le \mathbb{E}\left[\exp\left(\sup_{v\in\mathbb{S}^{d-1}}\frac{2r\lambda}{n}\sum_{i=1}^{n}\varepsilon_i y_i^2(\boldsymbol{x}_i^\top v)(\boldsymbol{x}_i^\top u_j)\right)\middle|\mathcal{E}\right], \qquad (55)
\end{aligned}
$$

where we define $v := \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$. Again, we can apply the $1/2$-covering number argument to bound this by

$$
\sum_{k=1}^{M}\mathbb{E}\left[\exp\left(\frac{4r\lambda}{n}\sum_{i=1}^{n}\varepsilon_i y_i^2(\boldsymbol{x}_i^\top u_k)(\boldsymbol{x}_i^\top u_j)\right)\middle|\mathcal{E}\right].
$$

Note that $y_i(\boldsymbol{x}_i^\top u_j)|\mathcal{E}$ is sub-Gaussian with Orcliz norm $O(\tau(1+\|\boldsymbol{\beta}^*\|)) = O(\tau)$. Since the multiplication of two sub-Gaussian variables is sub-exponential, it implies that $y_i^2(\boldsymbol{x}_i^\top u_k)(\boldsymbol{x}_i^\top u_j)|\mathcal{E}$ is sub-exponential with Orcliz norm $O(\tau^2)$ [30]. Now we need the lemma for the exponential moment of sub-exponential random variables from [30].

Lemma 15 (Lemma 5.15 in [30]). *Let X be a centered sub-exponential random variable. Then, for t such that $t \le c/\|X\|_{\psi_1}$, one has*

$$
\mathbb{E}[\exp(tX)] \le \exp(Ct^2\|X\|_{\psi_1}^2),
$$

for some universal constant $c, C > 0$.

Finally, note that $\varepsilon_i y_i^2(\boldsymbol{x}_i^\top v)(\boldsymbol{x}_i^\top u_1)$ is a centered sub-exponential random variable with the same Orcliz norm. Equipped with the lemma, we can obtain that

$$
\mathbb{E}\left[\exp\left(4\lambda r\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i^2(\boldsymbol{x}_i^\top u_k)(\boldsymbol{x}_i^\top u_1)\right)\middle|\mathcal{E}\right] \le \exp(C\lambda^2 r^2\tau^4/n), \qquad \forall|\lambda r/n| \le c/\tau^2,
$$

which yields

$$
\mathbb{E}\left[\exp\left(\sup_{\|\boldsymbol{\beta}\|\le r}\frac{\lambda}{n}\left\|\sum_{i=1}^{n}\varepsilon_i y_i\boldsymbol{x}_i\tanh\left(y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right\|\right)\middle|\mathcal{E}\right] \le \exp\left(C\lambda^2 r^2\tau^4/n + C'd\right),\ \forall|\lambda| \le n/c\tau^2 r,
$$

where we used $\log M = O(d)$ with some $C, C', c > 0$. Combining all the above, we have that

$$
\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}(\boldsymbol{\beta}^*,r)}\|Z(\boldsymbol{\beta})\| \ge t\right) \le \exp\left(C_0\lambda^2 r^2\tau^4/n + C_1 d - \lambda t/2\right) + \mathbb{P}(\mathcal{E}^c).
$$

From here, we can optimize for $\lambda = O(t/r^2\tau^4)$ with setting $t = O\left(r\sqrt{d\tau^4/n}\right)$. Since $t = O\left(r\sqrt{d\log^2(n/\delta)/n}\right)$, this concludes the proof. □

A.4 Lower Bound on the Norm: Proof of Lemma 13

Let $\alpha = \angle(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$. We recall here that $b_1^* = \boldsymbol{\beta}^*\cos\alpha$, $b_2^* = \boldsymbol{\beta}^*\sin\alpha$.

Case (i): $\cos\alpha \le 0.2$. This case we essentially give a norm bound for $\cos\alpha = 0$. Suppose that $\|\boldsymbol{\beta}\| \le \|\boldsymbol{\beta}^*\|/10$. We can first check that

$$
\begin{aligned}
\|\boldsymbol{\beta}'\| &\ge |\langle\boldsymbol{\beta}', v_1\rangle| = \mathbb{E}_{\alpha_1,\alpha_2,Y}[\tanh(Y\alpha_1\|\boldsymbol{\beta}\|)Y\alpha_1]\\
&= \mathbb{E}_{\alpha_1,\alpha_2,z}[\tanh((\alpha_1 b_1^* + \alpha_2 b_2^* + z)\alpha_1\|\boldsymbol{\beta}\|)(\alpha_1 b_1^* + \alpha_2 b_2^* + z)\alpha_1],
\end{aligned}
$$

where $\alpha_1, \alpha_2, z \sim \mathcal{N}(0,1)$. The above quantity is larger than the following $b_1^* = 0$ case:

$$
\mathbb{E}_{\alpha_1,\alpha_2,z}[\tanh((\alpha_2 b_2^* + z)\alpha_1\|\boldsymbol{\beta}\|)(\alpha_2 b_2^* + z)\alpha_1] = \mathbb{E}_{\alpha_1,\bar{z}}[\tanh(\bar{z}\alpha_1\|\boldsymbol{\beta}\|)\bar{z}\alpha_1],
$$

where $\bar{z} \sim \mathcal{N}(0, 1 + (b_2^*)^2) = \mathcal{N}(0, \sigma_2^2)$. We can lower bound the following quantity such that

$$
\mathbb{E}_{\alpha_1,\bar{z}}[\tanh(\bar{z}\alpha_1\|\boldsymbol{\beta}\|)\bar{z}\alpha_1] \ge \sigma_2\mathbb{E}_{\alpha_1,z}[\tanh(\sigma_2 z\alpha_1\|\boldsymbol{\beta}\|)z\alpha_1]
$$

$$\geq \sigma_2 \mathbb{E}_{\alpha_1,z}[\tanh(z\alpha_1\|\boldsymbol{\beta}\|)z\alpha_1].$$

If $\|\boldsymbol{\beta}\| > 0.5$, then through the numerical integration we can check that $\mathbb{E}_{\alpha_1,z}[\tanh(0.5z\alpha_1)z\alpha_1] > 1/\pi$. Hence, we immediately have that

$$|\langle \boldsymbol{\beta}', v_1\rangle| \geq \frac{1}{\pi}\sigma_2 \geq \frac{\sin\alpha}{\pi}\|\boldsymbol{\beta}^*\| \geq \frac{1}{5}\|\boldsymbol{\beta}^*\|,$$

since $\sin\alpha > 0.9$ in this case. Since we are considering the case when $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}^*\|/10$, clearly we have

$$\|\boldsymbol{\beta}'\| \geq \|\boldsymbol{\beta}\|(1 + 1 \cdot \min(1, \|\boldsymbol{\beta}\|^2)).$$

If $\|\boldsymbol{\beta}\| < 0.5$, then we get a lower bound using Taylor expansion:

$$\begin{aligned}\mathbb{E}_{\alpha_1,\bar{z}}[\tanh(\bar{z}\alpha_1\|\boldsymbol{\beta}\|)\bar{z}\alpha_1] &\geq \sigma_2\left(\mathbb{E}_{\alpha_1,z}[\|\boldsymbol{\beta}\|(z\alpha_1)^2] - \frac{1}{3}\mathbb{E}_{\alpha_1,z}[\|\boldsymbol{\beta}\|^3(z\alpha_1)^4]\right)\\ &= \sigma_2\|\boldsymbol{\beta}\|(1 - 3\|\boldsymbol{\beta}\|^2) = \|\boldsymbol{\beta}\|\sqrt{1+0.96\eta^2}(1-3\|\boldsymbol{\beta}\|^2),\end{aligned}$$

where $\|\boldsymbol{\beta}^*\| = \eta$. Here, we consider three cases when $\eta \geq 5$, $5 \geq \eta \geq 1$, $1 \geq \eta$. When $\eta \geq 5$, then we immediately have $|\langle \boldsymbol{\beta}', v_1\rangle| \geq 1.25\|\boldsymbol{\beta}\|$. In case $5 \geq \eta \geq 1$, we first note that since $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}^*\|/10$, we check the value of

$$\|\boldsymbol{\beta}\|\sqrt{1+0.96\eta^2}(1-0.03\eta^2).$$

We can again, numerically check that $\sqrt{1+0.96\eta^2}(1-0.03\eta^2) \leq 1.25$ for $1 \leq \eta \leq 5$. Finally, when $\eta \leq 1$, then a simple algebra shows that

$$\|\boldsymbol{\beta}\|\sqrt{1+0.96\eta^2}(1-0.03\eta^2) \geq \|\boldsymbol{\beta}\|(1+0.3\eta^2).$$

Combining all, we can conclude that when $\|\boldsymbol{\beta}\| \leq \frac{\|\boldsymbol{\beta}^*\|}{10}$

$$\|\boldsymbol{\beta}'\| \geq \|\boldsymbol{\beta}\|(1 + 0.25 \cdot \min(1, \|\boldsymbol{\beta}^*\|^2)) \geq \|\boldsymbol{\beta}\|(1 + 0.25 \cdot \min(1, \|\boldsymbol{\beta}\|^2)).$$

Now note that $\langle \boldsymbol{\beta}', v_1\rangle$ increases in $\|\boldsymbol{\beta}\|$, hence for all $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$, it holds that

$$\|\boldsymbol{\beta}'\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}(1 + 0.25 \cdot \min(1, \|\boldsymbol{\beta}^*\|^2)).$$

Case (ii): $\cos\alpha \geq 0.2$. Again, we can only consider when $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}^*\|/10$ since the other case will immediately follow. Their claim in this case is that $|\langle \boldsymbol{\beta}', v_1\rangle| \geq \min\left(\sigma_2^2\|\boldsymbol{\beta}\|, b_1^*\right)$. Hence we consider two cases when $\sigma_2^2\|\boldsymbol{\beta}\| = (1+\eta^2\sin^2\alpha)\|\boldsymbol{\beta}\| \leq b_1^* = \|\boldsymbol{\beta}^*\|\cos\alpha$ and the other case.

In the first case when $\sigma_2^2\|\boldsymbol{\beta}\| \leq b_1^*$, it can be shown that (see equation (39))

$$b_1^* - \langle \boldsymbol{\beta}', v_1\rangle \leq \kappa^3(b_1^* - \sigma_2^2\|\boldsymbol{\beta}\|),$$

where $\kappa \leq \sqrt{1+(b_1^*)^2}^{-1}$. Rearranging this inequality, we have

$$\begin{aligned}\langle \boldsymbol{\beta}', v_1\rangle &\geq \|\boldsymbol{\beta}^*\|(1-\kappa^3)\cos\alpha + \kappa^3(1+\eta^2\sin^2\alpha)\|\boldsymbol{\beta}\|\\ &\geq \|\boldsymbol{\beta}\|2(1-\kappa^3) + \kappa^3(1+\eta^2\sin^2\alpha)\|\boldsymbol{\beta}\|\\ &\geq \|\boldsymbol{\beta}\| + (1-\kappa^3)\|\boldsymbol{\beta}\|.\end{aligned}$$

Note that $1-\kappa^3 \geq c_1\min(1, b_1^2)$ for some constant $c_1 > 0$. On the other side, if $\sigma_2^2\|\boldsymbol{\beta}\| \geq b_1^*$, then we immediately have

$$\langle \boldsymbol{\beta}', v_1\rangle \geq \|\boldsymbol{\beta}^*\|/5 \geq \frac{\|\boldsymbol{\beta}^*\|}{10}(1 + 1 \cdot \min(1, \|\boldsymbol{\beta}^*\|^2)) \geq \|\boldsymbol{\beta}\|(1 + 1 \cdot \min(1, \|\boldsymbol{\beta}\|^2)).$$

Combining two cases, we have that

$$\|\boldsymbol{\beta}'\| \geq \|\boldsymbol{\beta}\|(1 + c_1 \cdot \min(1, \|\boldsymbol{\beta}\|^2)).$$

Now similarly to *Case (i)*, since $\langle \boldsymbol{\beta}', v_1\rangle$ is increasing in $\|\boldsymbol{\beta}\|$, when $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$, we have

$$\|\boldsymbol{\beta}'\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}(1 + c_2 \cdot \min(1, \|\boldsymbol{\beta}^*\|^2)),$$

where $c_2 = c_1/100$. Collecting all results in two cases, we have Lemma 13.

B Auxiliary Lemmas

B.1 Random Angle of a Gaussian

The following lemma characterizes the quality of a random initial iterate for EM.

Lemma 16 (Angle between a Gaussian and a fixed vector)**.** *Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with independent standard Gaussian entries. With probability at least $0.9 - 2\exp(-cd)$, we have $|\cos(\angle(X, \boldsymbol{e}_1)| = \boldsymbol{\beta}(1/\sqrt{d})$, where $\boldsymbol{e}_1 := (1, \ldots, 0)$ is the first standard basis vector in $\mathbb{R}^d$.*

Proof. Note that $\cos(\angle(\boldsymbol{X}, \boldsymbol{e}_1)) = \alpha_1/\sqrt{\sum_{i=1}^d \boldsymbol{x}_i^2}$. Since the $\boldsymbol{x}_i^2$'s are independent sub-exponential random variables, By standard concentration result ensures that

$$\mathbb{P}\left(\left|\sum_{i=1}^d \boldsymbol{x}_i^2 - d\right| > \delta d\right) \leq 2\exp(-cd\delta^2),$$

for some absolute constant $c > 0$. On the other hand, since $\alpha_1 \sim \mathcal{N}(0,1)$, we have $|\alpha_1| \in (0.01, 2)$ with probability at least 0.9. Combining, we conclude that with probability $0.9 - 2\exp(-0.01cd)$,

$$\frac{0.01}{\sqrt{1.1d}} \leq \frac{|\alpha_1|}{\sqrt{\sum_{i=1}^d \boldsymbol{x}_i^2}} \leq \frac{2}{\sqrt{0.9d}}.$$

□

B.2 Concentration of Second Order Moments

Lemma 17. *With probability at least $1 - \delta$, we have*

$$\|\frac{1}{n}\sum_{i=1}^n y_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top - I\|_{op} = O\left((\|\boldsymbol{\beta}^*\| + 1)^2 \sqrt{\frac{d\ln^2(n/\delta)}{n}}\right), \tag{56}$$

Proof. Let ε_i be an independent Rademacher variable and $z_i = \mathcal{N}(0,1)$. We can write $y_i = \varepsilon_i \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + z_i$. We use the truncation argument for the of concentration of higher order moments. First define the good event $\mathcal{E} := \{\forall i \in [n], |z_i| \leq \tau, |\boldsymbol{x}_i^\top \boldsymbol{\beta}^*| \leq \tau_2|\}$. We will decide the order of τ later such that $P(\mathcal{E}) \geq 1 - \delta$. Let $\widetilde{Y} \sim Y|\mathcal{E}, \widetilde{X} \sim X|\mathcal{E}$ and $(\widetilde{y}_i, \widetilde{\boldsymbol{x}}_i)$ be independent samples of $(\widetilde{Y}, \widetilde{\boldsymbol{X}})$. It is easy to check that $\widetilde{Y}\widetilde{\boldsymbol{X}}$ is a sub-Gaussian vector with Orlicz norm $O(\tau + \tau_2)$ [30]. To see this,

$$\left\|\widetilde{Y}\widetilde{\boldsymbol{X}}\right\|_{\psi_2} = \sup_{u \in \mathbb{S}^{d-1}} \sup_{p \geq 1} p^{-1/2} \mathbb{E}\left[|Y(\boldsymbol{X}^\top u)|^p | \mathcal{E}\right]^{1/p} \tag{57}$$

$$\leq (\tau + \tau_2) \sup_{u \in \mathbb{S}^{d-1}} \sup_{p \geq 1} p^{-1/2} \mathbb{E}\left[|\boldsymbol{X}^\top u|^p 1_{\mathcal{E}}\right]^{1/p} / \mathbb{P}(\mathcal{E})^{1/p} \tag{58}$$

$$\leq (\tau + \tau_2) K, \tag{59}$$

for some universal constant $K > 0$ and the last inequality comes from the p^{th} moments of Gaussian is $O((2p)^{p/2})$ and $P(\mathcal{E}) \geq 1 - \delta$.

Now we decompose the probability as the following:

$$\begin{aligned}
\mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^n y_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top - I\|_{\text{op}} \geq t\right) &\leq \mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^n y_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top - I\|_{\text{op}} \geq t | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq \underbrace{\mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^n \widetilde{y}_i^2 \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top - \mathbb{E}[\widetilde{y}^2 \widetilde{\boldsymbol{x}} \widetilde{\boldsymbol{x}}^\top]\|_{\text{op}} \geq t/2\right)}_{(a)} \\
&\quad + \underbrace{\mathbb{P}\left(\|\mathbb{E}[\widetilde{Y}^2 \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^\top] - I\|_{\text{op}} \geq t/2\right)}_{(b)} + \underbrace{\mathbb{P}(\mathcal{E}^c)}_{(c)}.
\end{aligned}$$

We can use a measure of concentration for random matrices for (a) given that $n \geq Cd$ for sufficiently large $C > 0$ [30], and bound by $\exp\left(-\frac{nt^2}{C(\tau+\tau_2)^4} + C'd\right)$ for some constants $C, C' > 0$. The bound for (c) is given by $n\exp(-\tau^2)$, hence we set

$$\tau = \Theta\left(\sqrt{\log(n/\delta)}\right), \tau_2 = \|\boldsymbol{\beta}^*\|\tau.$$

Finally, for (b), we first note that

$$\mathbb{E}[Y^2\boldsymbol{X}\boldsymbol{X}^\top] = \mathbb{E}[\widetilde{Y}^2\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top]P(\mathcal{E}) + \mathbb{E}[Y^2\boldsymbol{X}\boldsymbol{X}^\top 1_{\mathcal{E}^c}].$$

Rearranging the terms,

$$\begin{aligned}\|\mathbb{E}[\widetilde{Y}^2\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top] - I\|_{\mathrm{op}} &\leq \|\mathbb{E}[\widetilde{Y}^2\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top]\|_{\mathrm{op}}P(\mathcal{E}^c) + \sqrt{\sup_{u\in\mathbb{S}^d}\mathbb{E}[Y^4(\boldsymbol{X}^\top u)^4]}\sqrt{P(\mathcal{E}^c)} \\ &\leq (\tau+\tau_2)^2 n\exp(-\tau^2/2) + 3(\tau+\tau_2)^2\sqrt{n}\exp(-\tau^2/4) \leq \sqrt{1/n}.\end{aligned}$$

We can set $t = O\left((\|\boldsymbol{\beta}^*\|+1)^2\sqrt{d\log^2(n/\delta)/n}\right)$ and get the desired result. □

B.3 Initialization with Spectral Methods

Lemma 18. *Let $M = \frac{1}{n}\sum_{i=1}^n y_i^2\boldsymbol{x}_i\boldsymbol{x}_i^\top - I$. Let the largest eigenvalue and corresponding eigenvector of M be $(\lambda_1, \boldsymbol{v}_1)$. Then, there exists universal constants $c_0, c_1 > 0$ such that*

$$|\lambda_1 - \|\boldsymbol{\beta}^*\|^2| \leq c_0(\|\boldsymbol{\beta}^*\|^2 + 1)\sqrt{\frac{d\log^2(n/\delta)}{n}}.$$

Furthermore, if $\|\boldsymbol{\beta}^*\| \geq c_1(d\log^2(n/\delta)/n)^{1/4}$, then

$$\sin\angle(\boldsymbol{v}_1, \boldsymbol{\beta}^*) \leq c_0\left(1 + \frac{1}{\|\boldsymbol{\beta}^*\|^2}\right)\sqrt{\frac{d\log^2(n/\delta)}{n}} \leq \frac{1}{10}.$$

Proof. The lemma is a direct consequence of Lemma 17 and matrix perturbation theory [31]. Note that $\mathbb{E}[y_i^2\boldsymbol{x}_i\boldsymbol{x}_i^\top] = I + 2\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}$ (e.g., see Lemma 1 in [41]). □

The above lemma states that when $\|\boldsymbol{\beta}^*\|$ is not too small, we can always start from the well-initialized point where it is well aligned with ground truth $\boldsymbol{\beta}^*$. In low SNR regime where $\|\boldsymbol{\beta}^*\|^2 \lesssim (d/n)^{1/2}$, we cannot guarantee such a well-alignment with $\boldsymbol{\beta}^*$ since the eigenvector is perturbed too much. However, the largest eigenvalue can still serve as an indicator that $\|\boldsymbol{\beta}^*\|$ is small. Hence in all cases, we can initialize the estimator with $\widetilde{\boldsymbol{\beta}}_0 = \max\{0.2, \sqrt{\lambda_1}\}v_1$ to satisfy the initialization condition that we required in Phase 1.