

Structural Properties, Cycloid Trajectories and Non-Asymptotic Guarantees of EM Algorithm for Mixed Linear Regression

Zhankun Luo[✉], *Graduate Student Member, IEEE*, and Abolfazl Hashemi[✉], *Member, IEEE*

Abstract

This work investigates the structural properties, cycloid trajectories, and non-asymptotic convergence guarantees of the Expectation-Maximization (EM) algorithm for two-component Mixed Linear Regression (2MLR) with unknown mixing weights and regression parameters. Recent studies have established global convergence for 2MLR with known balanced weights and super-linear convergence in noiseless and high signal-to-noise ratio (SNR) regimes. However, the theoretical behavior of EM in the fully unknown setting remains unclear, with its trajectory and convergence order not yet fully characterized. We derive explicit EM update expressions for 2MLR with unknown mixing weights and regression parameters across all SNR regimes and analyze their structural properties and cycloid trajectories. In the noiseless case, we prove that the trajectory of the regression parameters in EM iterations traces a cycloid by establishing a recurrence relation for the sub-optimality angle, while in high SNR regimes we quantify its discrepancy from the cycloid trajectory. The trajectory-based analysis reveals the order of convergence: linear when the EM estimate is nearly orthogonal to the ground truth, and quadratic when the angle between the estimate and ground truth is small at the population level. Our analysis establishes non-asymptotic guarantees by sharpening bounds on statistical errors between finite-sample and population EM updates, relating EM's statistical accuracy to the sub-optimality angle, and proving convergence with arbitrary initialization at the finite-sample level. This work provides a novel trajectory-based framework for analyzing EM in Mixed Linear Regression.

Index Terms

Expectation-Maximization, Mixed Linear Regression, Cycloid Trajectory, Sub-optimality Angle, Convergence Analysis

I. INTRODUCTION

Mixture models of parameterized distributions, including Mixture of Linear Regression (MLR) and Gaussian Mixture Models (GMM), provide a flexible framework for modeling complex dependencies in practice. Its flexibility makes it well-suited for modeling datasets affected by corrupted observations, incomplete data, and latent variable structures [2]. In this work, we consider the symmetric two-component mixed linear regression (2MLR), defined as

$$y = (-1)^{z+1} \langle x, \theta^* \rangle + \varepsilon, \quad (1)$$

where $s := (x, y) \in \mathbb{R}^d \times \mathbb{R}$ denotes the observation, consisting of the covariate $x \in \mathbb{R}^d$ and the response $y \in \mathbb{R}$. The latent variable $z \in \{1, 2\}$, indicating the component label, follows the categorical distribution $\mathcal{CAT}(\pi^*)$ with probabilities $\mathbb{P}[z = 1] = \pi^*(1)$ and $\mathbb{P}[z = 2] = \pi^*(2)$. The additive noise ε is Gaussian with zero mean and variance σ^2 . The regression parameters are $\theta^* \in \mathbb{R}^d$, and the mixing weights are $\pi^* = (\pi^*(1), \pi^*(2))$, satisfying $\pi^*(1), \pi^*(2) \geq 0$ and $\pi^*(1) + \pi^*(2) = 1$. The ground-truth mixing weights $\pi^* = (\pi^*(1), \pi^*(2))$ and regression parameters θ^* of the data-generating process are unknown. The signal-to-noise ratio (SNR) is defined as $\eta := \|\theta^*\|/\sigma$. We observe i.i.d. samples $\mathcal{S} = \{s_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ drawn from this data-generating process.

Maximum Likelihood Estimation (MLE) provides a principled procedure for parameter inference of the above model. It seeks estimates (θ, π) of the ground-truth parameters (θ^*, π^*) by maximizing the likelihood. Given estimated parameters (θ, π) , where $\theta \in \mathbb{R}^d$ and $\pi = (\pi(1), \pi(2))$ with $\pi(1), \pi(2) \geq 0$ and $\pi(1) + \pi(2) = 1$, the population log-likelihood is defined as $\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} [\log p(s | \theta, \pi)]$, while in practice we maximize the empirical log-likelihood $\frac{1}{n} \sum_{i=1}^n \log p(s_i | \theta, \pi)$ based on the observed samples $\mathcal{S} = \{s_i\}_{i=1}^n$.

However, computing the Maximum Likelihood Estimate (MLE) for high-dimensional data is intractable due to its non-convexity and numerous spurious local maxima. A PCA-based technique [3] links the underlying geometric structure and probabilistic perspective with Gaussian covariates and noise. A meta-learning strategy [4], [5] was applied to estimate the parameters of MLR from small batches. The method in [6] introduced an iterative adaptation of least trimmed squares to cope with corrupted MLR data. Another alternative is the moment-based approach [7], which is combined with gradient descent.

Manuscript submitted to *IEEE Transactions on Information Theory* on Nov 6, 2025. The authors are with the School of Electrical and Computer Engineering, Purdue University, IN 47907, USA. (E-mail: {luo333, abolfazl}@purdue.edu.) An early version of this work have been presented in part at the 2024 International Conference on Machine Learning (ICML), July 2024, Vienna, Austria [1]. (*Corresponding author: Zhankun Luo.*)

Moreover, the Expectation Maximization (EM) method is particularly advantageous due to its computational efficiency and practical implementation simplicity. For the model in (1), EM estimates the regression parameters and mixing weights from observations through a two-step iterative process: The E-step evaluates the expected log-likelihood under the current parameter estimates, whereas the M-step updates the parameters by maximizing this expectation. This iterative process maximizes the lower bound on the MLE objective until convergence.

A foundational work [8] presented the modern EM algorithm and demonstrated its likelihood to be monotonically increasing with EM updates. The theoretical analysis in [9] established the global convergence of a unimodal likelihood under some regularity conditions. Empirical evidence [10]–[12] demonstrated EM success in the MLR problem. A comprehensive framework of EM for MLR involving unknown latent variables [13] was also introduced.

A. Motivation

As motivating applications, we emphasize the tasks of haplotype assembly in bioinformatics and genomics [14] and phase retrieval, which emerges in numerous fields such as optics, acoustics, and quantum information [15], demonstrating the practical implications of our study.

Haplotype Assembly. Haplotypes comprise sequences of chromosomal variations in genome that are pivotal for determining the human disease susceptibility. Haplotype assembly requires piecing together these sequences from a mixture of sequenced chromosome fragments. Particularly, humans carry two haplotypes, i.e., they are diploid organisms (see [14] for a clear mathematical formulation of the problem). For diploids, the main challenge is to reassemble two different haplotypes (binary sequences of single nucleotide polymorphisms (SNPs)) from corrupted short sequencing reads. Each read relates to a nearby window of the genome but comes from one of the pair of chromosomes. The uncertainty in the haplotype origin of each read can be characterized as a symmetric two-component mixed linear regression (2MLR), which is the model discussed in our work. As explained in [14], let $\theta^* \in \{-1, +1\}^d$ denote one haplotype, and the other haplotype is its opposite $-\theta^*$. The sequencing data consists of n corrupted reads $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \{0, 1\}^d$ is a sparse binary vector representing these SNPs covered by the read, and $y_i \in \mathbb{R}$ is the measured read signal (e.g., aggregate allele value), represented as

$$y_i = (-1)^{z_i+1} \langle x_i, \theta^* \rangle + \varepsilon_i, \quad (2)$$

where $z_i \in \{1, 2\}$ represents a latent variable showing whether the read comes from haplotype θ^* or $-\theta^*$, and ε_i stands for the additive noise. The research in [16] shows the effectiveness of EM-type methods for this problem, as the authors note that “this iterative update rule is reminiscent of the class of Expectation Maximization algorithms.”

Phase Retrieval. In the context of phase retrieval, as highlighted in Section 3 of [17] and elaborated in Section 3.5 of [18], a well-recognized connection exists between the phase retrieval problem and the symmetric 2MLR. Notably, squaring the response variable y and subtracting the variance σ^2 yields:

$$y' := y^2 - \sigma^2 = |\langle x, \theta^* \rangle|^2 + \xi. \quad (3)$$

Essentially, this reduces to the phase retrieval model with a heteroskedastic, zero-mean error $\xi := 2(-1)^{z+1} \langle x, \theta^* \rangle + (\varepsilon^2 - \sigma^2)$. Thus, convergence guarantees for the phase retrieval problem follow directly from the analysis of symmetric 2MLR.

B. Problem Setup and EM Updates

To develop a theoretical understanding of EM behavior for the 2MLR model, we adopt the standard assumption on the covariates $x \sim \mathcal{N}(0, I_d)$ (see also page 1 of [17], page 6 of [19], page 4 of [20]; and page 3 of [1]). The special case of 2MLR with known mixing weights has been intensively studied in a line of prior works [17], [19], [21], [22]. The population EM update rule for 2MLR with known balanced mixing weights $\pi = \pi^* = (\frac{1}{2}, \frac{1}{2})$ is given by [19] as follows:

$$M(\theta) = \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta \rangle}{\sigma^2}\right) yx. \quad (4)$$

We extend our analysis from the special case of known balanced mixing weights to the general case of unknown mixing weights by introducing an auxiliary variable ν to characterize the imbalance of the estimated mixing weights, i.e., $\tanh \nu = \pi(1) - \pi(2)$. The corresponding ground truth value ν^* satisfies the analogous relation $\tanh \nu^* = \pi^*(1) - \pi^*(2)$. Formally, these ν and ν^* are given by

$$\nu := \frac{\ln \pi(1) - \ln \pi(2)}{2}, \quad \nu^* := \frac{\ln \pi^*(1) - \ln \pi^*(2)}{2}. \quad (5)$$

With the help of ν , the population EM update rule for regression parameters θ with unknown mixing weights π is extended as follows:

$$M(\theta, \nu) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta \rangle}{\sigma^2} + \nu\right) yx. \quad (6)$$

In the absence of the assumption that the covariates satisfy $x \sim \mathcal{N}(0, I_d)$, so that $\mathbb{E}[xx^\top] = I_d$, the population EM update rule becomes $M(\theta, \nu) = \mathbb{E}[xx^\top]^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta \rangle}{\sigma^2} + \nu\right) yx$. Furthermore, the corresponding population EM update rule for the imbalance of mixing weights ν is given by:

$$N(\theta, \nu) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta \rangle}{\sigma^2} + \nu\right). \quad (7)$$

The finite-sample EM update rules with n samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ are:

$$\begin{aligned} M_n(\theta, \nu) &:= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \tanh\left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu\right) y_i x_i \right), \\ N_n(\theta, \nu) &:= \frac{1}{n} \sum_{i=1}^n \tanh\left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu\right). \end{aligned} \quad (8)$$

To facilitate theoretical analysis, we adopt the Easy EM method with the update rule of regression parameters as follows:

$$M_n^{\text{easy}}(\theta, \nu) := \frac{1}{n} \sum_{i=1}^n \tanh\left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu\right) y_i x_i. \quad (9)$$

Derivations of both the population and finite-sample EM update rules are provided in Appendix D, while similar derivations of EM updates for 2GMM can be found on pages 5-6 of [23].

C. Related Works

Both GMM and MLR can be interpreted as instances of subspace clustering, which naturally leads to similarities in their EM analysis. For GMM, well-separated spherical Gaussians with parameterization can be learned to nearly optimal accuracy using a variant of EM [24]. Moreover, the linear global convergence of EM under well-separated spherical Gaussians with initialization inside a neighborhood of the truth has been demonstrated [25]. For GMM with $k \geq 3$ components, EM with random initialization is often trapped in local minima with high probability, and the likelihood at local maxima can be arbitrarily worse than that of any global maximum [26]. The possible types of local minima for EM and k -means (i.e., EM with hard labels) in GMM under a separation condition have been identified [27], [28]. Building on these characterized structures of local minima, a general framework was introduced to escape them [29], [30], thereby unifying variants of k -means from a geometric standpoint. Finally, the relationship between EM and the method of moments for GMM, established through an asymptotic expansion of the log-likelihood in the low signal-to-noise ratio (SNR) regime, has been uncovered [31]. The particular case of GMM with $k = 2$ components (2GMM) has been extensively investigated. Global convergence of EM with random initialization for spherical 2GMM was demonstrated in [32]–[34]. This convergence property was analyzed across all SNR regimes in [35], while restricting the initialization to a very small neighborhood. The convergence study was additionally generalized from spherical Gaussians to rotation-invariant log-concave distributions in [36], [37]. The critical SNR threshold for exact recovery of 2GMM was explored in [38].

Similarly, EM for MLR with two components (2MLR) and random initialization achieves global convergence. The global convergence of EM for 2MLR with proper initialization within a ball around the true parameters was first established in [19]. This convergence guarantee was extended to the high SNR setting for the case where the cosine angle between the initial parameters and the ground truth is sufficiently large [17]. It was additionally verified in [21], [39] that EM for 2MLR converges from random initialization with high probability. Bounds on the statistical error of EM for 2MLR across varying SNR regimes were derived in [40]. Generalization error bounds for the log-likelihood of the first-order EM for 2MLR were presented in [41]. The statistical error and convergence rate of EM for 2MLR under all SNR regimes were further investigated in [22]. Alternating Minimization (AM), a hard-label variant of EM, for 2MLR in the noiseless case was studied in [42], [43]. Accordingly, a super-linear convergence rate of AM for 2MLR in the noiseless setting within a specific convergence region was shown in [44]. The noiseless scenario was generalized to the high SNR regime while maintaining super-linear convergence in [22]. Finally, a convergence analysis of EM for MLR with multiple components, covering the most general settings, was given in [45]. Prior works on the convergence analysis of EM for 2MLR have neglected unbalanced mixing weights and focused on the balanced scenario. The works in [46]–[48] incorporated unbalanced mixing weights of 2GMM/2MLR and demonstrated a pronounced difference in statistical error and convergence rate between the unbalanced and balanced settings, for the specific case without parameter separation, namely when the location/regression parameters are degenerate, $\theta^* = \vec{0}$, in 2GMM/2MLR.

Therefore, it remains an open question to investigate the convergence of EM for 2MLR with unknown mixing weights and regression parameters. Our work addresses this gap by providing a comprehensive analysis of the properties of EM updates, convergence guarantees, and statistical errors of EM iterations for 2MLR with unknown mixing weights and regression parameters.

D. Main Contributions

In this work, we provide explicit expressions for the log-likelihood function (Proposition 1) of the symmetric two-component Mixed Linear Regression (2MLR) model with unknown mixing weights, and subsequently derive the EM update rules (Equations (6), (7), (8)). We establish the connection between these EM updates and the gradient-descent dynamics of the log-likelihood (Proposition 3), and obtain closed-form expressions of the EM update rules with unknown mixing weights and regression parameters (Theorem 4).

We characterize the structural properties of the EM updates by bounding the length of EM iterations across all signal-to-noise ratio (SNR) regimes (Proposition 5), identifying their fixed points and proving contraction around them (Propositions 6, 7). In the noiseless setting, we show that the EM iterations follow a cycloid trajectory (Proposition 11), derive this from the explicit noiseless updates (Corollary 8), and quantify the deviation from this trajectory when the SNR is large but finite (Proposition 9).

In the noiseless setting, we establish a recurrence relation for the sub-optimality angle (Proposition 10), prove linear growth when the EM estimate is nearly orthogonal to the ground truth (Proposition 12), and show quadratic convergence when the angle is small (Proposition 13). We relate the accuracy of EM estimate (for regression parameters and mixing weights) to the sub-optimality angle (Proposition 14), yielding population-level convergence guarantees (Theorem 15).

At the finite-sample level, we couple the population-level analysis with statistical error bounds to establish convergence guarantees for the EM algorithm under arbitrary initialization. Specifically, we derive projected and statistical error bounds for the regression parameters (Propositions 16, 17), which are independent of the mixing weights, and a statistical error bound for the mixing weights (Proposition 18) that depends on the sub-optimality angle and the ground-truth mixing weights. We then characterize the connection between the statistical accuracy of the EM updates and the sub-optimality angle (Proposition 19) and establish convergence guarantees for the sub-optimality angle at the finite-sample level (Proposition 21) by combining the population-level convergence analysis with the bounds on the statistical errors and statistical accuracy of the EM updates. Consequently, by leveraging the connection between the statistical accuracy of the EM updates and the sub-optimality angle, we obtain finite-sample convergence guarantees for the EM algorithm with arbitrary initialization (Proposition 20) of unknown mixing weights and regression parameters (Theorem 22).

Overall, our results correspond directly to the three main components of this work:

- **Structural Properties:** We provide a unified characterization of the structural properties of the EM updates for the symmetric 2MLR model with unknown mixing weights and regression parameters across all SNR regimes (Section II);
- **Cycloid Trajectories:** We introduce a trajectory-based framework that reveals the cycloid trajectory of EM iterations and connects their geometric behavior to the accuracy of the EM updates (Section III); and
- **Non-Asymptotic Guarantees:** We establish convergence guarantees for finite-sample EM updates under arbitrary initialization of the mixing weights and regression parameters in the noiseless setting (Section IV).

E. Notations

Imbalance of Mixing Weights. The absolute value of the imbalance $\tanh|\nu|$ corresponds exactly to the ℓ_1 distance between the mixing weights $\pi = (\pi(1), \pi(2))$ and the balanced weights $\frac{\mathbf{1}}{2} = (\frac{1}{2}, \frac{1}{2})$, where $\mathbf{1} = (1, 1)$ is the all-ones vector. These quantities are related to the minimum of $\pi(1)$ and $\pi(2)$ by:

$$\tanh|\nu| = \left\| \pi - \frac{\mathbf{1}}{2} \right\|_1, \quad \frac{1 - \tanh|\nu|}{2} = \min(\pi(1), \pi(2)). \quad (10)$$

The ground truth counterparts follow the same relations, namely $\tanh|\nu^*| = \left\| \pi^* - \frac{\mathbf{1}}{2} \right\|_1$ and $\frac{1 - \tanh|\nu^*|}{2} = \min(\pi^*(1), \pi^*(2))$.

Sub-optimality Angles. The ℓ_2 norm of a vector is denoted by $\|\cdot\|$, and $\langle \cdot, \cdot \rangle$ represents the inner product. We define the unit direction vectors for the estimated and ground truth regression parameters as $\vec{e}_1 = \theta/\|\theta\|$ and $\hat{e}_1 = \theta^*/\|\theta^*\|$, respectively. The corresponding orthogonal direction vectors are $\vec{e}_2 = (\theta^* - \vec{e}_1 \vec{e}_1^\top \theta)/\|\theta^* - \vec{e}_1 \vec{e}_1^\top \theta\|$ and $\hat{e}_2 = (\theta - \hat{e}_1 \hat{e}_1^\top \theta^*)/\|\theta - \hat{e}_1 \hat{e}_1^\top \theta^*\|$. The cosine of the angle between the estimated regression parameters θ and the ground truth θ^* is $\rho := \langle \theta, \theta^* \rangle / \|\theta\| \|\theta^*\|$, with its sign denoted by $\text{sgn}(\rho)$ and $\text{sgn}(0) = 1$. From this, we define the sub-optimality angles φ and ϕ :

$$\varphi := \frac{\pi}{2} - \arccos|\rho|, \quad \phi := 2 \arccos|\rho|. \quad (11)$$

The ratio of the norms of the regression parameters is denoted by $k := \|\theta\|/\|\theta^*\|$.

Auxiliary Definitions. We define A_η and φ_η in terms of the SNR η , the ratio k , and the sub-optimality angles φ, ϕ :

$$A_\eta := k\eta^2 \sqrt{1 + \eta^{-2}}, \quad \varphi_\eta := \arcsin \frac{\sin \varphi}{\sqrt{1 + \eta^{-2}}} \quad (12)$$

with the convention $\sqrt{1 + \eta^{-2}} = \sin \varphi / \sin \varphi_\eta$ when $\varphi = 0$. To express results compactly, we use Gaussian random variables $g, g_\eta, g' \sim \mathcal{N}(0, 1)$ with the following correlation structure:

$$\mathbb{E}[gg'] = \sin \varphi, \quad \mathbb{E}[g_\eta g'] = \sin \varphi_\eta, \quad \mathbb{E}[gg_\eta] = \frac{1}{\sqrt{1 + \eta^{-2}}} = \frac{\sin \varphi_\eta}{\sin \varphi}. \quad (13)$$

The modified Bessel function of the second kind of order 0, denoted $K_0(\cdot)$, is introduced to aid our analysis. The probability density function of the product of two independent Gaussian random variables with zero mean and unit variance is exactly $K_0(|\cdot|)/\pi$ (see page 50, Section 4.4 Bessel Function Distributions, Chapter 12 Continuous Distributions (General) of [49]).

We use the asymptotic notations $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, and $\Theta(\cdot)$ (see page 528 of [50]) to help our analysis. The notation $f = \Omega(g)$, which can also be written as $f \gtrsim g$, indicates $g = \mathcal{O}(f)$, equivalently $g \lesssim f$; that is, there exists a universal constant $C > 0$ such that $|f(x)| \geq C|g(x)|$ for all sufficiently large x . The notation $f = \Theta(g)$, or $f \asymp g$, means that both $f = \mathcal{O}(g)$ and $g = \mathcal{O}(f)$ hold. The symbols $a \vee b$ and $a \wedge b$ denote the maximum and minimum of a and b , respectively.

F. Organization

We organize the paper as follows:

In Section II, we provide the explicit expressions of the negative log-likelihood and EM update rules for the symmetric 2MLR model with unknown mixing weights and regression parameters, and characterize the structural properties of the EM updates for the symmetric 2MLR model across all SNR regimes. The detailed derivations of the negative log-likelihood and EM update rules are given in Appendix D, while the proofs of the structural properties of the EM updates are given in Appendix A.

In Section III, We analyze the population EM updates in the noiseless setting, showing that the population EM iterations evolve along a cycloid trajectory governed by a recurrence relation of the sub-optimality angle, exhibiting linear growth and quadratic convergence of the sub-optimality angle, and relating the accuracy of mixing weights and regression parameters to the sub-optimality angle. The proofs of the results in this section are given in Appendix B.

In Section IV, we establish the non-asymptotic convergence guarantees of the finite-sample EM updates in the noiseless setting by deriving statistical error bounds for both regression parameters and mixing weights, characterizing the connection between the statistical accuracy of EM updates and the sub-optimality angle, and showing convergence guarantees of the sub-optimality angle under arbitrary initialization. The proofs of the results in this section are given in Appendix C.

In Section V, we provide numerical experiments to validate the theoretical results. We first show the cycloid trajectory of the EM iterations empirically in the noiseless setting, and then validate the theoretical findings on the convergence guarantees for the EM algorithm with arbitrary initialization of unknown mixing weights and regression parameters at finite-sample level. The code for empirical experiments is available at https://github.com/dassein/cycloid_em_tit.

II. EXPLICIT EM UPDATE EXPRESSIONS AND PROPERTIES OF EM UPDATE RULES

In this section, we focus on the general structural properties of EM updates for 2MLR across all SNR regimes, which consists of three parts: In the first part, we give the expressions of the negative log-likelihood (Proposition 1) at the population level and the finite-sample level by using expectations with respect to the ground truth distribution of the observed data $s = (x, y)$. Then we rewrite the EM update rules (Lemma 2) by using expectations with respect to the ground truth distribution of the observation data $s = (x, y)$. Finally, we establish the connection between the EM update rules and the gradient descent of the negative log-likelihood function (Proposition 3).

In the second part, we characterize the properties of the EM update rules by starting from giving the explicit expectation expressions of the EM update rules with respect to only two Gaussian random variables (Theorem 4). Based on the explicit expressions, we establish bounds for the length of the EM update rules for regression parameters (Proposition 5). Then we give fixed points for the EM update rules (Propositions 6) and provide an analysis of the contraction of the EM update rules around the fixed points (Proposition 7).

In the third part, we give the EM update rules with only the sub-optimality angle φ in the noiseless setting (Corollary 8), i.e., when the SNR $\eta \rightarrow \infty$, which will be shown to follow a cycloid trajectory in the next section (Proposition 11 in Section III). Furthermore, we bound the difference between the EM update rule and its limit of the cycloid trajectory in the finite high SNR regime (Proposition 9). The proofs of the results in this section are provided in Appendix A, and the detailed derivations of EM update rules are provided in Appendix D.

Proposition 1 (Population and Finite-Sample Negative Log-Likelihood). *Let $f(\theta, \pi) := -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\ln p(s | \theta, \pi)]$ be the negative log-likelihood function at the population level, and $f_n(\theta, \pi) := -\frac{1}{n} \sum_{i=1}^n \ln p(s_i | \theta, \pi)$ be the negative log-likelihood function at the finite-sample level for the dataset $\mathcal{S} = \{s_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples. Then, $f(\theta, \pi)$ and $f_n(\theta, \pi)$ can be expressed as:*

$$\begin{aligned} f(\theta, \pi) &= \frac{1}{2\sigma^2} \langle \theta, \mathbb{E}[xx^\top] \cdot \theta \rangle - \mathbb{E} \left[\ln \frac{\cosh(y\langle x, \theta \rangle / \sigma^2 + \nu)}{\cosh \nu} \right] + C \\ f_n(\theta, \pi) &= \frac{1}{2\sigma^2} \left\langle \theta, \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \cdot \theta \right\rangle - \frac{1}{n} \sum_{i=1}^n \ln \frac{\cosh(y_i \langle x_i, \theta \rangle / \sigma^2 + \nu)}{\cosh \nu} + C_n \end{aligned} \quad (14)$$

where $\mathbb{E}[\cdot] = \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\cdot]$ is the expectation over the ground truth distribution $p(s | \theta^*, \pi^*)$, and C and C_n are constants that are independent of θ and π .

Remark. For the 2MLR model with Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, we have shown that the negative log-likelihood function consists of three terms. The first term is quadratic in the regression parameters θ , involving the covariance matrix of the covariates x at the population level, or the sample covariance matrix at the finite-sample level. Under the assumption that the covariates $x \sim \mathcal{N}(0, I_d)$, we have $\mathbb{E}[xx^\top] = I_d$ at the population level, and the first term simplifies to $\|\theta\|^2/2\sigma^2$. The third term at the population level, $C = \frac{1}{2\sigma^2} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[y^2] + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\ln p(x)] - \frac{1}{2} \ln(2\pi\sigma^2)$, is a constant term that depends on the ground truth distribution of the observed data $s = (x, y) \sim p(s | \theta^*, \pi^*)$. Similarly, the third term at the finite-sample level $C_n = \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^n y_i^2 + \frac{1}{n} \sum_{i=1}^n \ln p(x_i) - \frac{1}{2} \ln(2\pi\sigma^2)$ is a constant that depends on the samples $\mathcal{S} = \{s_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(s_i | \theta^*, \pi^*)$ drawn from the ground truth distribution. Also, f is the KL divergence between the ground truth distribution $p(s | \theta^*, \pi^*)$ and the model distribution $p(s | \theta, \pi)$, plus a constant term (see Appendix D).

Lemma 2 (EM Update Rules and Gradients). *Let $U(\theta, \nu) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \ln \cosh(y\langle x, \theta \rangle / \sigma^2 + \nu)$ at population level and $U_n(\theta, \nu) := \frac{1}{n} \sum_{i=1}^n \ln \cosh(y_i \langle x_i, \theta \rangle / \sigma^2 + \nu)$ at finite-sample level, then the population EM update rules $M(\theta, \nu), N(\theta, \nu)$ and finite-sample EM update rules $M_n(\theta, \nu), N_n(\theta, \nu)$ for regression parameters θ and imbalance of mixing weights $\tanh \nu$ are:*

$$\begin{aligned} M(\theta, \nu) &= \mathbb{E}[xx^\top]^{-1} \sigma^2 \nabla_\theta U(\theta, \nu), & N(\theta, \nu) &= \nabla_\nu U(\theta, \nu), \\ M_n(\theta, \nu) &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \sigma^2 \nabla_\theta U_n(\theta, \nu), & N_n(\theta, \nu) &= \nabla_\nu U_n(\theta, \nu). \end{aligned} \quad (15)$$

Remark. The EM algorithm maximizes the evidence lower bound (ELBO) on the observed log-likelihood, namely by minimizing the surrogate functions of the negative log-likelihood functions f and f_n at the population level and the finite-sample level, respectively (see Appendix D for the expressions of the surrogate functions). The minimizers of the surrogate functions yield the EM update rules for regression parameters θ and the imbalance of mixing weights $\tanh \nu = \pi(1) - \pi(2)$: $M(\theta, \nu), N(\theta, \nu)$ and $M_n(\theta, \nu), N_n(\theta, \nu)$ at the population level and the finite-sample level, respectively (see equations (6), (7), (8) for EM update rules). By Leibniz's rule, we exchange the order of taking expectation and gradient of $\ln(\cosh(y\langle x, \theta \rangle / \sigma^2 + \nu))$, and noting that $\frac{d}{dt} \ln \cosh(t) = \tanh(t)$, we have established the above relation between EM update rules and the functions $U(\theta, \nu), U_n(\theta, \nu)$.

Proposition 3 (Connection between EM Update Rules and Gradient Descent). *The population/finite-sample EM update rules give:*

$$\begin{aligned} M(\theta, \nu) &= \theta - \mathbb{E}[xx^\top]^{-1} \sigma^2 \nabla_\theta f(\theta, \pi), & N(\theta, \nu) &= \tanh \nu - \nabla_\nu f(\theta, \pi), \\ M_n(\theta, \nu) &= \theta - \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \sigma^2 \nabla_\theta f_n(\theta, \pi), & N_n(\theta, \nu) &= \tanh \nu - \nabla_\nu f_n(\theta, \pi). \end{aligned} \quad (16)$$

Remark. Note that the negative log-likelihood functions f, f_n can be expressed in terms of the functions U, U_n as $f(\theta, \pi) = \frac{1}{2\sigma^2} \langle \theta, \mathbb{E}[xx^\top] \cdot \theta \rangle + \ln \cosh \nu - U(\theta, \nu)$ and $f_n(\theta, \pi) = \frac{1}{2\sigma^2} \langle \theta, \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \cdot \theta \rangle + \ln \cosh \nu - U_n(\theta, \nu)$. The first term depends only on the regression parameters θ , the second term depends only on the imbalance of mixing weights $\tanh \nu$, and the third term, given by $-U$ or $-U_n$, is a function of both θ and ν . By taking the gradients of the negative log-likelihood functions f and f_n with respect to θ and ν , and using the above relation between the EM update rules and the functions U, U_n , we establish the connection between the EM update rules and the gradient descent dynamics for the negative log-likelihood functions f and f_n (Proposition 3). This relation implies that $(\theta, \tanh \nu)$ is a fixed point of the population EM update rules $M(\theta, \nu), N(\theta, \nu)$ if and only if $(\theta, \tanh \nu)$ is a stationary point of the negative log-likelihood function f , namely $\nabla_\theta f = \vec{0}, \nabla_\nu f = 0$ at this point. See also the discussion of the fixed points of EM update rules in Proposition 6 and Proposition 7. Our results extend the connection between EM update rules only for regression parameters and gradient descent of the negative log-likelihood under known balanced mixing weights $\pi = \pi^* = (\frac{1}{2}, \frac{1}{2})$ (see Lemma 7 on page 19 of [21], Lemma 2 on page 6523 of [51]) to the more general case with unknown mixing weights for EM updates of regression parameters and mixing weights.

Theorem 4 (Explicit EM Update Expressions). *Let $A_\eta = k\eta^2 \sqrt{1 + \eta^{-2}}, k = \frac{\|\theta\|}{\|\theta^*\|}, \eta = \frac{\|\theta^*\|}{\sigma}$, and $\varphi = \frac{\pi}{2} - \arccos |\rho| \in (0, \pi/2]$, $\rho = \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$, and $\varphi_\eta = \arcsin(\sin \varphi / \sqrt{1 + \eta^{-2}})$, namely $\sqrt{1 + \eta^{-2}} = \sin \varphi / \sin \varphi_\eta$, and Gaussian random variables $g_\eta, g' \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[g_\eta g'] = \sin \varphi_\eta$, then population EM update rules for regression parameters θ and imbalance of mixing weights $\tanh \nu$ are:*

$$\begin{aligned} M(\theta, \nu) / \|\theta^*\| &= \tilde{e}_1 \frac{\sin \varphi}{\sin \varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g' + \text{sgn}(\rho)(-1)^{z+1} \nu) g_\eta g'] \\ &\quad + \tilde{e}_2 \frac{\cos \varphi}{\cos^2 \varphi_\eta} \text{sgn}(\rho) \mathbb{E}[\tanh(A_\eta g_\eta g' + \text{sgn}(\rho)(-1)^{z+1} \nu) g_\eta (g_\eta - \sin \varphi_\eta \cdot g')] \\ N(\theta, \nu) &= \text{sgn}(\rho) \mathbb{E}[\tanh((-1)^{z+1} A_\eta g_\eta g' + \text{sgn}(\rho) \nu)]. \end{aligned} \quad (17)$$

Remark. The above explicit expressions of EM update rules still hold when $\rho = 0$, namely when $\langle \theta, \theta^* \rangle = 0$, under the convention that $\text{sgn}(\rho) = \text{sgn}(0) = 1, \sin \varphi / \sin \varphi_\eta = \sqrt{1 + \eta^{-2}}$. Moreover, the above explicit expressions of EM update rules show that the EM update for regression parameters $M(\theta, \nu) \in \text{span}\{\theta, \theta^*\}$ must lie in the space spanned by θ, θ^* . The population EM update rules for the normalized regression parameters and the imbalance of mixing weights, $M(\theta, \nu) / \|\theta^*\|$

and $N(\theta, \nu)$, are determined by $\rho = \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$, $k = \frac{\|\theta\|}{\|\theta^*\|}$, $\tanh \nu = \pi(1) - \pi(2)$, and the SNR $\eta = \|\theta^*\|/\sigma$, while the other parameters $A_\eta = k\eta^2 \sqrt{1 + \eta^{-2}}$ and $\sin \varphi_\eta = \sin \varphi / \sqrt{1 + \eta^{-2}}$ are determined by k, η , and $\rho = \text{sgn}(\rho) \sin \varphi$. The application of Theorem 4 enables us to establish the following structural properties of EM update rules on boundedness (Proposition 5), fixed points (Proposition 6), the contraction property (Proposition 7), and the cycloid trajectory (Corollary 8, Proposition 9).

Proposition 5 (Boundedness of EM Update Rule). *The EM update rule for θ is bounded by the following bound, which depends on SNR $\eta = \|\theta^*\|/\sigma$:*

$$\|M(\theta, \nu)\| \leq \frac{\arctan \eta}{\pi/2} \|\theta^*\| + \frac{2}{\pi} \sigma. \quad (18)$$

Remark. The above bound for the EM update rule characterizes the range of the EM update rule for regression parameters θ . When the SNR η is small, the noise term dominates the magnitude of the EM update. In particular, when $\eta \rightarrow 0$ and the magnitude of the previous iteration $\|\theta\|/\sigma \rightarrow \infty$, we have $A_\eta \rightarrow \infty$ and $\varphi_\eta \rightarrow 0$, and therefore $\|M(\theta, \nu)\| \rightarrow \frac{2}{\pi} \sigma$. When the SNR η is large, the $\|\theta\|$ term dominates the noise term. Specifically, as $\eta \rightarrow \infty$, $|\rho| \rightarrow 1$ and $\|\theta\| \neq 0$, we have $A_\eta \rightarrow \infty$ and $\varphi_\eta \rightarrow \varphi = \frac{\pi}{2}$, and hence $\|M(\theta, \nu)\| \rightarrow \|\theta^*\|$. Therefore, our bound on the magnitude of the EM update rule $\|M(\theta, \nu)\|$ is tight in the limiting SNR regimes, both as $\eta \rightarrow 0$ and $\eta \rightarrow \infty$. Our result on the boundedness of the EM update rule for regression parameters is consistent with the previous bound $3\sqrt{\|\theta^*\|^2 + \sigma^2}$ obtained under known balanced mixing weights $\pi = \pi^* = (\frac{1}{2}, \frac{1}{2})$ (see Lemma 22 on page 52 of [21]), since $\frac{\arctan \eta}{\pi/2} \|\theta^*\| + \frac{2}{\pi} \sigma \leq \|\theta^*\| + \frac{2}{\pi} \sigma \leq \sqrt{(1 + 4/\pi^2)(\|\theta^*\|^2 + \sigma^2)} \leq 1.2\sqrt{\|\theta^*\|^2 + \sigma^2} \leq 3\sqrt{\|\theta^*\|^2 + \sigma^2}$.

Proposition 6 (Distinct Fixed Points of EM Update Rules). *The EM update rules $M(\theta, \nu)$, $N(\theta, \nu)$ for regression parameters θ and the imbalance of mixing weights $\tanh \nu$ have the following three distinct fixed points (θ^*, ν^*) , $(-\theta^*, -\nu^*)$, $(\vec{0}, 0)$:*

$$\begin{pmatrix} M(\theta^*, \nu^*) \\ N(\theta^*, \nu^*) \end{pmatrix} = \begin{pmatrix} \theta^* \\ \tanh \nu^* \end{pmatrix}, \quad \begin{pmatrix} M(-\theta^*, -\nu^*) \\ N(-\theta^*, -\nu^*) \end{pmatrix} = \begin{pmatrix} -\theta^* \\ \tanh(-\nu^*) \end{pmatrix}, \quad \begin{pmatrix} M(\vec{0}, 0) \\ N(\vec{0}, 0) \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \tanh 0 \end{pmatrix}. \quad (19)$$

Remark. Among these three distinct fixed points, the first two are the ground truth parameters (θ^*, ν^*) and $(-\theta^*, -\nu^*)$, and they are the only two global minimizers of the negative log-likelihood function $f(\theta, \pi)$ at the population level. Since f is equal to the KL divergence between the ground truth distribution $p(s | \theta^*, \pi^*)$ and the model distribution $p(s | \theta, \pi)$, plus a constant term (see Appendix D), these two fixed points uniquely minimize f , achieving a KL divergence value of zero. Apart from these three distinct fixed points, there exist two additional fixed points of the EM update rules whose regression parameters lie in the plane $\text{span}\{\theta, \theta^*\}$, as shown in Proposition 7. These points exhibit a contraction property along the direction of θ when θ is not in the same/opposite direction of ground truth θ^* . This result is also consistent with the previous findings on the fixed points of the EM update rule for regression parameters only, under known balanced mixing weights $\pi = \pi^* = (\frac{1}{2}, \frac{1}{2})$ (see Theorem 1 on page 6 of [21] and Theorem 2 on page 6523 of [51]).

Proposition 7 (Contraction Property of EM Update Rule around the Fixed Point). *Suppose that the unit direction vector $\vec{e}^\perp$ is orthogonal to the ground truth θ^* of regression parameters, namely $\langle \theta^*, \vec{e}^\perp \rangle = 0$, $\|\vec{e}^\perp\| = 1$, there exists $k^*(\eta) > 0$ which is determined by $\eta = \|\theta^*\|/\sigma$ such that the $(k^*(\eta)\|\theta^*\|\vec{e}^\perp, 0)$ is a fixed point of the EM update rule $M(\theta, \nu)$, $N(\theta, \nu)$ for regression parameters θ and the imbalance of mixing weights $\tanh \nu$ given any SNR $\eta = \|\theta^*\|/\sigma > 0$:*

$$\begin{pmatrix} M(k^*(\eta)\|\theta^*\|\vec{e}^\perp, 0) \\ N(k^*(\eta)\|\theta^*\|\vec{e}^\perp, 0) \end{pmatrix} = \begin{pmatrix} k^*(\eta)\|\theta^*\|\vec{e}^\perp \\ \tanh(0) \end{pmatrix}. \quad (20)$$

For any $k > 0$, $N(k\|\theta^*\|\vec{e}^\perp, 0) = \tanh(0)$ and $M(k\|\theta^*\|\vec{e}^\perp, 0), \vec{e}^\perp$ have the same direction,

$$\begin{cases} 0 < \|M(k\|\theta^*\|\vec{e}^\perp, 0)\|/\|\theta^*\| - k < k^*(\eta) - k & \text{if } k < k^*(\eta), \\ 0 = \|M(k\|\theta^*\|\vec{e}^\perp, 0)\|/\|\theta^*\| - k = k^*(\eta) - k & \text{if } k = k^*(\eta), \\ 0 > \|M(k\|\theta^*\|\vec{e}^\perp, 0)\|/\|\theta^*\| - k > k^*(\eta) - k & \text{if } k > k^*(\eta). \end{cases} \quad (21)$$

Moreover, the bounds of $k^*(\eta)$ are:

$$\frac{1}{\sqrt{3}} < k^*(\eta) < \min\left(\frac{2}{\pi} \sqrt{1 + \eta^{-2}}, 1\right), \forall \eta > 0, \quad (22)$$

with $\lim_{\eta \rightarrow 0^+} k^*(\eta) = \frac{1}{\sqrt{3}}$ and $\lim_{\eta \rightarrow \infty} k^*(\eta) = \frac{2}{\pi}$.

Remark. This indicates that there is a unique fixed point in the direction of $\vec{e}^\perp$ (distinct from the fixed point in the opposite direction), which is orthogonal to the ground truth θ^* of regression parameters. Furthermore, if the previous iteration lies in the direction of $\vec{e}^\perp$, the population EM update rule will also remain in this direction of $\vec{e}^\perp$, and the EM iterations get progressively closer to the fixed point, showing the contraction property of the EM update rule along $\vec{e}^\perp$. Moreover, we demonstrate that the bounds for the normalized length of the EM update rule $k^*(\eta)$ are tight across all SNR regimes. In particular, we have shown that the fixed point $k^*(\eta) \rightarrow \frac{2}{\pi}$ as the SNR $\eta \rightarrow \infty$, which aligns with the result that the diameter of the rolling circle of the cycloid trajectory for EM iterations is $\frac{2}{\pi} \|\theta^*\|$ in the noiseless setting (Proposition 11 in Section III).

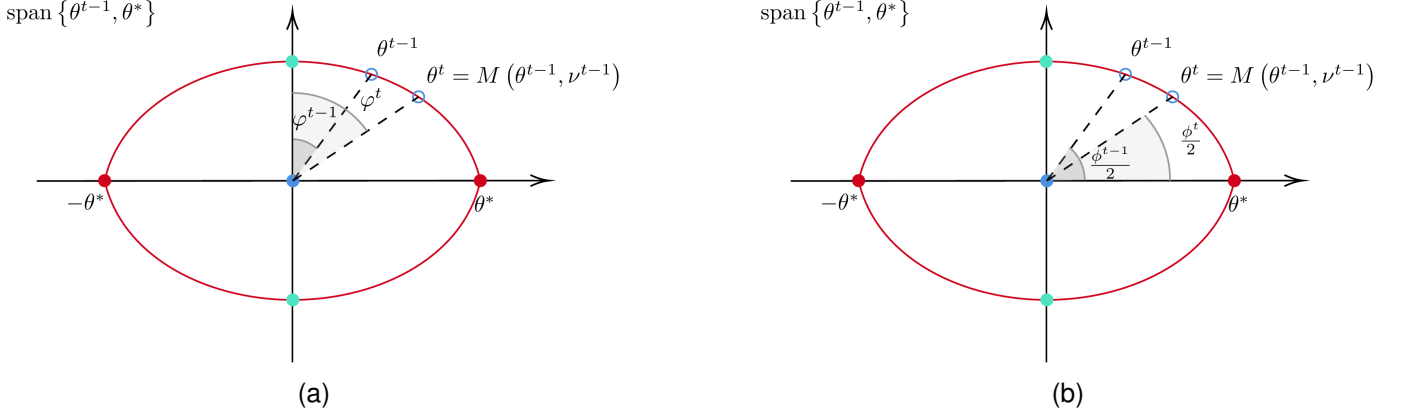


Fig. 1: The cycloid trajectory for the EM update rule $M(\theta, \nu)$ of regression parameters in the noiseless setting (SNR $\eta \rightarrow \infty$), and the fixed points of the population EM update rules are shown in the figure: these two red points stand for the ground truth parameters θ^* and $-\theta^*$, the blue point stands for the unstable fixed point $\vec{0}$ as distinct fixed points of the EM update rules in Proposition 6, and the green points stand for the two saddle points $\pm \lim_{\eta \rightarrow \infty} k^*(\eta) \|\theta^*\| \hat{e}_2 = \pm \frac{2}{\pi} \|\theta^*\| \hat{e}_2$ as fixed points of the EM update rules on the plane $\text{span}\{\theta, \theta^*\}$ in Proposition 7 with contraction property along the direction of $\pm \hat{e}_2$ orthogonal to the ground truth θ^* . (a) Sub-optimality angle φ : the angle between the unit direction vector $\hat{e}_2$ and the regression parameters θ ; φ^t, φ^{t-1} correspond to the sub-optimality angles at the t -th and $(t-1)$ -th EM iterations, where the regression parameters take the values θ^t and θ^{t-1} , respectively. (b) Sub-optimality angle ϕ : twice the minimum angle between θ and $\pm \theta^*$, i.e., $\phi = 2 \arccos |\langle \theta, \theta^* \rangle| / (\|\theta\| \|\theta^*\|)$; in the noiseless setting, θ^t follows a cycloid trajectory with rolling radius $\|\theta^*\|/\pi$, where the rolling angle is determined by the previous sub-optimality angle ϕ^{t-1} (Proposition 11).

Corollary 8 (EM Updates in Noiseless Setting, Corollary 3.3 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, the population EM update rules $M(\theta, \nu), N(\theta, \nu)$ for regression parameters θ and imbalance of mixing weights $\tanh \nu$ are:*

$$\begin{aligned} \frac{M(\theta, \nu)}{\|\theta^*\|} &= \frac{2}{\pi} \left[\text{sgn}(\rho) \varphi \frac{\theta^*}{\|\theta^*\|} + \cos \varphi \frac{\theta}{\|\theta\|} \right] \\ N(\theta, \nu) &= \text{sgn}(\rho) \frac{2}{\pi} \varphi \cdot \tanh \nu^* \end{aligned} \quad (23)$$

where $\varphi := \frac{\pi}{2} - \arccos |\rho|$, $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$.

Remark. In the noiseless setting, the first equation shows that the EM update rule is clearly a linear combination of the ground truth θ^* and the current estimate θ . It also indicates that the length of the normalized EM update rule reaches its minimum value $\frac{2}{\pi}$ when $\varphi = 0$, namely $\langle \theta, \theta^* \rangle = 0$; and the length of the normalized EM update rule reaches its maximum value 1 when $\varphi = \frac{\pi}{2}$, namely θ^*, θ are in the same/opposite direction. Using the first equation, we show that the EM iterations of regression parameters θ follow a cycloid trajectory, whose rolling diameter is $\frac{2}{\pi} \|\theta^*\|$ (Proposition 11 in Section III). The second equation shows that the EM update rule for the imbalance of mixing weights $\tanh \nu$ is independent of the previous iteration $\tanh \nu$, and depends only on the ground truth imbalance of mixing weights $\tanh \nu^*$ and the sub-optimality angle φ , which is determined by the ground truth θ^* and the current estimate θ . This suggests that once the evolution of the sub-optimality angle φ is characterized in the noiseless setting, then the evolution of the imbalance of mixing weights $\tanh \nu$ can also be determined.

Proposition 9 (Deviation from Cycloid Limit of EM Updates in High SNR Regime). *In the finite high SNR regime given by $\eta \gtrsim \frac{1}{\min(1, \sqrt{k}) \cos \varphi} \vee 1$, if the mixing weights are known balanced $\pi = \pi^* = (\frac{1}{2}, \frac{1}{2})$, then the difference between the EM update rule and its limit is bounded by:*

$$\left\| \frac{M(\theta)}{\|\theta^*\|} - \lim_{\eta \rightarrow \infty} \frac{M(\theta)}{\|\theta^*\|} \right\| = \mathcal{O} \left(\eta^{-2} \vee \cos^2 \varphi \frac{\log \Lambda}{\Lambda^4} \right) \quad (24)$$

where $\Lambda := \eta \sqrt{k} \cos \varphi$, $k := \frac{\|\theta\|}{\|\theta^*\|}$ and $\varphi := \frac{\pi}{2} - \arccos |\rho|$, $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$.

Remark. This result provides an explicit bound on the difference between the EM update rule and its limit (cycloid trajectory) in the finite high-SNR regime, which generalizes the result obtained in the noiseless setting. When the ratio $k = \frac{\|\theta\|}{\|\theta^*\|}$ is large enough and the angle $\frac{\pi}{2} - \varphi$ is large enough, meaning that the direction of the current iteration θ is away from that of the ground truth θ^* , the bound is dominated by η^{-2} . It also implies that the difference becomes smaller as the SNR η increases.

III. POPULATION LEVEL ANALYSIS IN NOISELESS SETTING

In this section, we provide a comprehensive analysis of population EM updates in the noiseless setting, which consists of three parts: In the first part, we derive a recurrence relation of the sub-optimality angle φ^t for population EM updates (Proposition 10), based on the EM update rule with sub-optimality angle φ^t in the noiseless setting (Corollary 8) from the

previous section (Section II). We then demonstrate that the EM update iterations follow a cycloid trajectory by formulating the parametric equation of the sub-optimality angle φ^t (Proposition 11) based on the recurrence relation.

In the second part, we show the linear growth of the sub-optimality angle φ^t (Proposition 12) when the angle between the EM iteration θ^t and the ground truth parameters θ^* is large, i.e., when the sub-optimality angle $\varphi^t = \frac{\pi}{2} - \arccos |\langle \theta^t, \theta^* \rangle| / (\|\theta^t\| \|\theta^*\|)$ is small. We then establish the quadratic convergence of the sub-optimality angle $\phi^t = 2 \arccos |\langle \theta^t, \theta^* \rangle| / (\|\theta^t\| \|\theta^*\|)$ (Proposition 13) to zero when $\varphi^t \leq 1.4$, that is, when the angle $\arccos |\langle \theta^t, \theta^* \rangle| / (\|\theta^t\| \|\theta^*\|) \leq 0.7 \approx 40^\circ$ between the EM iteration θ^t and the ground truth parameters θ^* is sufficiently small.

In the third part, we provide bounds on the accuracy of regression parameters and mixing weights in the noiseless setting, in terms of the sub-optimality angle ϕ^t (Proposition 14), based on the parametric equation of the cycloid trajectory for regression parameters (Proposition 11), and the EM update rule for mixing weights in terms of the sub-optimality angle φ^t (Corollary 8) from the previous section (Section II), respectively. Finally, we establish the convergence guarantees of the EM updates in the noiseless setting (Theorem 15) by combining the linear growth and quadratic convergence of sub-optimality angles φ^t, ϕ^t , and the bounds on the accuracy of regression parameters and mixing weights in terms of the sub-optimality angle ϕ^t . The proofs of the results in this section are provided in Appendix B.

Proposition 10 (Recurrence Relation of Sub-optimality Angle, Proposition 4.3 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, if the sub-optimality angle $\varphi^t \neq \frac{\pi}{2}$, then the recurrence relation of the sub-optimality angle φ^t for population EM updates is:*

$$\tan \varphi^{t+1} = \tan \varphi^t + \varphi^t (\tan^2 \varphi^t + 1) \quad (25)$$

where $\varphi^t := \frac{\pi}{2} - \arccos |\rho^t|$, $\rho^t := \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \|\theta^*\|}$.

Remark. Following Corollary 8 of EM update rules in the noiseless setting shown in the previous section (Section II), we can derive a recurrence relation for the sub-optimality angle φ^t in the population EM updates. In particular, when φ^t is small, the recurrence relation can be viewed as a discretized version of the differential equation $d \tan \varphi = \varphi (\tan^2 \varphi + 1) dt$, which gives a linear growth of $\tan \varphi(t) \approx \varphi(t) = C \exp(t)$ for small $\varphi(t)$ with $C \geq 0$. When φ^t becomes large, approaching $\frac{\pi}{2}$, i.e., $\varphi \rightarrow \frac{\pi}{2}$, the second term $\varphi^t (\tan^2 \varphi^t + 1)$ dominates the first term $\tan \varphi^t$, causing the growth of $\tan \varphi^t$ to exceed linear growth. In this regime, we have $\tan \varphi^{t+1} \approx \frac{\pi}{2} \tan^2 \varphi^t$, which suggests quadratic growth of $\tan \varphi^t$.

Proposition 11 (Parametric Equation for Cycloid Trajectory of EM Updates, Proposition 4.4 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, the coordinates $\mathbf{x}^t, \mathbf{y}^t$ of normalized vector $\frac{\theta^t}{\|\theta^*\|} = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2 = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2^0, \forall t \in \mathbb{Z}_+$ for population EM updates can be parameterized by the sub-optimality angle ϕ^{t-1} as follows:*

$$\begin{aligned} 1 - \operatorname{sgn}(\rho^0) \mathbf{x}^t &= \frac{1}{\pi} [\phi^{t-1} - \sin \phi^{t-1}] \\ \mathbf{y}^t &= \frac{1}{\pi} [1 - \cos \phi^{t-1}] \end{aligned} \quad (26)$$

where $\varphi^{t-1} := \frac{\pi}{2} - \arccos |\rho^{t-1}|$, $\rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \|\theta^*\|}$. Hence, the trajectory of EM iterations θ^t is on the cycloid with a parameter $\frac{\|\theta^*\|}{\pi}$, on the plane $\operatorname{span}\{\theta^0, \theta^*\}$.

Remark. By applying Corollary 8 for the EM update rule of regression parameters θ in the noiseless setting, as shown in the previous section (Section II), together with the recurrence relation of φ^t (Proposition 10), we show that EM iterations of regression parameters follow a cycloid trajectory, with a rolling radius $\frac{1}{\pi} \|\theta^*\|$. The rolling angle at the t -th EM iteration is determined by the sub-optimality angle ϕ^{t-1} from the previous $(t-1)$ -th EM iteration. Therefore, the cycloid trajectory of population EM iterations can fully characterized once the sub-optimality angles ϕ^t or $\varphi^t = \frac{\pi - \phi^t}{2}$ are fully characterized for the population EM updates.

Proposition 12 (Linear Growth of Sub-optimality Angle). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, $\tan \varphi^t$ of the sub-optimality angle φ^t grows at least linearly:*

$$\tan \varphi^{t+1} \geq 2 \cdot \tan \varphi^t \quad (27)$$

where $\varphi^t := \frac{\pi}{2} - \arccos |\rho^t|$, $\rho^t := \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \|\theta^*\|}$.

Remark. By analyzing the recurrence relation of φ^t above, we show that the linear growth guarantee of $\tan \varphi^t$ occurs in the worst case, which is consistent with our discussion in the previous remark on the recurrence relation of φ^t . In fact, by applying the trigonometric identity $\sin(2\varphi^t) = 2 \tan \varphi^t / (1 + \tan^2 \varphi^t)$, we obtain $\tan \varphi^{t+1} = (1 + \frac{2\varphi^t}{\sin(2\varphi^t)}) \tan \varphi^t \geq 2 \cdot \tan \varphi^t$. When φ^t is small enough and away from $\frac{\pi}{2}$, we have $\sin(2\varphi^t) \approx 2\varphi^t$, and thus the growth of $\tan \varphi^t$ is linear. However, as φ^t increases and approaches $\frac{\pi}{2}$, the growth of $\tan \varphi^t$ exceeds linear behavior, and we instead obtain quadratic convergence of ϕ^t to zero when $\phi^t = \pi - 2\varphi^t$ is sufficiently small. This result strengthens our previous finding on the linear growth of $\tan \varphi^t$, namely $\tan \varphi^t \geq (\frac{1+\sqrt{5}}{2}) \cdot \tan \varphi^t$ in the conference version of our paper (see Proposition 4.5 of [1]), by showing the optimal factor $2 \geq \frac{1+\sqrt{5}}{2}$ in the linear growth rate $\tan \varphi^{t+1} \geq 2 \cdot \tan \varphi^t \geq (\frac{1+\sqrt{5}}{2}) \cdot \tan \varphi^t$.

Proposition 13 (Quadratic Convergence of Sub-optimality Angle). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, the sub-optimality angle ϕ^t converges quadratically to zero when $\phi^t \leq 1.4$ is small enough:*

$$\frac{\phi^{t+1}}{\pi} \leq \left[\frac{\phi^t}{\pi} \right]^2 \quad (28)$$

where $\phi^t := 2 \arccos |\rho^t|$, $\rho^t := \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \|\theta^*\|}$.

Remark. When ϕ^t is small enough, namely $\varphi^t = \frac{\pi - \phi^t}{2}$ is close to $\pi/2$, we have $\tan \varphi^t = \cot \frac{\phi^t}{2} \approx \frac{2}{\phi^t}$, and the term $\tan^2 \varphi^t$ dominates the term φ^t in the recurrence relation. Therefore, we have $2/\phi^{t+1} \approx \tan \varphi^{t+1} \approx \frac{\pi}{2} \tan^2 \varphi^t \approx \frac{\pi}{2} (2/\phi^t)^2 = 2\pi/[\phi^t]^2$. Through a rigorous derivation, we can show that the above inequality $\phi^{t+1}/\pi \leq [\phi^t/\pi]^2$ holds for sufficiently small $\phi^t \leq 1.4$, which guarantees the quadratic convergence of ϕ^t to zero. While previous work (see page 5, Lemma 1 in [22]) established the quadratic convergence rate when $\|\theta - \text{sgn}(\rho)\theta^*\|/\|\theta^*\| \leq 1/10$ under high SNR, our analysis extends the region of quadratic convergence from $1/10$ to $\frac{1}{\pi} \sqrt{(\phi - \sin \phi)^2 + (1 - \cos \phi)^2} \big|_{\phi=1.4} \approx 0.30$.

Proposition 14 (Accuracy of Population EM Updates and Sub-optimality Angle). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$,*

$$\begin{aligned} \frac{\|\theta^t - \text{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} &\leq \frac{[\phi^{t-1}]^2}{2\pi} \\ \|\pi^t - \bar{\pi}^*\|_1 &= \frac{\phi^{t-1}}{\pi} \cdot \left\| \pi^* - \frac{\mathbb{1}}{2} \right\|_1 \end{aligned} \quad (29)$$

where $\phi^{t-1} := 2(\frac{\pi}{2} - \arccos |\rho^{t-1}|)$, $\rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \|\theta^*\|}$, and $\bar{\pi}^* := \frac{1}{2} + \text{sgn}(\rho^0)(\pi^* - \frac{1}{2})$, $\mathbb{1} := (1, 1)$.

Remark. The first inequality provides a bound on the accuracy of regression parameters θ^t with respect to the ground truth θ^* , which depends only on the sub-optimality angle ϕ^{t-1} from the previous $(t-1)$ -th EM iteration. The second identity quantifies the accuracy of mixing weights π^t with respect to the ground truth π^* , which depends only on the sub-optimality angle ϕ^{t-1} and the imbalance of ground truth mixing weights from the previous $(t-1)$ -th EM iteration. These two relations suggest that we can establish the convergence guarantee of the population EM updates once we establish the convergence guarantee of the sub-optimality angle ϕ^t .

Theorem 15 (Population Level Convergence, Theorem 4.1 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, if the initial sup-optimality angle cosine $\rho^0 := \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \|\theta^*\|} \neq 0$, then with the number of total iterations at most $T = \mathcal{O}(\log \frac{1}{|\rho^0|} \vee \log \log \frac{1}{\varepsilon})$, the error of EM update at the population level is bounded by: $\frac{\|\theta^{T+1} - \text{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} < \varepsilon$ and $\|\pi^{T+1} - \bar{\pi}^*\|_1 = \|\pi^* - \frac{1}{2}\|_1 \mathcal{O}(\sqrt{\varepsilon})$. where $\bar{\pi}^* := \frac{1}{2} + \text{sgn}(\rho^0)(\pi^* - \frac{1}{2})$, $\mathbb{1} := (1, 1)$.*

Remark. If the initial sub-optimality angle $\varphi^0 \geq 1$, then $\phi^t \leq \pi - 2\varphi^0 < 1.4$ converges quadratically to zero (thus $T_1 = 0$). Otherwise, when the initial sub-optimality angle $\varphi^0 < 1$, it takes at most $T_1 = \mathcal{O}(\log \frac{1}{\tan \varphi^0}) = \mathcal{O}(\log \frac{1}{|\rho^0|})$ EM iterations to ensure quadratic convergence of $\phi^t \leq 1.4$ for all $t \geq T_1$. Subsequently, the population EM updates require at most another $T_2 = \mathcal{O}(\log \log \frac{1}{\varepsilon})$ EM iterations to ensure that the sub-optimality angle $\phi^t < \sqrt{2\pi\varepsilon}$ is sufficiently small. Therefore, using the above expressions for the errors of regression parameters and mixing weights, we establish the convergence guarantee of the population EM updates with a total number of iterations at most $T = T_1 + T_2 = \mathcal{O}(\log \frac{1}{|\rho^0|} \vee \log \log \frac{1}{\varepsilon})$.

IV. FINITE-SAMPLE LEVEL ANALYSIS IN NOISELESS SETTING

In this section, we provide a comprehensive analysis of finite-sample EM updates in the noiseless setting, which consists of two parts: In the first part, we bound the statistical errors of regression parameters and mixing weights in the finite-sample setting (Propositions 16, 17 and 18), showing that the statistical error of regression parameters is on the order of $(d/n)^{1/2}$, while the statistical error of mixing weights depends on the sub-optimality angle ϕ and the ground truth mixing weights π^* . Furthermore, we bound the statistical accuracy of EM updates for regression parameters and mixing weights (Proposition 19), showing that the statistical accuracy of EM updates for regression parameters is on the order of $\phi^2 \vee \phi^{\frac{3}{2}} \sqrt{d/n} \vee \phi \sqrt{\log \frac{1}{\varepsilon}/n}$ and the statistical accuracy of EM updates for mixing weights depends on the sub-optimality angle ϕ and the ground truth mixing weights π^* .

In the second part, we establish the convergence guarantees of the sub-optimality angle ϕ^t (Proposition 21) by leveraging the linear growth and quadratic convergence of the sub-optimality angles (Propositions 12 and 13) at the population level from the previous section (Section III) and combining them with the bounds on the statistical errors and the statistical accuracy of regression parameters and mixing weights in the finite-sample setting. Finally, we establish the convergence guarantees of the EM updates in the finite-sample setting (Theorem 22), by combining the convergence guarantees of the sub-optimality angle ϕ^t (Proposition 21) and the connection between the sub-optimality angle ϕ^t and the statistical accuracy of regression parameters and mixing weights (Proposition 19) obtained in the first part. Moreover, the initialization of unknown mixing weights and

regression parameters can be arbitrary, enabled by Easy EM followed by the standard EM update rules (Proposition 20). Therefore, the convergence guarantees for the finite-sample EM updates of the regression parameters and mixing weights are established and characterized. The proofs of the results in this section are provided in Appendix C.

Proposition 16 (Projected Error of Easy EM Update for Regression Parameters, Proposition 5.2 in [1]). *In the noiseless setting, the projection on $\text{span}\{\theta, \theta^*\}$ for the statistical error of θ satisfies*

$$\frac{\|P_{\theta, \theta^*}[M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)]\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n}\right), \quad (30)$$

with probability at least $1 - \delta$, where $M_n(\theta, \nu), M(\theta, \nu)$ are the EM update rules for θ at the Finite-sample level and the population level respectively, and the orthogonal projection matrix P_{θ, θ^*} satisfies $\text{span}(P_{\theta, \theta^*}) = \text{span}\{\theta, \theta^*\}$.

Remark. We can show that the projected error of the EM update rule for regression parameters is the average of n i.i.d. sub-exponential two-dimensional random vectors. Therefore, by Bernstein's inequality for sub-exponential random variables (see Corollary 2.8.3 on page 38 of [52]), the bound for the projection error of the EM update rule is obtained as shown above.

Proposition 17 (Statistical Error of EM Update for Regression Parameters, Proposition 5.3 in [1]). *In the noiseless setting, the statistical error of θ for finite-sample EM updates with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples satisfies*

$$\frac{\|M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}\right), \quad \frac{\|M_n(\theta, \nu) - M(\theta, \nu)\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}\right), \quad (31)$$

with probability at least $1 - \delta$, $M_n(\theta, \nu), M(\theta, \nu)$ denote the EM update rules for θ at the Finite-sample level and the Population level.

Remark. For the statistical error of the Easy EM update rule for regression parameters, we obtain a tight bound by decomposing the error into two parts: the projected error and the error orthogonal to the space $\text{span}\{\theta, \theta^*\}$, the projected error is already bounded in Proposition 16, while the orthogonal error is bounded by applying the rotational invariance of Gaussians and rewriting the ℓ_2 norm of the error as the geometric mean of two Chi-square random variables, and leveraging the concentration inequality for Chi-square distribution (see Lemma 1, page 1325 in [53]). For the statistical error of the standard EM update rule for regression parameters, we obtain similar bounds by separately bounding the decomposed errors (see the detailed derivation in Appendix C). Our analysis yields a tighter bound $\mathcal{O}(\sqrt{(d \vee \log \frac{1}{\delta})/n})$ for the statistical error compared to the previously established bound $\mathcal{O}(\sqrt{d/n} \log \frac{n}{\delta})$, which includes a multiplicative logarithmic factor $\log \frac{n}{\delta}$ arising from the standard symmetrization technique and the Ledoux-Talagrand contraction argument (Appendix E, Lemma 11, page 17 in [22] and [19]).

Proposition 18 (Statistical Error of EM Update for Mixing Weights). *In the noiseless setting, the statistical error of mixing weights for finite-sample EM updates satisfies*

$$|N_n(\theta, \nu) - N(\theta, \nu)| = \mathcal{O}\left(\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right) \quad (32)$$

with probability at least $1 - \delta$, where $N_n(\theta, \nu), N(\theta, \nu)$ denote the EM update rules for imbalance $\tanh \nu$ of the mixing weights at the Finite-sample level and the Population level, and $p := \|\pi^* - \frac{1}{2}\|_1 \frac{\phi}{2\pi} + \min(\pi^*(1), \pi^*(2))$, $\phi = 2 \arccos |\rho|$, $\rho = \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$.

Remark. We bound the statistical error of mixing weights by rewriting it as an average of n i.i.d. Bernoulli random variables with the success probability $p := \|\pi^* - \frac{1}{2}\|_1 \frac{\phi}{2\pi} + \min(\pi^*(1), \pi^*(2))$. By applying the concentration inequality for Bernoulli r.v.'s (see Appendix E), we obtain the bound for the statistical error of mixing weights as shown above. In particular, when $\pi^* \rightarrow (1, 0)$ or $\pi^* \rightarrow (0, 1)$ and the sub-optimality angle $\phi \rightarrow 0$, then the probability $p = \|\pi^* - \frac{1}{2}\|_1 \frac{\phi}{2\pi} + \min(\pi^*(1), \pi^*(2)) \rightarrow 0$, therefore the statistical error of mixing weights $|N_n(\theta, \nu) - N(\theta, \nu)|$ also approaches zero. In general, for fixed ground truth mixing weights π^* , the statistical error of mixing weights $|N_n(\theta, \nu) - N(\theta, \nu)|$ decreases as the sub-optimality angle ϕ decreases.

Proposition 19 (Statistical Accuracy of EM Updates for Regression Parameters and Mixing Weights). *In the noiseless setting, the finite-sample EM with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples achieves the statistical accuracy of regression parameters and mixing weights:*

$$\begin{aligned} \frac{\|M_n(\theta, \nu) - \text{sgn}(\rho)\theta^*\|}{\|\theta^*\|} &= \mathcal{O}\left(\phi^2 \vee \phi^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right) \\ |N_n(\theta, \nu) - \text{sgn}(\rho) \tanh \nu^*| &= \mathcal{O}\left(\phi \left\|\pi^* - \frac{1}{2}\right\|_1 \vee \left[\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right]\right) \end{aligned} \quad (33)$$

with probability at least $1 - \delta$, where $M_n(\theta, \nu)$ denotes the EM update rule for regression parameters at the Finite-sample level, and $\phi := 2 \arccos |\rho|$, $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$ and $p := \frac{\phi}{2\pi} \|\pi^* - \frac{1}{2}\|_1 + \min(\pi^*(1), \pi^*(2))$.

Remark. The above results show that the statistical accuracy of regression parameters satisfies $\phi^{t+1} \leq \pi \sin \frac{\phi^{t+1}}{2} \leq \pi \|\theta^{t+1} - \text{sgn}(\rho^{t+1})\theta^*\|/\|\theta^*\| \lesssim [\phi^t]^2 \vee [\phi^t]^{\frac{3}{2}} \sqrt{d/n} \vee \phi^t \sqrt{\log \frac{1}{\delta}/n}$. This implies the quadratic convergence rate for sub-optimality angle ϕ^t and therefore the quadratic convergence rate for the accuracy of normalized regression parameters $\|\theta^t - \text{sgn}(\rho^0)\theta^*\|/\|\theta^*\|$, when the sub-optimality angle ϕ^t is large enough, i.e., $\phi^t \gtrsim d/n \vee \sqrt{\log \frac{1}{\delta}/n}$. This aligns with our theoretical finding in Proposition 13 at the population level with infinite samples $n \rightarrow \infty$. If the sub-optimality angle is very small such that $\phi^t \lesssim \log \frac{1}{\delta}/n$, then the term $\phi^t \sqrt{\log \frac{1}{\delta}/n}$ dominates the other terms, and the sub-optimality angle ϕ^t (and therefore the regression parameter accuracy $\|\theta^t - \text{sgn}(\rho^0)\theta^*\|/\|\theta^*\|$) converges to zero linearly at a rate of $\sqrt{\log \frac{1}{\delta}/n}$. By the above analysis, and noting that the normalized regression parameter accuracy $\|\theta^t - \text{sgn}(\rho^0)\theta^*\|/\|\theta^*\| \asymp \phi^t$, we establish the quadratic convergence rate for the regression parameter accuracy $\|\theta^t - \text{sgn}(\rho^0)\theta^*\|/\|\theta^*\|$ when the normalized regression parameter accuracy is large enough in the noiseless setting without the requirements on the mixing weights. While the previous work [44] showed the quadratic convergence rate for the accuracy of normalized regression parameters when the accuracy $\gtrsim \min(\pi^*(1), \pi^*(2))$ for a variant of EM in the noiseless setting, our results remove restrictions on mixing weights and still establish quadratic convergence rate.

Proposition 20 (Initialization with Easy EM, Proposition 5.4 in [1]). *In the noiseless setting, suppose we run the sample-splitting finite-sample Easy EM with $n' \asymp \frac{n}{\log \frac{1}{\delta}}$ fresh samples for each iteration, then after at most $T_0 = \mathcal{O}(\log \frac{1}{\delta})$ iterations, it satisfies $\varphi^{T_0} \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$ with probability at least $1 - \delta$.*

Remark. The above result guarantees that, when running the EM algorithm, arbitrary initialization can still satisfy the condition for angle convergence $\varphi^0 \gtrsim \sqrt{\log \frac{1}{\delta}/n}$ in the next stage, after only a few iterations of Easy EM with high probability. Our theoretical result avoids both using the spectral method for initialization and requiring the comparison between the sample size n and $d^2 \log^2 \frac{n}{\delta}$ as in previous works (see page 6526 of [51] and page 3518 of [17]).

Proposition 21 (Convergence of Sub-optimality Angle). *In the noiseless setting, suppose $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, given a positive number $\varepsilon \lesssim \frac{\log \frac{1}{\delta}}{n}$, we run Easy finite-sample EM for $T_1 = \mathcal{O}\left(\log \frac{d}{\log \frac{1}{\delta}}\right)$ iterations followed by the standard finite-sample EM for at most $T' = \mathcal{O}\left(\left[\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}\right] \vee \log\left[\log \frac{n}{\log \frac{1}{\delta}} / \log \frac{n}{d}\right] \vee \left[\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}}\right]\right)$ iterations with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples, then it satisfies*

$$\phi^T \leq \varepsilon, \quad (34)$$

with probability at least $1 - T\delta$, where $T := T_1 + T'$, $\varphi^0 := \frac{\pi}{2} - \arccos \left| \frac{\langle \theta^0, \theta^ \rangle}{\|\theta^0\| \cdot \|\theta^*\|} \right|$ and $\varphi^T := \frac{\pi}{2} - \arccos \left| \frac{\langle \theta^T, \theta^* \rangle}{\|\theta^T\| \cdot \|\theta^*\|} \right|$.*

Remark. Instead of analyzing the convergence of regression parameters and mixing weights at the finite-sample level directly, we first establish the convergence of the sub-optimality angle ϕ^t (see detailed proof in Appendix C), then later establish the convergence of EM update rules by employing the convergence of the sub-optimality angle ϕ^t together with the bounds on the errors of regression parameters and mixing weights in the finite-sample setting. The analysis for the convergence of the sub-optimality angle ϕ^t proceeds in four stages.

In the first stage, Easy EM is run for $T_1 = \mathcal{O}(\log \frac{\text{statistical error of } \theta}{\text{projected error of } \theta}) = \mathcal{O}(\log(d/\log \frac{1}{\delta}))$ iterations to ensure that the sub-optimality angle φ^t grows from $\varphi^0 \gtrsim \sqrt{\log \frac{1}{\delta}/n}$ to $\varphi^{T_1} \gtrsim \sqrt{(d \vee \log \frac{1}{\delta})/n}$, which matches the statistical error of normalized regression parameters. In this stage, the sub-optimality angle φ^t of Easy EM exhibits a linear growth rate, with an error on the order of $\sqrt{\log \frac{1}{\delta}/n}$ which matches the bound on the projected error of the normalized regression parameters, as indicated by Proposition 12 at the population level and Proposition 16 at the finite-sample level.

In the second stage, standard EM is run for $T_2 = \mathcal{O}(\log(1/\text{statistical error of } \theta)) = \mathcal{O}(\log(n/d) \wedge \log(n/\log \frac{1}{\delta}))$ iterations, continuing the linear growth from $\varphi^{T_1} \gtrsim \sqrt{(d \vee \log \frac{1}{\delta})/n}$ until $\varphi^{T_1+T_2} \geq 1$ (ensuring $\phi^{T_1+T_2} \leq \pi - 2\varphi^{T_1+T_2} < 1.4$). In this stage, the linear growth of the sub-optimality angle φ^t of standard EM is guaranteed with an error on the order of $\sqrt{(d \vee \log \frac{1}{\delta})/n}$, which matches the statistical error of the normalized regression parameters, by leveraging Proposition 12 at the population level and Proposition 17 at the finite-sample level.

In the third stage, standard EM is run for $T_3 = \mathcal{O}(\log[\log(n/d) \wedge \log(n/\log \frac{1}{\delta})]) = \mathcal{O}(\log \log(1/\text{statistical error of } \theta))$ iterations, exhibiting a quadratic convergence rate of the sub-optimality angle ϕ^t from $\phi^{T_1+T_2} < 1.4$ to $\phi^{T_1+T_2+T_3} \lesssim \sqrt{(d \vee \log \frac{1}{\delta})/n}$. In this stage, the quadratic convergence of the sub-optimality angle ϕ^t of standard EM is shown with an error on the order of $\sqrt{(d \vee \log \frac{1}{\delta})/n}$, which matches the statistical error of the normalized regression parameters, by applying Proposition 21 at the population level and Proposition 17 at the finite-sample level.

In the fourth stage, standard EM is run for $T_4 = \mathcal{O}(\log[\log(n/\log \frac{1}{\delta})/\log(n/d)] \vee [\log(1/\varepsilon)/\log(n/\log \frac{1}{\delta})])$ iterations to achieve a final sub-optimality angle $\phi^{T_1+T_2+T_3+T_4} \leq \varepsilon$ (for a positive number $\varepsilon \lesssim \log \frac{1}{\delta}/n$). This stage begins with $\phi^{T_1+T_2+T_3+1} \lesssim \sqrt{(d \vee \log \frac{1}{\delta})/n} \phi^{T_1+T_2+T_3} \lesssim (d \vee \log \frac{1}{\delta})/n$, after which the convergence rate becomes conditional. The sub-optimality angle ϕ^t exhibits super-linear convergence of order 3/2 if $n \lesssim d^2/\log^2 \frac{1}{\delta}$ and $\phi^t \gtrsim \log \frac{1}{\delta}/d$; otherwise, it demonstrates a linear convergence with $\phi^{t+1} \lesssim \sqrt{\log \frac{1}{\delta}/n} \phi^t$, by applying Proposition 19 at the finite-sample level.

Therefore, this four-stage analysis (Easy EM for T_1 iterations, standard EM for $T' = T_2 + T_3 + T_4$ iterations) establishes the convergence of the sub-optimality angle ϕ^t . Our finer analysis reduces T_1 from $\mathcal{O}(\log(n/\log \frac{1}{\delta}))$ to $\mathcal{O}(\log(d/\log \frac{1}{\delta}))$.

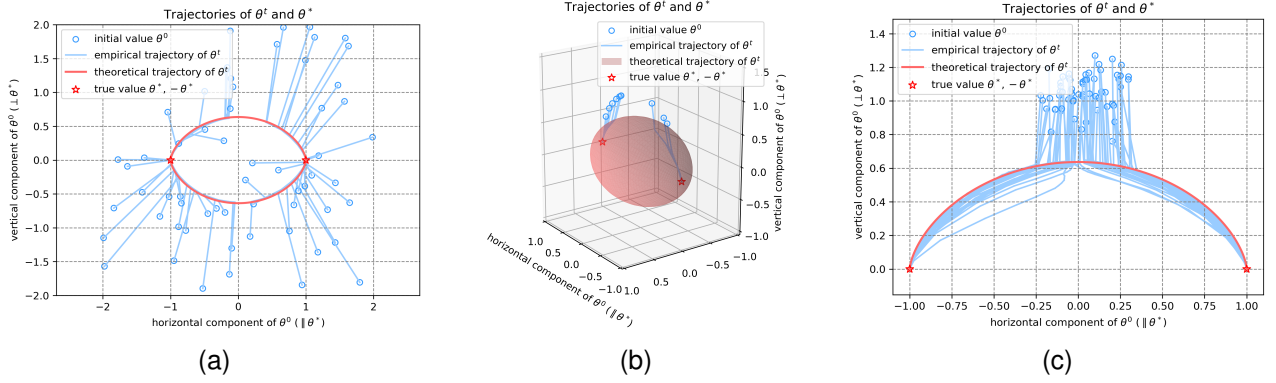


Fig. 2: Cycloid trajectories of EM iterations for regression parameters θ^t : we run 100 iterations of Finite-sample EM at $\text{SNR}=10^8$ for varying dimensions ($d = 2, 3, 50$). (a) $d = 2$, trajectories of θ^t across 60 trials with $\theta^* = (1, 0)$, $\pi^* = (0.7, 0.3)$; initial values θ^0 and π^0 are uniformly sampled from $[-2, 2]^2$ and $[0, 1]$, respectively. (b) $d = 3$, trajectories of θ^t across 10 trials, where θ^*, θ^0 are sampled from three-dimensional unit sphere, and π^*, π^0 are drawn uniformly from $[0, 1]$. (c) $d = 50$, trajectories of θ^t across 60 trials, with θ^*, θ^0 sampled from $\mathcal{N}(0, I_d)$, and π^*, π^0 uniformly drawn from $[0, 1]$.

Furthermore, by establishing Proposition 19 and applying it in a new fourth stage (both absent in our conference version [1]), we show that the final accuracy of sub-optimality angle ϕ^t can be arbitrary small (up to ε), thereby strengthening Proposition 5.5 in the conference version of our paper [1].

Theorem 22 (Finite-Sample Level Convergence). *In the noiseless setting, suppose any initial mixing weights π^0 and any initial regression parameters $\theta^0 \in \mathbb{R}^d$ ensuring that $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$. Given a positive number $\varepsilon \lesssim \frac{\log \frac{1}{\delta}}{n}$, we run finite-sample Easy EM for at most $T_1 = \mathcal{O}\left(\log \frac{d}{\log \frac{1}{\delta}}\right)$ iterations followed by the finite-sample standard EM for at most $T' = \mathcal{O}\left(\left[\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}\right] \vee \log\left[\log \frac{n}{\log \frac{1}{\delta}} / \log \frac{n}{d}\right] \vee \left[\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}}\right]\right)$ iterations with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples, then*

$$\frac{\|\theta^{T+1} - \text{sgn}(\rho^{T+1})\theta^*\|}{\|\theta^*\|} = \mathcal{O}\left(\varepsilon \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right), \quad \|\pi^{T+1} - \bar{\pi}^*\|_1 = \mathcal{O}\left(\varepsilon \left\|\frac{1}{2} - \pi^*\right\|_1 \vee \left[\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p(\varepsilon, \pi^*)}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right]\right), \quad (35)$$

with probability at least $1 - T\delta$, where $T := T_1 + T'$, $\varphi^0 := \frac{\pi}{2} - \arccos\left|\frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|}\right|$, $\rho^{T+1} := \frac{\langle \theta^{T+1}, \theta^* \rangle}{\|\theta^{T+1}\| \cdot \|\theta^*\|}$, $\bar{\pi}^* := \frac{1}{2} + \text{sgn}(\rho^{T+1})(\pi^* - \frac{1}{2})$, and $p(\varepsilon, \pi^*) := \varepsilon \left\|\pi^* - \frac{1}{2}\right\|_1 + \min(\pi^*(1), \pi^*(2))$.

Remark. Based on Proposition 21, which establishes the convergence of sub-optimality angle ϕ^t such that $\phi^t \leq \varepsilon \lesssim \log \frac{1}{\delta}/n$, and Proposition 19, which provides the statistical accuracy of the EM updates for regression parameters and mixing weights, we obtain the following bound on the accuracy of normalized regression parameters: $\|\theta^{T+1} - \text{sgn}(\rho^{T+1})\theta^*\|/\|\theta^*\| \lesssim \varepsilon^2 + \varepsilon^{3/2}\sqrt{d/n} + \varepsilon\sqrt{\log \frac{1}{\delta}/n} \asymp \varepsilon\sqrt{\log \frac{1}{\delta}/n}$. Consequently, the finite-sample EM update $M_n(\theta, \nu)$ for regression parameters achieves exact recovery as $\varepsilon \rightarrow 0$, and acts as a variant of Ordinary Least Squares (OLS) (see Chapter 3 of [54]) which also guarantees exact recovery in the noiseless setting. Similarly, from Proposition 19 for the statistical accuracy of EM updates, noting that $\frac{\phi^T}{2\pi} \|\pi^* - \frac{1}{2}\|_1 + \min(\pi^*(1), \pi^*(2)) \lesssim \varepsilon \|\pi^* - \frac{1}{2}\|_1 + \min(\pi^*(1), \pi^*(2)) \equiv p(\varepsilon, \pi^*)$, the bound on the accuracy of mixing weights is obtained as shown above.

When the ground truth mixing weights π^* are close to $(1, 0)$ or $(0, 1)$, we have $p(\varepsilon, \pi^*) \rightarrow 0$ as $\varepsilon \rightarrow 0$, implying that the final accuracy of mixing weights $\|\pi^{T+1} - \bar{\pi}^*\|_1 \rightarrow 0$ as $\varepsilon \rightarrow 0$. This means that, after running EM for a sufficient number of iterations, the estimation error of EM estimates for mixing weights and regression parameters can be arbitrary small (up to ε) in the noiseless setting, when the ground truth mixing weights π^* are either $(1, 0)$ or $(0, 1)$. In this case, 2MLR degenerates to a linear regression model, and EM update rules also guarantee exact recovery of mixing weights in the noiseless setting.

When the ground truth mixing weights $\pi^* = (\frac{1}{2}, \frac{1}{2})$ are balanced, we have $p(\varepsilon, \pi^*) \rightarrow \varepsilon \left\|\pi^* - \frac{1}{2}\right\|_1 + \min(\pi^*(1), \pi^*(2)) \asymp \varepsilon$, hence the final accuracy of mixing weights is on the order of $\mathcal{O}(\sqrt{\log \frac{1}{\delta}/n})$, which is independent of the dimension d of the regression parameters. This result strengthens Theorem 5.1 in the conference version of our paper [1], as it demonstrates the exact recovery of regression parameters and characterizes the dependence of the final accuracy of mixing weights on the ground truth mixing weights π^* through the newly established bound on the statistical error of mixing weights in Proposition 18.

V. EXPERIMENTS

In this section of numerical experiments, we confirm the theoretical results established in the previous sections. A total of 5,000 independently and identically distributed (i.i.d.) d -dimensional covariates, denoted by $\{x_i\}_{i=1}^n$, are sampled from a normal distribution $\mathcal{N}(0, I_d)$. The ground truth parameters θ^* are selected at random from a d -dimensional unit sphere.

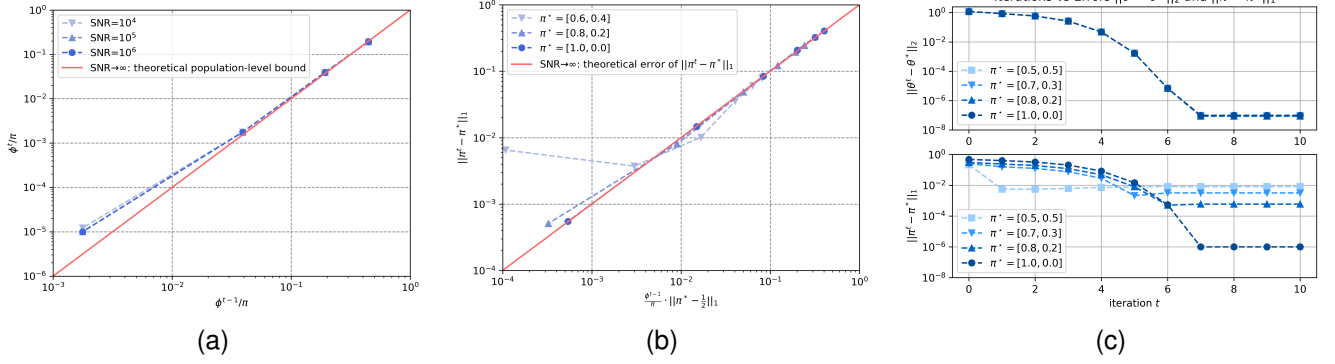


Fig. 3: Left and middle panels illustrate the quadratic convergence of the sub-optimality angle ϕ^t and its correlation with the mixing-weight error. Both use θ^* and θ^0 sampled from the $d = 50$ unit sphere, with $\phi^0 = 1.4$ (equivalently, $\varphi^0 = (\pi - 1.4)/2$) in (a) and $\varphi^0 = 0.3$ in (b). The right panel (c) shows the accuracy of the EM estimates for the regression parameters and mixing weights over ten EM iterations with $d = 50$, $\varphi^0 = 0.3$ at $\text{SNR} = 10^6$, and varying true mixing weights $\pi^* = (0.5, 0.5), (0.6, 0.4), (0.8, 0.2)$, and $(1 - 10^{-6}, 10^{-6})$.

(a) Quadratic convergence of the sub-optimality angle ϕ^t with all EM iterations starting with $\phi^0 = 1.4$ and $\pi^*(1), \pi^0(1)$ drawn uniformly from $[0, 1]$.

(b) Correlation between the mixing-weight error $\|\pi^t - \pi^*\|_1$ and the preceding sub-optimality angle $\phi^{t-1} = \arccos | \langle \theta^{t-1}, \theta^* \rangle | / (\|\theta^{t-1}\| \|\theta^*\|)$.

(c) Accuracy of the EM estimates for the regression parameters and mixing weights versus iteration for different ground-truth mixing weights.

Then, the ground truth mixture weights π^* for these two components are specified manually or randomly. We use π^* to generate latent variable samples $\{z_i\}_{i=1}^n$ from a categorical distribution $\mathcal{CAT}(\pi^*)$. Gaussian noise ε_i is then added to the linear regression $(-1)^{z_i+1} \langle \theta^*, x_i \rangle$ indicated by these latent variables, resulting in output response samples $\{y_i\}_{i=1}^n$. In all experiments, the full dataset is used for EM updates at each iteration. Every point in the plots of Fig. 3 represents an average over 50 runs with different initialization values for EM updates. The code for empirical experiments is available at https://github.com/dassein/cycloid_em_tit.

Cycloid Trajectory of Regression Parameters. At the population level, we demonstrate that the output of the t -th iteration lies on the cycloid in the space $\text{span}\{\theta^{t-1}, \theta^*\}$ in the noiseless setting. For the corresponding experiments, we set the signal-to-noise ratio (SNR) to 10^8 and examine different dimensions d (2, 3, and 50). As shown in Fig. 2, all iterations remain close to the theoretical cycloid. Therefore, the empirical results confirm our theoretical findings in Proposition 11.

Super-linear Convergence of Sub-Optimality Angle. We demonstrate the quadratic convergence of ϕ^t/π in Fig. 3(a) in high SNR regime, when ϕ^t is large enough. The experiments are conducted in the dimension ($d=50$), considering different high SNR values ($10^4, 10^5, 10^6$). The initial values of the first component of mixing weights and the regression parameters are drawn uniformly from an interval $[0, 1]$ and a unit sphere, respectively. Each point corresponding to one of 4 EM iterations in Fig. 3(a) represents an average over 50 independent runs with different initializations. The slope of the plot shows the convergence rate exponent, and the slopes observed at various SNR levels consistently remain close to 2 when ϕ^t is large enough. That observation is consistent with our theoretical result establishing a quadratic convergence rate in Proposition 13, and aligns with our analysis in the remark of Proposition 19.

Accuracy of EM Estimate for Mixing Weights and Sub-optimality Angle. In the noiseless setting, we show that the accuracy of EM estimate for mixing weights $\|\pi^t - \pi^*\|_1$ is proportional to the sub-optimality angle ϕ^{t-1} at population level in Proposition 14. We illustrate the linear correlation between the accuracy of EM estimate for mixing weights and the sub-optimality angle in Fig. 3(b). For the experimental setup, we set the dimension at $d = 50$, and examine different choices of ground truth mixing weights $\pi^* = (0.6, 0.4), (0.8, 0.2)$, and $(1, 0)$. It is observed that the accuracy of EM estimate for mixing weights at t -th iteration is exactly characterized by the sub-optimality angle ϕ^{t-1} from the preceding iteration. Therefore, our empirical results confirm Proposition 14, which is consistent with our analysis in the remark of Proposition 19.

Finite-Sample Level Convergence with Different Mixing Weights. In the noiseless setting, Corollary 8 and Proposition 10 establish that the EM update for regression parameters θ^t is unaffected by the ground truth mixing weights π^* . The first subplot of Fig. 3(c) shows that, at a high SNR (10^6), the accuracy of normalized regression parameters (measured in ℓ_2 norm) remains nearly unchanged across different choices of ground truth mixing weights $\pi^* = (0.5, 0.5), (0.7, 0.3), (0.8, 0.2)$, and $(1 - 10^{-6}, 10^{-6})$, thereby confirming our theoretical analysis in Proposition 14.

Theorem 22 further proves that the final accuracy of EM estimate for mixing weights (in ℓ_1 norm) depends on both the accuracy of normalized regression parameters and the ground truth mixing weights. In particular, when the accuracy of normalized regression parameters is small, the closer the ground truth mixing weights are to $(1, 0)$ or $(0, 1)$, the smaller the final accuracy of EM estimate for mixing weights. To verify this result, we investigate statistical accuracy of EM estimate for regression parameters θ (measured in ℓ_2 norm) and mixing weights π (measured in ℓ_1 norm) under different ground truth mixing weights $\pi^* = (0.5, 0.5), (0.7, 0.3), (0.8, 0.2)$, and $(1 - 10^{-6}, 10^{-6})$. The second subplot of Fig. 3(c) depicts the dependence of the statistical accuracy of EM estimate for mixing weights on π^* , providing further support for Theorem 22. In our experiments, θ^* and θ^0 are drawn from the 50-dimensional unit sphere with initial sub-optimality angle $\varphi^0 = 0.3$.

VI. CONCLUSION

We derive explicit expressions for the EM updates in the two-component Mixed Linear Regression (2MLR) model with unknown mixing weights and regression parameters across all SNR regimes. We then characterize the properties of EM updates based on the explicit expressions, establishing their structural behavior and boundedness, and showing that in the noiseless setting, they follow a cycloid trajectory derived via a recurrence relation for the sub-optimality angle. In finite high-SNR regimes, we further bound the deviation of the EM updates from this cycloid trajectory. At the population level, the trajectory-based analysis reveals the order of convergence: linear convergence when the EM estimate is nearly orthogonal to the ground truth regression parameters, and quadratic convergence when the angle between the estimate and the ground truth is small. Furthermore, our work provides a novel trajectory-based framework that establishes non-asymptotic guarantees by tightening bounds of the statistical errors between the finite-sample and population EM updates, revealing the connection between EM's statistical accuracy and the sub-optimality angle, and establishing convergence guarantees with arbitrary initialization at the finite-sample level.

CONTENTS

I	Introduction	1
I-A	Motivation	2
I-B	Problem Setup and EM Updates	2
I-C	Related Works	3
I-D	Main Contributions	4
I-E	Notations	4
I-F	Organization	5
II	Explicit EM Update Expressions and Properties of EM Update Rules	5
III	Population Level Analysis in Noiseless Setting	8
IV	Finite-Sample Level Analysis in Noiseless Setting	10
V	Experiments	13
VI	Conclusion	15
	Appendix A: Explicit EM Update Expressions and Properties of EM Update Rules	17
A-A	Connection between EM Update Rules and Gradient Descent of Negative Log-Likelihood	17
A-B	Explicit EM Update Expressions	18
A-C	Bound of the EM Update Rule in General SNR Regime	19
A-D	Analysis for the Fixed Points of EM Updates in General SNR Regime	19
A-E	EM Update Rule in the Noiseless Setting	23
A-F	Deviation from Cycloid Limit of EM Updates in High SNR Regime	23
	Appendix B: Population Level Analysis in Noiseless Regime	25
B-A	Recurrence Relation and Cycloid Trajectory of EM Iterations in Noiseless Setting	25
B-B	Linear Growth and Quadratic Convergence of Sub-optimality Angles in Noiseless Setting	27
B-C	Accuracy of Regression Parameters and Mixing Weights in Noiseless Setting	27
B-D	Convergence Guarantee of EM Updates at the Population Level in Noiseless Setting	28
	Appendix C: Finite-Sample Level Analysis in Noiseless Setting	29
C-A	Statistical Error, Statistical Accuracy of EM Updates in Noiseless Setting	29
C-B	Initialization of Easy EM Updates and Perturbation in Sub-Optimality Angle	33
C-C	Convergence Guarantees for Finite-Sample EM Updates in Noiseless Setting	33
	Appendix D: Derivations for EM Update Rules	37
D-A	Negative Log-Likelihood, Surrogate Function, and Q Function	37
D-B	Negative Log-Likelihood, Surrogate Function for General MLR and 2MLR	38
D-C	Derivations of EM Update Rules for 2MLR	41
	Appendix E: Helper Lemmas Used in the Proofs of Results	42
E-A	Asymptotic Analysis of Integrals	42
E-B	Identities and Inequalities for Useful Expectations	45
E-C	Limit Behavior and Non-Increasing Property of Some Helper Functions	46
E-D	Probability Concentration Inequality for Bernoulli Random Variables	47
	References	49
	Biographies	50
	Zhankun Luo	50
	Abolfazl Hashemi	50

APPENDIX A

EXPLICIT EM UPDATE EXPRESSIONS AND PROPERTIES OF EM UPDATE RULES

A. Connection between EM Update Rules and Gradient Descent of Negative Log-Likelihood

Theorem A.1 (Proposition 1: Population and Finite-Sample Negative Log-Likelihood). *Let $f(\theta, \pi) := -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} [\ln p(s|\theta, \pi)]$ be the negative log-likelihood function at the population level, and $f_n(\theta, \pi) := -\frac{1}{n} \sum_{i=1}^n \ln p(s_i|\theta, \pi)$ be the negative log-likelihood function at the finite-sample level for the dataset $\mathcal{S} = \{s_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples. Then, $f(\theta, \pi)$ and $f_n(\theta, \pi)$ can be expressed as:*

$$\begin{aligned} f(\theta, \pi) &= \frac{1}{2\sigma^2} \langle \theta, \mathbb{E}[xx^\top] \cdot \theta \rangle - \mathbb{E} \left[\ln \frac{\cosh(y \langle x, \theta \rangle / \sigma^2 + \nu)}{\cosh \nu} \right] + \mathbf{C} \\ f_n(\theta, \pi) &= \frac{1}{2\sigma^2} \left\langle \theta, \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \cdot \theta \right\rangle - \frac{1}{n} \sum_{i=1}^n \ln \frac{\cosh(y_i \langle x_i, \theta \rangle / \sigma^2 + \nu)}{\cosh \nu} + \mathbf{C}_n \end{aligned}$$

where $\mathbb{E}[\cdot] = \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\cdot]$ is the expectation over the ground truth distribution $p(s|\theta^*, \pi^*)$, and $\mathbf{C}$ and $\mathbf{C}_n$ are constants that are independent of θ and π .

Proof. By the detailed derivation of negative log-likelihood in Appendix D (which is adapted from pages 20-25, Appendix B of [1]), we have shown the above expressions for $f(\theta, \pi)$ and $f_n(\theta, \pi)$ with these constants $\mathbf{C}$ and $\mathbf{C}_n$.

$$\begin{aligned} \mathbf{C} &= \frac{1}{2\sigma^2} \mathbb{E}[y^2] - \mathbb{E}[\ln p(x)] - \frac{1}{2} \ln(2\pi\sigma^2) \stackrel{x \sim \mathcal{N}(0, I_d)}{=} \frac{1}{2\sigma^2} (\|\theta^*\|^2 + \sigma^2) - \frac{d}{2} (1 + \ln(2\pi)) - \frac{1}{2} \ln(2\pi\sigma^2) \\ \mathbf{C}_n &= \frac{1}{2\sigma^2} \cdot \frac{1}{n} \sum_{i=1}^n y_i^2 - \frac{1}{n} \sum_{i=1}^n \ln p(x_i) - \frac{1}{2} \ln(2\pi\sigma^2) \end{aligned}$$

□

Theorem A.2 (Lemma 2: EM Update Rules and Gradients). *Let $U(\theta, \nu) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \ln \cosh(y \langle x, \theta \rangle / \sigma^2 + \nu)$ at population level and $U_n(\theta, \nu) := \frac{1}{n} \sum_{i=1}^n \ln \cosh(y_i \langle x_i, \theta \rangle / \sigma^2 + \nu)$ at finite-sample level, then the population EM update rules $M(\theta, \nu)$, $N(\theta, \nu)$ and finite-sample EM update rules $M_n(\theta, \nu)$, $N_n(\theta, \nu)$ for regression parameters θ and imbalance of mixing weights $\tanh \nu$ are:*

$$\begin{aligned} M(\theta, \nu) &= \mathbb{E}[xx^\top]^{-1} \sigma^2 \nabla_\theta U(\theta, \nu), & N(\theta, \nu) &= \nabla_\nu U(\theta, \nu), \\ M_n(\theta, \nu) &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \sigma^2 \nabla_\theta U_n(\theta, \nu), & N_n(\theta, \nu) &= \nabla_\nu U_n(\theta, \nu). \end{aligned}$$

Proof. By the derivation of EM update rules in Appendix D (which is adapted from pages 20-25, Appendix B of [1]), and the Leibniz rule such that $\nabla_\theta \mathbb{E}[\ln \cosh(\cdot)] = \mathbb{E}[\nabla_\theta \ln \cosh(\cdot)]$, $\nabla_\nu \mathbb{E}[\ln \cosh(\cdot)] = \mathbb{E}[\nabla_\nu \ln \cosh(\cdot)]$ and $\frac{d}{dt} \ln \cosh(t) = \tanh(t)$, we have the population EM update rules as follows, and it is worth noting that $\mathbb{E}[xx^\top] = I_d$ for $x \sim \mathcal{N}(0, I_d)$:

$$\begin{aligned} M(\theta, \nu) &= \mathbb{E}[xx^\top]^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) yx = \mathbb{E}[xx^\top]^{-1} \nabla_\theta U(\theta, \nu) = \mathbb{E}[xx^\top]^{-1} \sigma^2 \nabla_\theta U(\theta, \nu) \\ N(\theta, \nu) &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) = \nabla_\nu U(\theta, \nu) \end{aligned}$$

Similarly, we can establish the results for the finite-sample EM update rules by substituting f_n, U_n for f, U and $\frac{1}{n} \sum_{i=1}^n, s_i := (x_i, y_i)$ for $\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}, s := (x, y)$ into the above expressions. □

Theorem A.3 (Proposition 3: Connection between EM and Gradient Descent). *The population/finite-sample EM update rules give:*

$$\begin{aligned} M(\theta, \nu) &= \theta - \mathbb{E}[xx^\top]^{-1} \sigma^2 \nabla_\theta f(\theta, \pi), & N(\theta, \nu) &= \tanh \nu - \nabla_\nu f(\theta, \pi), \\ M_n(\theta, \nu) &= \theta - \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \sigma^2 \nabla_\theta f_n(\theta, \pi), & N_n(\theta, \nu) &= \tanh \nu - \nabla_\nu f_n(\theta, \pi). \end{aligned}$$

Proof. Noticing that $\frac{d}{dt} \ln \cosh(t) = \tanh(t)$ and the direct computation of the gradients of the negative log-likelihood function $f(\theta, \pi), f_n(\theta, \pi)$ with respect to θ and ν and comparing with the EM update rules $M(\theta, \nu), N(\theta, \nu)$ and $M_n(\theta, \nu), N_n(\theta, \nu)$ gives the results. It is worth noting that $\mathbb{E}[xx^\top] = I_d$ for the random variable $x \sim \mathcal{N}(0, I_d)$. □

B. Explicit EM Update Expressions

Theorem A.4 (Theorem 4: Explicit EM Update Expressions). *Let $A_\eta = k\eta^2\sqrt{1+\eta^{-2}}$, $k = \frac{\|\theta\|}{\|\theta^*\|}$, $\eta = \frac{\|\theta^*\|}{\sigma}$, and $\varphi = \frac{\pi}{2} - \arccos|\rho| \in (0, \pi/2]$, $\rho = \frac{\langle\theta, \theta^*\rangle}{\|\theta\|\|\theta^*\|}$, and $\varphi_\eta = \arcsin(\sin\varphi/\sqrt{1+\eta^{-2}})$, namely $\sqrt{1+\eta^{-2}} = \sin\varphi/\sin\varphi_\eta$, and Gaussian random variables $g_\eta, g' \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[g_\eta g'] = \sin\varphi_\eta$, then population EM update rules for regression parameters θ and imbalance of mixing weights $\tanh\nu$ are:*

$$\begin{aligned} M(\theta, \nu)/\|\theta^*\| &= \vec{e}_1 \frac{\sin\varphi}{\sin\varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g' + \text{sgn}(\rho)(-1)^{z+1}\nu) g_\eta g'] \\ &\quad + \vec{e}_2 \frac{\cos\varphi}{\cos^2\varphi_\eta} \text{sgn}(\rho) \mathbb{E}[\tanh(A_\eta g_\eta g' + \text{sgn}(\rho)(-1)^{z+1}\nu) g_\eta (g_\eta - \sin\varphi_\eta \cdot g')] \\ N(\theta, \nu) &= \text{sgn}(\rho) \mathbb{E}[\tanh((-1)^{z+1} A_\eta g_\eta g' + \text{sgn}(\rho)\nu)]. \end{aligned}$$

Proof. Let's start from the EM update rules, $M(\theta, \nu) = \mathbb{E}\left[\tanh\left(\frac{y\langle x, \theta \rangle}{\sigma^2} + \nu\right) yx\right]$, $N(\theta, \nu) = \mathbb{E}\left[\tanh\left(\frac{y\langle x, \theta \rangle}{\sigma^2} + \nu\right)\right]$, also let $k := \frac{\|\theta\|}{\|\theta^*\|}$, $\eta := \frac{\|\theta^*\|}{\sigma}$, $Z_\varepsilon = (-1)^{z+1}\varepsilon/\sigma \sim \mathcal{N}(0, 1)$, and denote cosine of angle $\rho := \frac{\langle\theta, \theta^*\rangle}{\|\theta\|\|\theta^*\|}$, a pair of orthonormal vectors $\vec{e}_1 := \theta/\|\theta\|$, $\vec{e}_2 = \frac{\theta^* - \langle\theta^*, \vec{e}_1\rangle\vec{e}_1}{\|\theta^* - \langle\theta^*, \vec{e}_1\rangle\vec{e}_1\|}$. By introducing $Z = \langle x, \vec{e}_1 \rangle$, $Z_* = \langle x, \vec{e}_2 \rangle$, and $x^\perp = x - Z\vec{e}_1 - Z_*\vec{e}_2$, then:

$$\begin{aligned} y/\|\theta^*\| &= (-1)^{z+1}[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon] \\ y\langle x, \theta \rangle/\sigma^2 &= (\|\theta^*\|^2/\sigma^2) \times (y\langle x, \theta \rangle/\|\theta^*\|^2) = \eta^2 \times (-1)^{z+1}[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon] \cdot kZ \\ yx/\|\theta^*\| &= (-1)^{z+1}[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon] \cdot [Z\vec{e}_1 + Z_*\vec{e}_2 + x^\perp] \end{aligned}$$

Therefore, by using the independence of Z, Z_*, Z_ε and $x^\perp$, then

$$\begin{aligned} M(\theta, \nu)/\|\theta^*\| &= \vec{e}_1 \mathbb{E}\left[\tanh\left(k\eta^2\left[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon\right] Z + (-1)^{z+1}\nu\right)\left[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon\right] Z\right] \\ &\quad + \vec{e}_2 \mathbb{E}\left[\tanh\left(k\eta^2\left[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon\right] Z + (-1)^{z+1}\nu\right)\left[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon\right] Z_*\right] \\ N(\theta, \nu) &= \mathbb{E}\left[\tanh\left((-1)^{z+1}k\eta^2\left[\rho Z + \sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon\right] Z + \nu\right)\right] \end{aligned}$$

By introducing $Z' := (\sqrt{1-\rho^2}Z_* + \eta^{-1} \cdot Z_\varepsilon)/\sqrt{(1-\rho^2) + \eta^{-2}}$, then $(Z_* - (\sqrt{1-\rho^2}/\sqrt{(1-\rho^2) + \eta^{-2}})Z') \perp Z', Z$:

$$\begin{aligned} M(\theta, \nu)/\|\theta^*\| &= \vec{e}_1 \mathbb{E}\left[\tanh\left(k\eta^2\left[\rho Z + \sqrt{(1-\rho^2) + \eta^{-2}}Z'\right] Z + (-1)^{z+1}\nu\right)\left[\rho Z + \sqrt{(1-\rho^2) + \eta^{-2}}Z'\right] Z\right] \\ &\quad + \vec{e}_2 \frac{\sqrt{1-\rho^2}}{\sqrt{(1-\rho^2) + \eta^{-2}}} \cdot \mathbb{E}\left[\tanh\left(k\eta^2\left[\rho Z + \sqrt{(1-\rho^2) + \eta^{-2}}Z'\right] Z + (-1)^{z+1}\nu\right)\left[\rho Z + \sqrt{(1-\rho^2) + \eta^{-2}}Z'\right] Z'\right] \\ N(\theta, \nu) &= \mathbb{E}\left[\tanh\left((-1)^{z+1}k\eta^2\left[\rho Z + \sqrt{(1-\rho^2) + \eta^{-2}}Z'\right] Z + \nu\right)\right] \end{aligned}$$

If $\rho \neq 0$, then $\varphi = \frac{\pi}{2} - \arccos|\rho| \in (0, \pi/2]$, therefore $\rho = \text{sgn}(\rho) \sin\varphi$, $\sqrt{1-\rho^2} = \cos\varphi$, we define $A_\eta = k\eta^2\sqrt{1+\eta^{-2}}$ and $\varphi_\eta := \arcsin(\sin\varphi/\sqrt{1+\eta^{-2}})$, namely $\sqrt{1+\eta^{-2}} = \sin\varphi/\sin\varphi_\eta$. By introducing $g' \equiv Z'' := \text{sgn}(\rho)Z \sim \mathcal{N}(0, 1)$ and $g := (\rho Z + \sqrt{1-\rho^2}Z') \sim \mathcal{N}(0, 1)$, $g_\eta := (\rho Z + \sqrt{(1-\rho^2) + \eta^{-2}}Z')/\sqrt{1+\eta^{-2}}$ with $\mathbb{E}[gg'] = \sin\varphi$, $\mathbb{E}[g_\eta g'] = \sin\varphi_\eta$:

$$\begin{aligned} M(\theta, \nu)/\|\theta^*\| &= \vec{e}_1 \frac{\sin\varphi}{\sin\varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g' + \text{sgn}(\rho)(-1)^{z+1}\nu) g_\eta g'] \\ &\quad + \vec{e}_2 \frac{\cos\varphi}{\cos^2\varphi_\eta} \text{sgn}(\rho) \mathbb{E}[\tanh(A_\eta g_\eta g' + \text{sgn}(\rho)(-1)^{z+1}\nu) g_\eta (g_\eta - \sin\varphi_\eta \cdot g')] \\ N(\theta, \nu) &= \text{sgn}(\rho) \mathbb{E}[\tanh((-1)^{z+1} A_\eta g_\eta g' + \text{sgn}(\rho)\nu)] \end{aligned}$$

If $\rho = 0$, then $\varphi = 0$, $\varphi_\eta = 0$, $\lim_{\varphi \rightarrow 0} \sin\varphi_\eta/\sin\varphi = 1/\sqrt{1+\eta^{-2}}$, $\cos\varphi = \cos\varphi_\eta = 1$, $A_\eta = k\eta^2\sqrt{1+\eta^{-2}}$ and $g = g' = g_\eta$ with $\mathbb{E}[gg'] = \mathbb{E}[g_\eta g'] = 0$. Therefore, the product $(-1)^{z+1}g_\eta g'$ of two independent standard normal random variables g_η, g' has a density involving Bessel function $X = (-1)^{z+1}g_\eta g' \sim f_X(x) = K_0(|x|)/\pi$ (see page 50, Section 4.4 Bessel Function Distributions, Chapter 12 Continuous Distributions (General) of [49]),

$$\begin{aligned} M(\theta, \nu)/\|\theta^*\| &= \vec{e}_1 \sqrt{1+\eta^{-2}} \mathbb{E}[\tanh(A_\eta X + \nu) X] + \vec{e}_2 \mathbb{E}[\tanh(A_\eta g_\eta g' + (-1)^{z+1}\nu) g_\eta^2] \\ &= \vec{e}_1 \sqrt{1+\eta^{-2}} \mathbb{E}[\tanh(A_\eta X + \nu) X] + \vec{e}_2 \mathbb{E}[(-1)^{z+1}] \mathbb{E}[\tanh(A_\eta g_\eta g' + \nu) g_\eta^2] \\ N(\theta, \nu) &= \mathbb{E}[\tanh(A_\eta X + \nu)] \end{aligned}$$

The above expressions when $\rho = 0$ also align with the expressions when $\rho \neq 0$ with the convention of $\text{sgn}(0) = 1$ and $(\sin\varphi_\eta/\sin\varphi)_{\rho=0} = \lim_{\varphi \rightarrow 0} \sin\varphi_\eta/\sin\varphi = 1/\sqrt{1+\eta^{-2}}$. $\square$

C. Bound of the EM Update Rule in General SNR Regime

Theorem A.5 (Proposition 5: Boundedness of EM Update Rule). *The EM update rule for θ is bounded by the following bound, which depends on SNR $\eta = \|\theta^*\|/\sigma$:*

$$\|M(\theta, \nu)\| \leq \frac{\arctan \eta}{\pi/2} \|\theta^*\| + \frac{2}{\pi} \sigma.$$

Proof. Since we can express $g' = \sin \varphi_\eta g_\eta + \cos \varphi_\eta h_\eta$ for some $h_\eta \sim \mathcal{N}(0, 1)$ with $h_\eta \perp g_\eta$, noting that $\mathbb{E}[g_\eta h_\eta] = \mathbb{E}[g_\eta] \cdot \mathbb{E}[h_\eta] = \frac{2}{\pi}$ and $\mathbb{E}[g_\eta g'] = \frac{2}{\pi} [\varphi_\eta \sin \varphi_\eta + \cos \varphi_\eta]$, then for $M(\theta, \nu) \in \text{span}\{\vec{e}_1, \vec{e}_2\}$:

$$\begin{aligned} |\langle M(\theta, \nu) / \|\theta^*\|, \vec{e}_1 \rangle| &\leq \frac{\sin \varphi}{\sin \varphi_\eta} \mathbb{E}[g_\eta g'] = \frac{2}{\pi} \sin \varphi \left[\varphi_\eta + \frac{\cos \varphi_\eta}{\sin \varphi_\eta} \right] \\ |\langle M(\theta, \nu) / \|\theta^*\|, \vec{e}_2 \rangle| &\leq \frac{\cos \varphi}{\cos \varphi_\eta} \mathbb{E}[g_\eta h_\eta] = \frac{2}{\pi} \frac{\cos \varphi}{\cos \varphi_\eta} \\ \|M(\theta, \nu) / \|\theta^*\| &= \sqrt{|\langle M(\theta, \nu) / \|\theta^*\|, \vec{e}_1 \rangle|^2 + |\langle M(\theta, \nu) / \|\theta^*\|, \vec{e}_2 \rangle|^2} \\ &= \frac{2}{\pi} \sqrt{H(\sin \varphi, \eta^{-2})} \leq \frac{2}{\pi} \sqrt{H(1, \eta^{-2})} = \frac{2}{\pi} (\arctan \eta + \eta^{-1}) \end{aligned}$$

where the last inequality is due to the fact (Lemma E.6) that $H(s, \delta)$ is non-decreasing with respect to $s = \sin \varphi \in [0, 1]$ given $\delta := \eta^{-2} \geq 0$, and $H(s, \delta)$ is defined as:

$$H(s, \delta) = \left(s \arcsin \frac{s}{\sqrt{1+\delta}} + \sqrt{1-s^2+\delta} \right)^2 + \frac{(1+\delta)(1-s^2)}{1-s^2+\delta}$$

Therefore, we establish the bound for $\|M(\theta, \nu)\|$, which depends on SNR $\eta = \|\theta^*\|/\sigma$.

$$\|M(\theta, \nu)\| \leq \frac{\arctan \eta}{\pi/2} \|\theta^*\| + \frac{2}{\pi} \sigma$$

□

D. Analysis for the Fixed Points of EM Updates in General SNR Regime

Theorem A.6 (Proposition 6: Distinct Fixed Points of EM Update Rules). *The EM update rules $M(\theta, \nu), N(\theta, \nu)$ for regression parameters θ and the imbalance of mixing weights $\tanh \nu$ have the following three distinct fixed points $(\theta^*, \nu^*), (-\theta^*, -\nu^*), (\vec{0}, 0)$:*

$$\begin{pmatrix} M(\theta^*, \nu^*) \\ N(\theta^*, \nu^*) \end{pmatrix} = \begin{pmatrix} \theta^* \\ \tanh \nu^* \end{pmatrix}, \quad \begin{pmatrix} M(-\theta^*, -\nu^*) \\ N(-\theta^*, -\nu^*) \end{pmatrix} = \begin{pmatrix} -\theta^* \\ \tanh(-\nu^*) \end{pmatrix}, \quad \begin{pmatrix} M(\vec{0}, 0) \\ N(\vec{0}, 0) \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \tanh(0) \end{pmatrix}.$$

Proof. For the point $(\theta, \nu) = (\vec{0}, 0)$, by the equations (6), (7) of the EM update rules, we have:

$$M(\vec{0}, 0) = \mathbb{E} \left[\tanh \left(\frac{y \langle x, \vec{0} \rangle}{\sigma^2} + 0 \right) yx \right] = \tanh(0) \cdot \mathbb{E}[yx] = \vec{0}, \quad N(\vec{0}, 0) = \mathbb{E} \left[\tanh \left(\frac{y \langle x, \vec{0} \rangle}{\sigma^2} + 0 \right) \right] = \tanh(0)$$

For the point $(\theta, \nu) = (\theta^*, \nu^*)$, we have $k = \|\theta\|/\|\theta^*\| = 1, \varphi = \pi/2, \text{sgn}(\rho) = 1, \vec{e}_1 = \theta/\|\theta\| = \theta^*/\|\theta^*\|$, then $\cos \varphi = 0, \sin \varphi = 1, \sin \varphi_\eta = \sin \varphi / \sqrt{1 + \eta^{-2}} = 1 / \sqrt{1 + \eta^{-2}}, \cos \varphi_\eta = \eta^{-1} / \sqrt{1 + \eta^{-2}}, A_\eta = k \eta^2 \sqrt{1 + \eta^{-2}} = \eta^2 \sqrt{1 + \eta^{-2}} = \sin \varphi_\eta / \cos^2 \varphi_\eta$, by the expressions of the EM update rules with $g_\eta, g' \sim \mathcal{N}(0, 1)$ and $\mathbb{E}[g_\eta g'] = \sin \varphi_\eta$:

$$\begin{aligned} M(\theta^*, \nu^*) &= \theta^* \frac{\cos^2 \varphi_\eta}{\sin \varphi_\eta} \mathbb{E} \left[\tanh \left(\sin \varphi_\eta \frac{g_\eta g'}{\cos^2 \varphi_\eta} + (-1)^{z+1} \nu^* \right) \frac{g_\eta g'}{\cos^2 \varphi_\eta} \right] \\ N(\theta^*, \nu^*) &= \mathbb{E} \left[\tanh \left((-1)^{z+1} \sin \varphi_\eta \frac{g_\eta g'}{\cos^2 \varphi_\eta} + \nu^* \right) \right] \end{aligned}$$

Note that the distribution of $g_\eta g' / \cos^2 \varphi_\eta$ has the density involving Bessel function $\cos \varphi_\eta \exp(\sin \varphi_\eta \cdot x) K_0(|x|) / \pi$ (see page 50, Section 4.4 Bessel Function Distributions, Chapter 12 Continuous Distributions (General) of [49]), therefore by introducing the random variable X with a simple probability density function $X \sim f_X(x) = K_0(|x|) / \pi$, we can write the expectations as:

$$\begin{aligned} \mathbb{E} \left[\tanh \left(\sin \varphi_\eta \frac{g_\eta g'}{\cos^2 \varphi_\eta} + (-1)^{z+1} \nu^* \right) \frac{g_\eta g'}{\cos^2 \varphi_\eta} \right] &= \cos \varphi_\eta \mathbb{E}_X \mathbb{E}_z \left[\exp(\sin \varphi_\eta X) \tanh(\sin \varphi_\eta X + (-1)^{z+1} \nu^*) X \right] \\ \mathbb{E} \left[\tanh \left((-1)^{z+1} \sin \varphi_\eta \frac{g_\eta g'}{\cos^2 \varphi_\eta} + \nu^* \right) \right] &= \cos \varphi_\eta \mathbb{E}_X \mathbb{E}_z \left[\exp(\sin \varphi_\eta X) \tanh((-1)^{z+1} \sin \varphi_\eta X + \nu^*) \right] \end{aligned}$$

Note that $\mathbb{P}[z = 1] = \frac{1}{2} + \frac{1}{2} \tanh \nu^* = \frac{\exp(\nu^*)}{2 \cosh \nu^*}$, $\mathbb{P}[z = -1] = \frac{1}{2} - \frac{1}{2} \tanh \nu^* = \frac{\exp(-\nu^*)}{2 \cosh \nu^*}$, X is a symmetric random variable, and $(\exp(t) + \exp(-t)) \tanh(t) = 2 \sinh(t)$, $2 \sinh(a+b) = \exp(b) \exp(a) - \exp(-b) \exp(-a)$, $2 \cosh(b) = \exp(b) + \exp(-b)$:

$$\begin{aligned} \mathbb{E}_X \mathbb{E}_z [\exp(\sin \varphi_\eta X) \tanh(\sin \varphi_\eta X + (-1)^{z+1} \nu^*) X] &= \frac{1}{\cosh \nu^*} \mathbb{E}_X [\sinh(\sin \varphi_\eta X + \nu^*) X] = \mathbb{E}_X [\exp(\sin \varphi_\eta X) X] \\ \mathbb{E}_X \mathbb{E}_z [\exp(\sin \varphi_\eta X) \tanh((-1)^{z+1} \sin \varphi_\eta X + \nu^*)] &= \frac{1}{\cosh \nu^*} \mathbb{E}_X [\sinh(\sin \varphi_\eta X + \nu^*)] = \tanh \nu^* \mathbb{E}_X [\exp(\sin \varphi_\eta X)] \end{aligned}$$

By letting $\nu \rightarrow 0$ in the third formula in table 6.611 with modified Bessel function K_ν , Section 6.61 Combinations of Bessel functions and exponentials, Page 703 of [55], applying Leibniz integral rule or the dominated convergence theorem to exchange the order of taking limit and taking expectations (see Theorem 1.5.8, page 24 of [56]) with $|\alpha| < 1$:

$$\mathbb{E}_X [\exp(\alpha X)] = \frac{1}{\sqrt{1-\alpha^2}} \implies \mathbb{E}_X [\exp(\alpha X) X] = \mathbb{E}_X [\partial_\alpha \exp(\alpha X)] = \frac{d}{d\alpha} \mathbb{E}_X [\exp(\alpha X)] = \frac{\alpha}{(1-\alpha^2)^{3/2}}$$

By substituting $\alpha \leftarrow \sin \varphi_\eta$ in the above expressions, we have:

$$\mathbb{E}_X [\exp(\sin \varphi_\eta X) X] = \frac{\sin \varphi_\eta}{\cos^3 \varphi_\eta}, \quad \mathbb{E}_X [\exp(\sin \varphi_\eta X)] = \frac{1}{\cos \varphi_\eta}$$

Putting the above expressions together, we have shown that:

$$\begin{aligned} M(\theta^*, \nu^*) &= \theta^* \frac{\cos^2 \varphi_\eta}{\sin \varphi_\eta} \cos \varphi_\eta \mathbb{E}_X [\exp(\sin \varphi_\eta X) X] = \theta^* \frac{\cos^2 \varphi_\eta}{\sin \varphi_\eta} \cos \varphi_\eta \frac{\sin \varphi_\eta}{\cos^3 \varphi_\eta} = \theta^* \\ N(\theta^*, \nu^*) &= \cos \varphi_\eta \tanh \nu^* \mathbb{E}_X [\exp(\sin \varphi_\eta X)] = \cos \varphi_\eta \tanh \nu^* \frac{1}{\cos \varphi_\eta} = \tanh \nu^* \end{aligned}$$

For the point $(\theta, \nu) = (-\theta^*, -\nu^*)$, we have $k = \|\theta\|/\|\theta^*\| = 1$, $\varphi = \pi/2$, $\text{sgn}(\rho) = -1$, $\vec{e}_1 = \theta/\|\theta\| = -\theta^*/\|\theta^*\|$, then $\cos \varphi = 0$, $\sin \varphi = 1$, $\sin \varphi_\eta = \sin \varphi / \sqrt{1+\eta^{-2}} = -1/\sqrt{1+\eta^{-2}}$, $\cos \varphi_\eta = \eta^{-1}/\sqrt{1+\eta^{-2}}$, $A_\eta = k\eta^2 \sqrt{1+\eta^{-2}} = \eta^2 \sqrt{1+\eta^{-2}} = \sin \varphi_\eta / \cos^2 \varphi_\eta$, by the expressions of the EM update rules with $g_\eta, g' \sim \mathcal{N}(0, 1)$ and $\mathbb{E}[g_\eta g'] = \sin \varphi_\eta$:

$$\begin{aligned} M(-\theta^*, -\nu^*) &= -\theta^* \frac{\cos^2 \varphi_\eta}{\sin \varphi_\eta} \mathbb{E} \left[\tanh \left(\sin \varphi_\eta \frac{g_\eta g'}{\cos^2 \varphi_\eta} + (-1)^{z+1} \nu^* \right) \frac{g_\eta g'}{\cos^2 \varphi_\eta} \right] = -M(\theta^*, \nu^*) = -\theta^* \\ N(-\theta^*, -\nu^*) &= -\mathbb{E} \left[\tanh \left((-1)^{z+1} \sin \varphi_\eta \frac{g_\eta g'}{\cos^2 \varphi_\eta} + \nu^* \right) \right] = -N(\theta^*, \nu^*) = \tanh(-\nu^*) \end{aligned}$$

Therefore, we have verified these three distinct fixed points (θ^*, ν^*) , $(-\theta^*, -\nu^*)$, $(\vec{0}, 0)$ of the EM update rules. $\square$

Theorem A.7 (Proposition 7: Contraction Property of EM Update Rule around the Fixed Point). *Suppose that the unit direction vector $\vec{e}^\perp$ is orthogonal to the ground truth θ^* of regression parameters, namely $\langle \theta^*, \vec{e}^\perp \rangle = 0$, $\|\vec{e}^\perp\| = 1$, there exists $k^*(\eta) > 0$ which is determined by $\eta = \|\theta^*\|/\sigma$ such that the $(k^*(\eta)\|\theta^*\|\vec{e}^\perp, 0)$ is a fixed point of the EM update rule $M(\theta, \nu)$, $N(\theta, \nu)$ for regression parameters θ and the imbalance of mixing weights $\tanh \nu$ given any SNR $\eta = \|\theta^*\|/\sigma > 0$:*

$$\begin{pmatrix} M(k^*(\eta)\|\theta^*\|\vec{e}^\perp, 0) \\ N(k^*(\eta)\|\theta^*\|\vec{e}^\perp, 0) \end{pmatrix} = \begin{pmatrix} k^*(\eta)\|\theta^*\|\vec{e}^\perp \\ \tanh(0) \end{pmatrix}.$$

For any $k > 0$, $N(k\|\theta^*\|\vec{e}^\perp, 0) = \tanh(0)$ and $M(k\|\theta^*\|\vec{e}^\perp, 0)$, $\vec{e}^\perp$ have the same direction,

$$\begin{cases} 0 < \|M(k\|\theta^*\|\vec{e}^\perp, 0)\|/\|\theta^*\| - k < k^*(\eta) - k & \text{if } k < k^*(\eta), \\ 0 = \|M(k\|\theta^*\|\vec{e}^\perp, 0)\|/\|\theta^*\| - k = k^*(\eta) - k & \text{if } k = k^*(\eta), \\ 0 > \|M(k\|\theta^*\|\vec{e}^\perp, 0)\|/\|\theta^*\| - k > k^*(\eta) - k & \text{if } k > k^*(\eta). \end{cases}$$

Moreover, the bounds of $k^*(\eta)$ are:

$$\frac{1}{\sqrt{3}} < k^*(\eta) < \min \left(\frac{2}{\pi} \sqrt{1+\eta^{-2}}, 1 \right), \forall \eta > 0.$$

with $\lim_{\eta \rightarrow 0^+} k^*(\eta) = \frac{1}{\sqrt{3}}$ and $\lim_{\eta \rightarrow \infty} k^*(\eta) = \frac{2}{\pi}$.

Proof. Suppose the regression parameters $\theta = k\|\theta^*\|\vec{e}^\perp$ for some $k > 0$ and the imbalance of mixing weights $\tanh \nu = 0$, then by proof of for the expressions of the EM update rules, we have the following expressions with $A_\eta = k\eta^2 \sqrt{1+\eta^{-2}}$:

$$\begin{aligned} M(k\|\theta^*\|\vec{e}^\perp, 0)/\|\theta^*\| &= \vec{e}_1 \sqrt{1+\eta^{-2}} \mathbb{E}[\tanh(A_\eta X + 0) X] + \vec{e}_2 \mathbb{E}[(-1)^{z+1}] \mathbb{E}[\tanh(A_\eta g_\eta g' + 0) g_\eta^2] \\ &= \vec{e}^\perp \sqrt{1+\eta^{-2}} \mathbb{E}[\tanh(A_\eta X) X] \\ N(k\|\theta^*\|\vec{e}^\perp, 0) &= \mathbb{E}[\tanh(A_\eta X + 0)] = \mathbb{E}[\tanh(A_\eta X)] = 0 \end{aligned}$$

due to the fact that $\vec{e}_1 = \theta/\|\theta\| = \vec{e}^\perp$ and $X = (-1)^{z+1}g_\eta g' \sim f_X(x) = K_0(|x|)/\pi$ is a symmetric random variable, and $g_\eta, g' \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[g_\eta g'] = 0$,

$$\mathbb{E}[\tanh(A_\eta g_\eta g') g_\eta^2] = \mathbb{E}_{g_\eta} \mathbb{E}_{g'}[\tanh(A_\eta g_\eta g') g_\eta^2] = \mathbb{E}_{g_\eta}[0] = 0$$

Moreover, the EM update $M(k\|\theta^*\|\vec{e}^\perp, 0)$ and $\vec{e}^\perp$ have the same direction, since

$$M(k\|\theta^*\|\vec{e}^\perp, 0) = \vec{e}^\perp \sqrt{1 + \eta^{-2}} \mathbb{E}[\tanh(A_\eta X) X] \in \text{span}\{\vec{e}^\perp\}, \left\langle \frac{M(k\|\theta^*\|\vec{e}^\perp, 0)}{\|\theta^*\|}, \vec{e}^\perp \right\rangle = \sqrt{1 + \eta^{-2}} \mathbb{E}[\tanh(A_\eta X) X] > 0$$

For the ease of analysis, we introduce the following function $G_\eta(k)$ with $k > 0$:

$$G_\eta(k) = \frac{\|M(k\|\theta^*\|\vec{e}^\perp, 0)\|}{\|\theta^*\|} - k = \sqrt{1 + \eta^{-2}} \mathbb{E}[\tanh(A_\eta X) X] - k = \sqrt{1 + \eta^{-2}} \mathbb{E}[\tanh(k\eta^2 \sqrt{1 + \eta^{-2}} X) X] - k$$

We obtain the derivative and second derivative of $G_\eta(k)$ as follows by applying Leibniz integral rule or the dominated convergence theorem to exchange the order of taking derivative and taking expectations (see Theorem 1.5.8, page 24 of [56]):

$$\begin{aligned} \frac{\partial G_\eta(k)}{\partial k} &= \sqrt{1 + \eta^{-2}} \mathbb{E}[\partial_k \tanh(k\eta^2 \sqrt{1 + \eta^{-2}} X) X] - 1 = \eta^2(1 + \eta^{-2}) \mathbb{E} \left[[1 - \tanh^2(k\eta^2 \sqrt{1 + \eta^{-2}} X)] X^2 \right] - 1 \\ \frac{\partial^2 G_\eta(k)}{\partial k^2} &= \eta^2(1 + \eta^{-2}) \mathbb{E} \partial_k \left[\frac{X^2}{\cosh^2(k\eta^2 \sqrt{1 + \eta^{-2}} X)} \right] = -2\eta^4(1 + \eta^{-2})^{3/2} \mathbb{E} \left[\frac{\tanh(k\eta^2 \sqrt{1 + \eta^{-2}} X) X^3}{\cosh^2(k\eta^2 \sqrt{1 + \eta^{-2}} X)} \right] < 0 \end{aligned}$$

The initial/final values of $G_\eta(k)$ and its derivative are obtained by applying the dominated convergence theorem to exchange the order of taking derivative and taking expectations (see Theorem 1.5.8, page 24 of [56]) and using the fact $\mathbb{E}[X^2] = 1$:

$$\begin{aligned} \lim_{k \rightarrow 0_+} G_\eta(k) &= \sqrt{1 + \eta^{-2}} \mathbb{E}[\lim_{k \rightarrow 0_+} \tanh(k\eta^2 \sqrt{1 + \eta^{-2}} X) X] - 0 = \mathbb{E}[0] - 0 = 0 \\ \lim_{k \rightarrow 0_+} \frac{\partial G_\eta(k)}{\partial k} &= \eta^2(1 + \eta^{-2}) \mathbb{E} \left[[1 - \lim_{k \rightarrow 0_+} \tanh^2(k\eta^2 \sqrt{1 + \eta^{-2}} X)] X^2 \right] - 1 = \eta^2(1 + \eta^{-2}) - 1 = \eta^2 > 0 \\ \lim_{k \rightarrow \infty} G_\eta(k) &= \sqrt{1 + \eta^{-2}} \mathbb{E}[\lim_{k \rightarrow \infty} \tanh(k\eta^2 \sqrt{1 + \eta^{-2}} X) X] - \lim_{k \rightarrow \infty} k = \mathbb{E}[|X|] - \infty = -\infty \\ \lim_{k \rightarrow \infty} \frac{\partial G_\eta(k)}{\partial k} &= \eta^2(1 + \eta^{-2}) \mathbb{E} \left[[1 - \lim_{k \rightarrow \infty} \tanh^2(k\eta^2 \sqrt{1 + \eta^{-2}} X)] X^2 \right] - 1 = \eta^2(1 + \eta^{-2}) \mathbb{E}[0] - 1 = -1 \end{aligned}$$

The signs of some special values of $G_\eta(k)$ are determined by Lemma E.4 with $A := \eta^2 \sqrt{1 + \eta^{-2}}$, $\frac{1}{\sqrt{1 + \eta^{-2}}} = \frac{\sqrt{4A^2 + 1} - 1}{2A}$ and $A' = \frac{1}{\sqrt{3}} \eta^2 \sqrt{1 + \eta^{-2}}$, $\frac{1}{\sqrt{1 + \eta^{-2}}} = \frac{\sqrt{12A'^2 + 1} - 1}{2\sqrt{3}A'}$ and noting the fact that $\tanh(x) < 1, \forall x > 0$ and $\mathbb{E}[|X|] = \frac{2}{\pi}$:

$$\begin{aligned} G_\eta \left(\frac{2}{\pi} \sqrt{1 + \eta^{-2}} \right) &< \sqrt{1 + \eta^{-2}} \mathbb{E}[|X|] - \frac{2}{\pi} \sqrt{1 + \eta^{-2}} = 0 \\ G_\eta(1) &= \sqrt{1 + \eta^{-2}} \left(\mathbb{E}[\tanh(AX) X] - \frac{\sqrt{4A^2 + 1} - 1}{2A} \right) < 0 \\ G_\eta \left(\frac{1}{\sqrt{3}} \right) &= \sqrt{1 + \eta^{-2}} \left(\mathbb{E}[\tanh(A'X) X] - \frac{\sqrt{12A'^2 + 1} - 1}{6A'} \right) > 0 \end{aligned}$$

Existence and uniqueness of $k^*(\eta)$: Therefore, there exists $k^*(\eta) \in (\frac{1}{\sqrt{3}}, \min(\frac{2}{\pi} \sqrt{1 + \eta^{-2}}, 1))$ such that $G_\eta(k^*(\eta)) = 0$ by the continuity of $G_\eta(k)$ and the intermediate value theorem (see 4.23 Theorem on page 93 of [57]). Moreover, the root $k^*(\eta)$ of $G_\eta(k) = 0$ is unique, otherwise it would contradict to the fact of $G_\eta(\frac{1}{\sqrt{3}}) > 0$ and $\frac{\partial^2 G_\eta(k)}{\partial k^2} < 0$. If there exists another root $\tilde{k}^*(\eta)$ such that $G_\eta(\tilde{k}^*(\eta)) = 0$ (we may assume $k^*(\eta) < \tilde{k}^*(\eta)$), then there exists $\bar{k} \in (k^*(\eta), \tilde{k}^*(\eta))$ such that $\frac{\partial G_\eta(k)}{\partial k} \big|_{k=\bar{k}} = 0 = \frac{G_\eta(\tilde{k}^*(\eta)) - G_\eta(k^*(\eta))}{\tilde{k}^*(\eta) - k^*(\eta)}$ by using the mean value theorem (see 5.10 Theorem on page 108 of [57]). Then by $\frac{\partial^2 G_\eta(k)}{\partial k^2} < 0$, we have $\frac{\partial G_\eta(k)}{\partial k} > \frac{\partial G_\eta(k)}{\partial k} \big|_{k=\bar{k}} = 0$ for $k \in (\frac{1}{\sqrt{3}}, \bar{k})$, since $k^*(\eta) \in (\frac{1}{\sqrt{3}}, \bar{k})$, we have $G_\eta(k^*(\eta)) > G_\eta(\frac{1}{\sqrt{3}}) > 0$, which contradicts to the fact of $G_\eta(k^*(\eta)) = 0$.

Contraction mapping with fixed point $k^*(\eta)$: Again, by applying the mean value theorem to $G_\eta(k)$ on $(0, k^*(\eta))$ and noting the fact that $\lim_{k \rightarrow 0_+} G_\eta(k) = G_\eta(k^*(\eta)) = 0$, there exists $\bar{k} \in (0, k^*(\eta))$ such that $G_\eta(\bar{k}) = 0$ and $\frac{\partial G_\eta(k)}{\partial k} \big|_{k=\bar{k}} = 0$. Noting $\frac{\partial^2 G_\eta(k)}{\partial k^2} < 0$, we have $\frac{\partial G_\eta(k)}{\partial k} > \frac{\partial G_\eta(k)}{\partial k} \big|_{k=\bar{k}} = 0$ for $k \in (0, \bar{k})$, and $\frac{\partial G_\eta(k)}{\partial k} < \frac{\partial G_\eta(k)}{\partial k} \big|_{k=\bar{k}} = 0$ for $k \in (\bar{k}, \infty)$, and therefore

$$\begin{aligned} G_\eta(k) &> \lim_{k \rightarrow 0_+} G_\eta(k) = 0 \quad \forall k \in (0, \bar{k}); \quad G_\eta(k) > G_\eta(k^*(\eta)) = 0 \quad \forall k \in (\bar{k}, k^*(\eta)) \\ \implies G_\eta(k) &> 0 \quad \forall k \in (0, k^*(\eta)); \quad G_\eta(k) < G_\eta(k^*(\eta)) = 0 \quad \forall k \in (k^*(\eta), \infty) \end{aligned}$$

Recalling the definition of $G_\eta(k) := \frac{\|M(k\|\theta^*\|\bar{e}^\perp, 0)\|}{\|\theta^*\|} - k$ for $k > 0$, noting $G_\eta(k) + k = \|M(k\|\theta^*\|\bar{e}^\perp, 0)\|/\|\theta^*\| = \sqrt{1 + \eta^{-2}}\mathbb{E}[\tanh(k\eta^2\sqrt{1 + \eta^{-2}}X)]$ is strictly increasing with respect to k , we have:

$$\begin{aligned} 0 &< \frac{\|M(k\|\theta^*\|\bar{e}^\perp, 0)\|}{\|\theta^*\|} - k = G_\eta(k) = (G_\eta(k) + k) - k < (G_\eta(k^*(\eta)) + k^*(\eta)) - k = k^*(\eta) - k \quad \text{if } k < k^*(\eta) \\ 0 &= \frac{\|M(k\|\theta^*\|\bar{e}^\perp, 0)\|}{\|\theta^*\|} - k = G_\eta(k) = (G_\eta(k) + k) - k = (G_\eta(k^*(\eta)) + k^*(\eta)) - k = k^*(\eta) - k \quad \text{if } k = k^*(\eta) \\ 0 &> \frac{\|M(k\|\theta^*\|\bar{e}^\perp, 0)\|}{\|\theta^*\|} - k = G_\eta(k) = (G_\eta(k) + k) - k > (G_\eta(k^*(\eta)) + k^*(\eta)) - k = k^*(\eta) - k \quad \text{if } k > k^*(\eta) \end{aligned}$$

Limit behavior of $k^*(\eta)$: To analyze the limit behavior of $k^*(\eta)$ as $\eta \rightarrow 0_+$ and $\eta \rightarrow \infty$, we notice that $A := \eta^2\sqrt{1 + \eta^{-2}} = \eta\sqrt{1 + \eta^2} \rightarrow 0_+$ as $\eta \rightarrow 0_+$ and $A \rightarrow \infty$ as $\eta \rightarrow \infty$, and there is a one-to-one increasing mapping $\eta(A)$ from A to η . Consequently, we write $k^*(\eta(A))$ as $k^*(A)$ for simplicity. From the definition of $k^*(\eta)$ such that $G_\eta(k^*(\eta)) = 0$, we show that $k = k^*(A)$ is the unique root of $F_1(A, k) = 0$ for any $A > 0$, and is also the unique root of $F_2(A, k) = 0$ for any $A > 0$, where $F_1(A, k), F_2(A, k)$ are defined as follows:

$$\begin{aligned} F_1(A, k) &= \left(\mathbb{E}[\tanh(kAX)X] - k \frac{\sqrt{4A^2 + 1} - 1}{2A} \right) / A^3 \\ F_2(A, k) &= \mathbb{E}[\tanh(kAX)X] - k \frac{\sqrt{4A^2 + 1} - 1}{2A} \end{aligned}$$

These $F_1(A, k), F_2(A, k)$ are continuous in A and k for any $A > 0, k > 0$, their limits as $A \rightarrow 0_+$ and $A \rightarrow \infty$ are as follows:

$$\begin{aligned} F_{1,0_+}(k) &= \lim_{A \rightarrow 0_+} F_1(A, k) = \lim_{A \rightarrow 0_+} 3k^3 - k + \mathcal{O}(A^2) = k(3k^2 - 1) \\ F_{2,\infty}(k) &= \lim_{A \rightarrow \infty} F_2(A, k) = \mathbb{E}[|X|] - k = \frac{2}{\pi} - k \end{aligned}$$

due to that $\tanh(t) = t - \frac{t^3}{3} + \mathcal{O}(t^5)$, $\frac{\sqrt{4A^2+1}-1}{2A} = A - A^3 + \mathcal{O}(A^5)$ and $\mathbb{E}[X^{2n}] = [(2n-1)!!]^2, \mathbb{E}[|X|] = \frac{2}{\pi}$ by Lemma E.3. We denote the unique positive root of $F_{1,0_+}(k) = 0$ as $k_{0_+}^*$ and the unique root of $F_{2,\infty}(k) = 0$ as k_∞^* .

$$k_{0_+}^* = \frac{1}{\sqrt{3}}, \quad k_\infty^* = \frac{2}{\pi}$$

We show that the limits of $k^*(A)$ as $A \rightarrow 0_+$ and $A \rightarrow \infty$ are $k_{0_+}^*$ and k_∞^* , respectively,

$$\lim_{\eta \rightarrow 0_+} k^*(\eta) = \lim_{A \rightarrow 0_+} k^*(A) = k_{0_+}^* = \frac{1}{\sqrt{3}}, \quad \lim_{\eta \rightarrow \infty} k^*(\eta) = \lim_{A \rightarrow \infty} k^*(A) = k_\infty^* = \frac{2}{\pi}$$

by checking the following conditions of $F_1(A, k), F_2(A, k)$ and $F_{1,0_+}(k), F_{2,\infty}(k)$ and applying Lemma E.5 for the rigorous analysis and substituting $A_0 \leftarrow 0_+, \infty, F \leftarrow F_1, F_2, k_{A_0}^* \leftarrow k_{0_+}^*, k_\infty^*$ in this lemma, respectively.

Conditions of $F_1(A, k), F_2(A, k)$ and $F_{1,0_+}(k), F_{2,\infty}(k)$: (1) $F_1(A, k), F_2(A, k)$ are continuous in k for any $A > 0$, since $\frac{\partial F_1(A, k)}{\partial k}$ and $\frac{\partial F_2(A, k)}{\partial k}$ exists for any $k, A > 0$. The root $k = k^*(A)$ of $F_1(A, k) = 0, F_2(A, k) = 0$ is unique for any $A > 0$. (2) $F_1(A, k), F_2(A, k)$ converge to $F_{1,0_+}(k), F_{2,\infty}(k)$ on compact interval $I := [\frac{1}{2}, 1]$ as $A \rightarrow 0_+$ and $A \rightarrow \infty$, respectively.

$$\sup_{k \in I} |F_1(A, k) - F_{1,0_+}(k)| \leq \sup_{k \in I} \max(|30k^5 A^2|, |2kA^2|) = 30A^2 < \varepsilon$$

for $0 < A < \sqrt{\varepsilon/30}, \forall \varepsilon > 0$, by using $t - \frac{1}{3}t^3 < \tanh(t) < t - \frac{t^3}{3} + \frac{2}{15}t^5, \forall t > 0, A - A^3 < \frac{\sqrt{4A^2+1}-1}{2A} < A - A^3 + 2A^5$ and $\mathbb{E}[X^{2n}] = [(2n-1)!!]^2$ in Lemma E.3. Similarly, we have

$$\begin{aligned} \sup_{k \in I} |F_2(A, k) - F_{2,\infty}(k)| &\leq \sup_{k \in I} |2\mathbb{E}[|X| \exp(-2kA|X|)]| + \left| \frac{k}{2}A^{-1} \right| < \frac{4}{\sqrt{2\pi}} \int_0^\infty \sqrt{x} \exp(-Ax) dx + \frac{1}{2}A^{-1} \\ &= \sqrt{2}A^{-\frac{3}{2}} + \frac{1}{2}A^{-1} \leq \frac{5}{2}A^{-1} < \varepsilon \end{aligned}$$

for $A > \max(\frac{1}{2}, \frac{5}{2\varepsilon}), \forall \varepsilon > 0$, by using $1 - 2\exp(-2t) < \tanh(t) < 1 - \exp(-2t), \forall t > 0, 1 - \frac{1}{2}A^{-1} < \frac{\sqrt{4A^2+1}-1}{2A} < 1 - \frac{1}{2}A^{-1} + \frac{1}{8}A^{-2} \leq 1 - \frac{1}{4}A^{-1}, \forall A \geq \frac{1}{2}$ and $K_0(x) < K_{\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2}} \frac{\exp(-x)}{\sqrt{x}}, \forall x > 0$ by the monotonicity property of K_ν with respect to $\nu \geq 0$ (see Section 10.37 Inequalities; Monotonicity; Section 10.39 Relation to Other Functions, Chapter 10 Bessel Function of [58]).

(3) The roots $k_{0_+}^* = \frac{1}{\sqrt{3}}, k_\infty^* = \frac{2}{\pi}$ of $F_{1,0_+}(k) = 0, F_{2,\infty}(k) = 0$ are the interior points of the compact interval $I = [\frac{1}{2}, 1]$, and $\frac{dF_{1,0_+}(k)}{dk} \big|_{k=k_{0_+}^*} = 2 \neq 0, \frac{dF_{2,\infty}(k)}{dk} \big|_{k=k_\infty^*} = -1 \neq 0$, and $\frac{dF_{1,0_+}(k)}{dk} = 9k^2 - 1, \frac{dF_{2,\infty}(k)}{dk} = -1$ are continuous in k on I . $\square$

E. EM Update Rule in the Noiseless Setting

Theorem A.8 (Corollary 8: EM Updates in Noiseless Setting, Corollary 3.3 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, the population EM update rules $M(\theta, \nu), N(\theta, \nu)$ for regression parameters θ and imbalance of mixing weights $\tanh \nu$ are:*

$$\begin{aligned} \frac{M(\theta, \nu)}{\|\theta^*\|} &= \frac{2}{\pi} \left[\text{sgn}(\rho) \varphi \frac{\theta^*}{\|\theta^*\|} + \cos \varphi \frac{\theta}{\|\theta\|} \right] \\ N(\theta, \nu) &= \text{sgn}(\rho) \frac{2}{\pi} \varphi \cdot \tanh \nu^* \end{aligned}$$

where $\varphi := \frac{\pi}{2} - \arccos |\rho|$, $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$.

Proof. We provide a completely new proof based on our previous newly established results rather than the old approach in [1]. Taking the limit $\eta \rightarrow +\infty$, then $\sqrt{1 + \eta^{-2}} \rightarrow 1$, $A_\eta \rightarrow +\infty$, $g_\eta \rightarrow g$ and $\varphi_\eta \rightarrow \varphi$, $\cos \varphi_\eta \rightarrow \cos \varphi$, $\sin \varphi_\eta \rightarrow \sin \varphi$, we have

$$\begin{aligned} \lim_{\eta \rightarrow \infty} M(\theta, \nu) / \|\theta^*\| &= \vec{e}_1 \mathbb{E}[gg'] + \vec{e}_2 \frac{1}{\cos \varphi} \text{sgn}(\rho) \mathbb{E}[g(g - \sin \varphi \cdot g') \text{sgn}(gg')] \\ &= \vec{e}_1 \mathbb{E}[gg'] + \vec{e}_2 \frac{\text{sgn}(\rho)}{\cos \varphi} (\mathbb{E}[g^2 \text{sgn}(gg')] - \sin \varphi \mathbb{E}[gg']) \\ \lim_{\eta \rightarrow \infty} N(\theta, \nu) &= \text{sgn}(\rho) \mathbb{E}[(-1)^{z+1}] \cdot \mathbb{E}[\text{sgn}(gg')] = \text{sgn}(\rho) \tanh \nu^* \mathbb{E}[\text{sgn}(gg')] \end{aligned}$$

By applying Lemma E.2, we have $\mathbb{E}[\text{sgn}(gg')] = \frac{2}{\pi} \varphi$, $\mathbb{E}[g^2 \text{sgn}(gg')] = \frac{2}{\pi} [\varphi + \sin \varphi \cos \varphi]$, and $\mathbb{E}[gg'] = \frac{2}{\pi} [\varphi \sin \varphi + \cos \varphi]$:

$$\begin{aligned} \lim_{\eta \rightarrow \infty} M(\theta, \nu) / \|\theta^*\| &= \frac{2}{\pi} [\varphi \sin \varphi + \cos \varphi] \vec{e}_1 + \text{sgn}(\rho) \frac{2}{\pi} \varphi \cos \varphi \vec{e}_2 \\ \lim_{\eta \rightarrow \infty} N(\theta, \nu) &= \text{sgn}(\rho) \frac{2}{\pi} \varphi \cdot \tanh \nu^* \end{aligned}$$

Noting that $\vec{e}_1 = \theta / \|\theta\|$, $\text{sgn}(\rho) \sin \varphi \vec{e}_1 + \cos \varphi \vec{e}_2 = \theta^* / \|\theta^*\|$, we have:

$$\lim_{\eta \rightarrow \infty} \frac{M(\theta, \nu)}{\|\theta^*\|} = \frac{2}{\pi} \left[\text{sgn}(\rho) \varphi \frac{\theta^*}{\|\theta^*\|} + \cos \varphi \frac{\theta}{\|\theta\|} \right]$$

□

F. Deviation from Cycloid Limit of EM Updates in High SNR Regime

Theorem A.9 (Proposition 9: Deviation from Cycloid Limit of EM Updates in High SNR Regime). *In the finite high SNR regime given by $\eta \gtrsim \frac{1}{\min(1, \sqrt{k}) \cos \varphi} \vee 1$, if the mixing weights are known balanced $\pi = \pi^* = (1/2, 1/2)$, then the difference between the EM update rule and its limit is bounded by:*

$$\left\| \frac{M(\theta)}{\|\theta^*\|} - \lim_{\eta \rightarrow \infty} \frac{M(\theta)}{\|\theta^*\|} \right\| = \mathcal{O} \left(\eta^{-2} \vee \cos^2 \varphi \frac{\log \Lambda}{\Lambda^4} \right)$$

where $\Lambda := \eta \sqrt{k} \cos \varphi$, $k := \frac{\|\theta\|}{\|\theta^*\|}$ and $\varphi := \frac{\pi}{2} - \arccos |\rho|$, $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$.

Proof. When the mixing weights are known balanced $\pi = \pi^* = (1/2, 1/2)$, we must have $\nu = \nu^* = 0$. Consequently, the EM update rules for the imbalance of mixing weights directly gives $N(\theta, \nu)_{\nu=0, \nu^*=0} = 0$ by using the symmetry of the data distributions with two balanced components. Furthermore, the EM update rule for regression parameters becomes

$$\frac{M(\theta)}{\|\theta^*\|} = \frac{M(\theta, \nu)_{\nu=0, \nu^*=0}}{\|\theta^*\|} = \vec{e}_1 \frac{\sin \varphi}{\sin \varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g') g_\eta g'] + \vec{e}_2 \frac{\cos \varphi}{\cos^2 \varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g')]$$

We further bound the difference between the EM update rule and its limit (the cycloid trajectory) when SNR goes to infinity.

$$\begin{aligned}
\frac{M(\theta)}{\|\theta^*\|} - \lim_{\eta \rightarrow \infty} \frac{M(\theta)}{\|\theta^*\|} &= \vec{e}_1 \left[\underbrace{\frac{\sin \varphi}{\sin \varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g') g_\eta g' - |g_\eta g'|]}_{T_{1,1}} + \underbrace{\mathbb{E} \left[\frac{\sin \varphi}{\sin \varphi_\eta} |g_\eta g'| - |g g'| \right]}_{T_{1,2}} \right] \\
&+ \vec{e}_2 \text{sgn}(\rho) \left[\underbrace{\frac{\cos \varphi}{\cos \varphi_\eta} \mathbb{E}[\tanh(A_\eta g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g') - \text{sgn}(g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g')]}_{T_{2,1}} \right] \\
&+ \vec{e}_2 \text{sgn}(\rho) \underbrace{\mathbb{E} \left[\frac{\cos \varphi}{\cos \varphi_\eta} \text{sgn}(g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g') / \cos \varphi_\eta - \text{sgn}(g g') g (g - \sin \varphi \cdot g') / \cos \varphi \right]}_{T_{2,2}}
\end{aligned}$$

Evaluation of the terms $T_{1,2}, T_{2,2}$: By applying the Lemma E.2 for $g_\eta, g, g' \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[g_\eta g'] = \sin \varphi_\eta$, $\mathbb{E}[g g'] = \sin \varphi$, we have $\mathbb{E}[|g_\eta g'|] = \frac{2}{\pi} [\varphi_\eta \sin \varphi_\eta + \cos \varphi_\eta]$, $\mathbb{E}[|g g'|] = \frac{2}{\pi} [\varphi \sin \varphi + \cos \varphi]$:

$$T_{1,2} = \mathbb{E} \left[\frac{\sin \varphi}{\sin \varphi_\eta} |g_\eta g'| - |g g'| \right] = \frac{\sin \varphi}{\sin \varphi_\eta} \frac{2}{\pi} [\varphi_\eta \sin \varphi_\eta + \cos \varphi_\eta] - \frac{2}{\pi} [\varphi \sin \varphi + \cos \varphi] = \frac{2}{\pi} \left[(\varphi_\eta - \varphi) \sin \varphi - \frac{\sin(\varphi_\eta - \varphi)}{\sin \varphi_\eta} \right]$$

Again, by applying the Lemma E.2 for g_η, g, g' , we have $\mathbb{E}[\text{sgn}(g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g') / \cos \varphi_\eta] = \frac{2}{\pi} \varphi_\eta \cos \varphi_\eta$ and $\mathbb{E}[\text{sgn}(g g') g (g - \sin \varphi \cdot g') / \cos \varphi] = \frac{2}{\pi} \varphi \cos \varphi$:

$$T_{2,2} = \mathbb{E} \left[\frac{\cos \varphi}{\cos \varphi_\eta} \text{sgn}(g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g') / \cos \varphi_\eta - \text{sgn}(g g') g (g - \sin \varphi \cdot g') / \cos \varphi \right] = \frac{2}{\pi} (\varphi_\eta - \varphi) \cos \varphi$$

We bound the following quantities to establish the asymptotic notations of $T_{1,1}, T_{2,1}$ in terms of SNR η and φ :

$$\begin{aligned}
0 \leq 1 - \frac{\sin \varphi_\eta}{\sin \varphi} &= 1 - (1 + \eta^{-2})^{-\frac{1}{2}} \leq \frac{\eta^{-2}}{2}, \quad 0 \leq \frac{\sin \varphi}{\sin \varphi_\eta} - 1 = \sqrt{1 + \eta^{-2}} - 1 \leq \frac{\eta^{-2}}{2} \\
0 \leq \varphi - \varphi_\eta &\leq \pi \sin \frac{\varphi - \varphi_\eta}{2} = \pi \frac{(1 - \frac{\sin \varphi_\eta}{\sin \varphi}) \sin \varphi}{2 \cos \frac{\varphi + \varphi_\eta}{2}} \leq \frac{\pi (\frac{\eta^{-2}}{2}) \sin \varphi}{2 \cos \varphi} = \frac{\pi \sin \varphi}{4 \cos \varphi} \eta^{-2} \\
&\leq \left| \frac{\sin(\varphi - \varphi_\eta)}{(\varphi - \varphi_\eta)} \cdot \frac{1}{\sin \varphi_\eta \sin \varphi} - 1 \right| \\
&\leq \frac{1}{\sin^2 \varphi} \left(\frac{\sin \varphi}{\sin \varphi_\eta} - 1 \right) \frac{\sin(\varphi_\eta - \varphi)}{(\varphi_\eta - \varphi)} + \frac{1}{\sin^2 \varphi} \left| \frac{\sin(\varphi_\eta - \varphi)}{(\varphi_\eta - \varphi)} - 1 \right| + \left(\frac{1}{\sin^2 \varphi} - 1 \right) \\
&\leq \frac{1}{\sin^2 \varphi} \left(\frac{\eta^{-2}}{2} + \frac{1}{6} \times \left[\frac{\pi \sin \varphi}{4 \cos \varphi} \eta^{-2} \right]^2 + \cos^2 \varphi \right) = \frac{\cos^2 \varphi}{\sin^2 \varphi} + \frac{\eta^{-2}}{\sin^2 \varphi} \mathcal{O} \left(1 \vee \frac{\sin^2 \varphi}{\cos^2 \varphi} \eta^{-2} \right)
\end{aligned}$$

Therefore, when $\eta \gtrsim \frac{1}{\cos \varphi}$, we have $\mathcal{O} \left(1 \vee \eta^{-2} \frac{\sin^2 \varphi}{\cos^2 \varphi} \right) = \mathcal{O}(1)$:

$$\begin{aligned}
|T_{1,2}| &= \frac{2}{\pi} \left| \frac{\sin(\varphi - \varphi_\eta)}{(\varphi - \varphi_\eta)} \frac{1}{\sin \varphi_\eta \sin \varphi} - 1 \right| (\varphi - \varphi_\eta) \sin \varphi \leq \frac{\cos \varphi \eta^{-2}}{2} + \frac{\eta^{-4}}{2 \cos \varphi} \mathcal{O} \left(1 \vee \frac{\eta^{-2} \sin^2 \varphi}{\cos^2 \varphi} \right) = \frac{\cos \varphi \eta^{-2}}{2} + \mathcal{O} \left(\frac{\eta^{-4}}{\cos \varphi} \right) \\
|T_{2,2}| &= \frac{2}{\pi} (\varphi - \varphi_\eta) \cos \varphi \leq \frac{\sin \varphi \eta^{-2}}{2}
\end{aligned}$$

Asymptotic notations of the terms $T_{1,1}, T_{2,1}$: Since we can express g_η, g' in terms of independent random variables R, U with $R \sim r \exp(-\frac{r^2}{2}) \mathbb{I}_{r \geq 0}$ (standard Rayleigh distribution with $\mathbb{E}[R^2] = 2$) and $U \sim \text{Unif}[0, 4\pi]$,

$$g_\eta = R \cos((U - \varphi_\eta)/2), \quad g' = R \sin((U + \varphi_\eta)/2)$$

Noting that $\varrho := R^2/2 \sim \text{Exp}(1)$ follows the standard exponential distribution with $\mathbb{E}[\varrho] = 1$, applying $g g' = \varrho [\sin U + \sin \varphi_\eta]$ and $\text{sgn}(g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g') / \cos \varphi_\eta = \varrho \text{sgn}(\sin(U + \varphi_\eta)) |\cos U + \cos \varphi_\eta|$, then:

$$\begin{aligned}
T_{1,1} &= \mathbb{E}[(\tanh(A_\eta |g_\eta g'|) - 1) |g_\eta g'|] = -2 \mathbb{E}_U \mathbb{E}_\varrho \left[\frac{|\sin U + \sin \varphi_\eta| \varrho}{\exp(2A_\eta |\sin U + \sin \varphi_\eta| \varrho) + 1} \right] \\
T_{2,1} &= \mathbb{E}[(\tanh(A_\eta |g_\eta g'|) - 1) \text{sgn}(g_\eta g') g_\eta (g_\eta - \sin \varphi_\eta \cdot g') / \cos \varphi_\eta] = -2 \mathbb{E}_U \mathbb{E}_\varrho \left[\frac{\text{sgn}(\sin(U + \varphi_\eta)) |\cos U + \cos \varphi_\eta| \varrho}{\exp(2A_\eta |\sin U + \sin \varphi_\eta| \varrho) + 1} \right]
\end{aligned}$$

Since $\exp(t) \leq \exp(t) + 1 \leq 2\exp(t)$ for all $t \geq 0$, and

$$\mathbb{E}_\varrho [\rho \exp(-2A_\eta |\sin U + \sin \varphi_\eta| \varrho)] = \frac{1}{(1 + 2A_\eta |\sin U + \sin \varphi_\eta|)^2}$$

we have shown that T_1, T_2 can be approximated by the following expectations upto some constant factors:

$$T_{1,1} \asymp \mathbb{E}_U \left[\frac{|\sin U + \sin \varphi_\eta|}{(1 + 2A_\eta |\sin U + \sin \varphi_\eta|)^2} \right], \quad T_{2,1} \asymp \mathbb{E}_U \left[\frac{\text{sgn}(\sin(U + \varphi_\eta)) |\cos U + \cos \varphi_\eta|}{(1 + 2A_\eta |\sin U + \sin \varphi_\eta|)^2} \right]$$

When $\Lambda' := \sqrt{A_\eta} \cos \varphi_\eta = \eta(1 + \eta^{-2})^{\frac{1}{4}} \sqrt{k} \cos \varphi_\eta \gtrsim 1$, applying Lemma E.1 for asymptotic notations of above expectations:

$$\begin{aligned} T_{1,1} &= \mathcal{O} \left(\frac{1}{A_\eta^2 \cos \varphi_\eta} \log(A_\eta \cos^2 \varphi_\eta) \right) = \cos^3 \varphi_\eta \mathcal{O} \left(\frac{\log \Lambda'}{[\Lambda']^4} \right) \\ T_{2,1} &= \mathcal{O} \left(\frac{\tan \varphi_\eta}{A_\eta^2 \cos \varphi_\eta} \log(A_\eta \cos^2 \varphi_\eta) \right) = \sin \varphi_\eta \cos^2 \varphi_\eta \mathcal{O} \left(\frac{\log \Lambda'}{[\Lambda']^4} \right) \end{aligned}$$

where the condition $\Lambda' \gtrsim 1$ is satisfied and therefore $\Lambda' \asymp \sqrt{k}(\eta \cos \varphi \vee 1)$ when SNR $\eta \gtrsim \frac{1}{\sqrt{k} \cos \varphi} \vee 1$ is large enough.

Bounds for the difference between the EM update rule and its limit: When SNR $\eta \gtrsim \frac{1}{\min(1, \sqrt{k}) \cos \varphi} \vee 1$ is large enough, and we set $\Lambda := \sqrt{k} \eta \cos \varphi \asymp \sqrt{k}(\eta \cos \varphi \vee 1) \asymp \Lambda'$, then $\sin \varphi / \sin \varphi_\eta = \sqrt{1 + \eta^{-2}} = \Theta(1)$, $\cos \varphi / \cos \varphi_\eta = \Theta(1)$:

$$\begin{aligned} & \left\| \frac{M(\theta)}{\|\theta^*\|} - \lim_{\eta \rightarrow \infty} \frac{M(\theta)}{\|\theta^*\|} \right\| = \left\| \left(\vec{e}_1 \frac{\sin \varphi}{\sin \varphi_\eta} T_{1,1} + \vec{e}_2 \frac{\cos \varphi}{\cos \varphi_\eta} T_{2,1} \right) + (\vec{e}_1 T_{1,2} + \vec{e}_2 \text{sgn}(\rho) T_{2,2}) \right\| \\ & \leq \|\vec{e}_1 \cos \varphi_\eta + \vec{e}_2 \sin \varphi_\eta\| \cos^2 \varphi \mathcal{O} \left(\frac{\log \Lambda}{\Lambda^4} \right) + \|\vec{e}_1 \cos \varphi + \vec{e}_2 \sin \varphi\| \frac{\eta^{-2}}{2} + \|\vec{e}_1\| \mathcal{O} \left(\frac{\eta^{-4}}{\cos \varphi} \right) \\ & = \frac{\eta^{-2}}{2} + \mathcal{O} \left(\frac{\eta^{-4}}{\cos \varphi} \vee \cos^2 \varphi \frac{\log \Lambda}{\Lambda^4} \right) \\ & = \mathcal{O} \left(\eta^{-2} \vee \cos^2 \varphi \frac{\log \Lambda}{\Lambda^4} \right) \end{aligned}$$

□

APPENDIX B

POPULATION LEVEL ANALYSIS IN NOISELESS REGIME

A. Recurrence Relation and Cycloid Trajectory of EM Iterations in Noiseless Setting

Theorem B.1 (Proposition 10: Recurrence Relation of Sub-optimality Angle, Proposition 4.3 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, if the sub-optimality angle $\varphi^t \neq \frac{\pi}{2}$, then the recurrence relation of the sub-optimality angle φ^t for population EM updates is:*

$$\tan \varphi^{t+1} = \tanh \varphi^t + \varphi^t (\tan^2 \varphi^t + 1)$$

where $\varphi^t := \frac{\pi}{2} - \arccos |\rho^t|$, $\rho^t := \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \|\theta^*\|}$.

Proof. The proof based on Corollary 3.3 of EM Updates in Noiseless Setting of [1] can be found on page 31, Appendix D of [1]. Let's start from the EM update rule for θ^t in previous Corollary, since ρ^0, ρ^{t-1} have the same sign (validated by checking the sign of $\langle \theta^t, \theta^* \rangle$)

$$\frac{\theta^t}{\|\theta^*\|} = \left(\frac{\pi}{2} \right)^{-1} \left[\varphi^{t-1} \frac{\text{sgn}(\rho^0) \theta^*}{\|\theta^*\|} + \cos \varphi^{t-1} \frac{\theta^{t-1}}{\|\theta^{t-1}\|} \right]$$

With $\left\langle \frac{\text{sgn}(\rho^0) \theta^*}{\|\theta^*\|}, \frac{\theta^{t-1}}{\|\theta^{t-1}\|} \right\rangle = |\rho^{t-1}| = \sin \varphi^{t-1}$

$$\begin{aligned} \sin \varphi^t \frac{\|\theta^t\|}{\|\theta^*\|} &= |\rho^t| \frac{\|\theta^t\|}{\|\theta^*\|} = \left\langle \frac{\theta^t}{\|\theta^*\|}, \frac{\text{sgn}(\rho^0) \theta^*}{\|\theta^*\|} \right\rangle = \left(\frac{\pi}{2} \right)^{-1} [\varphi^{t-1} + \cos \varphi^{t-1} \sin \varphi^{t-1}] \\ \frac{\|\theta^t\|}{\|\theta^*\|} &= \left(\frac{\pi}{2} \right)^{-1} \sqrt{[\varphi^{t-1}]^2 + \cos^2 \varphi^{t-1} + 2\varphi^{t-1} \cos \varphi^{t-1} \sin \varphi^{t-1}} \end{aligned}$$

Therefore

$$\sin \varphi^t = \frac{\varphi^{t-1} + \cos \varphi^{t-1} \sin \varphi^{t-1}}{\sqrt{[\varphi^{t-1}]^2 + \cos^2 \varphi^{t-1} + 2\varphi^{t-1} \cos \varphi^{t-1} \sin \varphi^{t-1}}}$$

Hence

$$\cos \varphi^t = \sqrt{1 - \sin^2 \varphi^t} = \frac{\cos^2 \varphi^{t-1}}{\sqrt{[\varphi^{t-1}]^2 + \cos^2 \varphi^{t-1} + 2\varphi^{t-1} \cos \varphi^{t-1} \sin \varphi^{t-1}}}$$

Thus, we obtain the recurrence relation for φ^t

$$\tan \varphi^t = \frac{\sin \varphi^t}{\cos \varphi^t} = \frac{\varphi^{t-1}}{\cos^2 \varphi^{t-1}} + \frac{\sin \varphi^{t-1}}{\cos \varphi^{t-1}} = \tan \varphi^{t-1} + \varphi^{t-1}[\tan^2 \varphi^{t-1} + 1]$$

□

Theorem B.2 (Proposition 11: Parametric Equation for Cycloid Trajectory of EM Updates, Proposition 4.4 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, the coordinates $\mathbf{x}^t, \mathbf{y}^t$ of normalized vector $\frac{\theta^t}{\|\theta^*\|} = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2 = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2^0, \forall t \in \mathbb{Z}_+$ for population EM updates can be parameterized by the sub-optimality angle ϕ^{t-1} as follows:*

$$\begin{aligned} 1 - \text{sgn}(\rho^0) \mathbf{x}^t &= \frac{1}{\pi} [\phi^{t-1} - \sin \phi^{t-1}] \\ \mathbf{y}^t &= \frac{1}{\pi} [1 - \cos \phi^{t-1}] \end{aligned}$$

where $\varphi^{t-1} := \frac{\pi}{2} - \arccos |\rho^{t-1}|$, $\rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \|\theta^*\|}$. Hence, the trajectory of EM iterations θ^t is on the cycloid with a parameter $\frac{\|\theta^*\|}{\pi}$, on the plane $\text{span}\{\theta^0, \theta^*\}$.

Proof. The proof based on Corollary 3.3 of EM Updates in Noiseless Setting and Proposition 4.3 of Recurrence Relation of [1] can be found on page 32, Appendix D of [1]. Since $\tan \varphi^t = \tan \varphi^{t-1} + \varphi^{t-1}(\tan^2 \varphi^{t-1} + 1)$ in the proven recurrence relation, it shows that $\tan \varphi^t \geq \tan \varphi^{t-1} \geq 0$, therefore $0 \leq \varphi^0 \leq \varphi^1 \leq \dots \leq \varphi^{t-1} \leq \varphi^t < \frac{\pi}{2}$.

Let $\hat{e}_1 := \frac{\theta^*}{\|\theta^*\|}$, and $\hat{e}_2 := \hat{e}_2 |_{\theta=\theta^t} = \frac{\theta - \hat{e}_1 \hat{e}_1^\top \theta}{\|\theta - \hat{e}_1 \hat{e}_1^\top \theta\|} |_{\theta=\theta^t} = \frac{\frac{\theta^t}{\|\theta^*\|} - [\rho^t] \frac{\theta^*}{\|\theta^*\|}}{\sqrt{1 - [\rho^t]^2}}$ and $\langle \hat{e}_1, \hat{e}_2^t \rangle = 0, \|\hat{e}_1\| = \|\hat{e}_2^t\| = 1$

$$\begin{aligned} \frac{(\frac{\pi}{2})}{\|\theta^*\| \|\theta^{t-1}\|} \langle \theta^{t-1} - \hat{e}_1 \hat{e}_1^\top \theta^{t-1}, \theta^t - \hat{e}_1 \hat{e}_1^\top \theta^t \rangle &= \frac{(\frac{\pi}{2})}{\|\theta^*\| \|\theta^{t-1}\|} \{ \langle \theta^{t-1}, \theta^t \rangle - \langle \theta^{t-1}, \hat{e}_1 \rangle \langle \hat{e}_1, \theta^t \rangle \} \\ &= [\varphi^{t-1} \sin \varphi^{t-1} + \cos \varphi^{t-1}] - \sin \varphi^{t-1} [\varphi^{t-1} + \cos \varphi^{t-1} \sin \varphi^{t-1}] \\ &= \cos^3 \varphi^{t-1} > 0 \end{aligned}$$

Hence, we conclude that $\langle \hat{e}_2^{t-1}, \hat{e}_2^t \rangle > 0$, With $\hat{e}_2^{t-1}, \hat{e}_2^t \perp \hat{e}_1$ and $\hat{e}_2^{t-1}, \hat{e}_2^t \in \text{span}\{\theta^t, \theta^{t-1}, \theta^*\} \subset \text{span}\{\theta^0, \theta^*\}$, $\|\hat{e}_2^{t-1}\| = \|\hat{e}_2^t\| = 1$, we validate $\hat{e}_2^0 = \dots = \hat{e}_2^{t-1} = \hat{e}_2^t$.

By the definition of $\hat{e}_2^t$, we obtain $\theta^t = \|\theta^t\| \{ \text{sgn}(\rho^0) \sin \varphi^t \hat{e}_1 + \cos \varphi^t \hat{e}_2^t \} = \|\theta^t\| \{ \text{sgn}(\rho^0) \sin \varphi^t \hat{e}_1 + \cos \varphi^t \hat{e}_2^0 \}$

Since $\theta^t \in \text{span}\{\theta^{t-1}, \theta^*\}$, then $\theta^t \in \text{span}\{\theta^0, \theta^*\}$, we can express $\frac{\theta^t}{\|\theta^*\|} = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2^t = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2^0$.

Comparing the expressions for θ^t , we derive the following result.

$$\frac{\theta^t}{\|\theta^*\|} = \mathbf{x}^t \hat{e}_1 + \mathbf{y}^t \hat{e}_2^0 = \left\{ \text{sgn}(\rho^0) \sin \varphi^t \cdot \frac{\|\theta^t\|}{\|\theta^*\|} \right\} \hat{e}_1 + \left\{ \cos \varphi^t \cdot \frac{\|\theta^t\|}{\|\theta^*\|} \right\} \hat{e}_2^0$$

With the recurrence relation $\sin \varphi^t \cdot \frac{\|\theta^t\|}{\|\theta^*\|} = (\frac{\pi}{2})^{-1} [\varphi^{t-1} + \cos \varphi^{t-1} \sin \varphi^{t-1}]$, $\cos \varphi^t \cdot \frac{\|\theta^t\|}{\|\theta^*\|} = (\frac{\pi}{2})^{-1} \cos^2 \varphi^{t-1}$, which we showed in the proof of Recurrence Relation, we derive the implicit equation of $\mathbf{x}^t, \mathbf{y}^t (t \geq 1)$ as follows:

$$\begin{aligned} \mathbf{x}^t &= \left\langle \frac{\theta^t}{\|\theta^*\|}, \hat{e}_1 \right\rangle = \left(\frac{\pi}{2} \right)^{-1} \text{sgn}(\rho^0) [\varphi^{t-1} + \cos \varphi^{t-1} \sin \varphi^{t-1}] \\ \mathbf{y}^t &= \left\langle \frac{\theta^t}{\|\theta^*\|}, \hat{e}_2^0 \right\rangle = \left\langle \frac{\theta^t}{\|\theta^*\|}, \hat{e}_2^t \right\rangle = \left(\frac{\pi}{2} \right)^{-1} \cos^2 \varphi^{t-1} \end{aligned}$$

Let's cancel out the parameter φ^{t-1} in the parameterized curve $\varphi^{t-1} \mapsto (\mathbf{x}^t, \mathbf{y}^t)$

$$\text{sgn}(\rho^0) \frac{\pi}{2} \mathbf{x}^t = \sqrt{\left(\frac{\pi}{2} \mathbf{y}^t \right) \left(1 - \frac{\pi}{2} \mathbf{y}^t \right)} + \arccos \sqrt{\frac{\pi}{2} \mathbf{y}^t}$$

Let $\phi := 2 \left(\frac{\pi}{2} - \varphi \right) \in (0, \pi]$, then we rewrite the implicit equations of $\mathbf{x}^t, \mathbf{y}^t (t \geq 1)$ as follows, which is a cycloid curve [59]:

$$\begin{aligned} 1 - \text{sgn}(\rho^0) \mathbf{x}^t &= \pi^{-1} [\phi - \sin \phi]_{\phi=\phi^{t-1}} \\ \mathbf{y}^t &= \pi^{-1} [1 - \cos \phi]_{\phi=\phi^{t-1}} \end{aligned}$$

□

B. Linear Growth and Quadratic Convergence of Sub-optimality Angles in Noiseless Setting

Theorem B.3 (Proposition 12: Linear Growth of Sub-optimality Angle). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, $\tan \varphi^t$ of the sub-optimality angle φ^t grows at least linearly:*

$$\tan \varphi^{t+1} \geq 2 \cdot \tan \varphi^t$$

where $\varphi^t := \frac{\pi}{2} - \arccos |\rho^t|$, $\rho^t := \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \|\theta^*\|}$.

Proof. We start from the recurrence relation of φ^t for the population EM update rule:

$$\tan \varphi^{t+1} = \tan \varphi^t + \varphi^t (\tan^2 \varphi^t + 1)$$

The inequality $\tan \varphi^{t+1} \geq 2 \cdot \tan \varphi^t$ holds when $\varphi^t = 0$. Let's consider the case when $\varphi^t > 0$. By the trigonometry identity $\sin(2\varphi^t) = 2 \tan \varphi^t / (1 + \tan^2 \varphi^t)$, and the inequality $2\varphi^t \geq \sin(2\varphi^t)$, we have:

$$\tan \varphi^{t+1} = (1 + \frac{2\varphi^t}{\sin(2\varphi^t)}) \tan \varphi^t \geq 2 \cdot \tan \varphi^t$$

□

Theorem B.4 (Proposition 13: Quadratic Convergence of Sub-optimality Angle). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, the sub-optimality angle ϕ^t converges quadratically to zero when $\phi^t \leq 1.4$ is small enough:*

$$\frac{\phi^{t+1}}{\pi} \leq \left[\frac{\phi^t}{\pi} \right]^2$$

where $\phi^t := 2 \arccos |\rho^t|$, $\rho^t := \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \|\theta^*\|}$.

Proof. If $\phi^t = 0$, then $\phi^{t+1} = 0$ by the EM updates in noiseless setting. Otherwise, we have $\phi^{t+1} > 0$ when $\phi^t > 0$ from the EM updates in noiseless setting. We start from the recursive relation of ϕ^t for the population EM update rule:

$$\tan \varphi^{t+1} = \tanh \varphi^t + \varphi^t (\tan^2 \varphi^t + 1)$$

By the definition $\phi^t = 2(\frac{\pi}{2} - \varphi^t)$ and apply the inequality $x \leq \tan x, \forall x \in [0, \frac{\pi}{2})$, we have:

$$\frac{1}{\tan \frac{\phi^{t+1}}{2}} = \frac{1}{\tan \frac{\phi^t}{2}} + \frac{\pi - \phi^t}{2} \cdot \frac{1}{\sin^2 \frac{\phi^t}{2}} \implies \frac{\phi^{t+1}}{\pi} \leq \frac{2}{\pi} \tan \frac{\phi^{t+1}}{2} = \left[\frac{\phi^t}{\pi} \right]^2 \cdot \frac{2\pi(1 - \cos \phi^t)/[\phi^t]^2}{\sin \phi^t + (\pi - \phi^t)}$$

By the inequality based on Taylor's theorem $\cos x \geq 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720}$, $\sin x \geq x - \frac{x^3}{6}, \forall x \geq 0$, we have:

$$\frac{2\pi(1 - \cos \phi^t)}{[\phi^t]^2} \leq \pi - \frac{\pi}{12}[\phi^t]^2 + \frac{\pi}{360}[\phi^t]^4 \leq \pi - \frac{1}{6}[\phi^t]^3 \leq \sin \phi^t + (\pi - \phi^t)$$

where the second inequality is due to $30 - \frac{60}{\pi}\phi^t - [\phi^t]^2 \geq 0$ when $0 \leq \phi^t \leq 1.4$. Therefore, when $0 \leq \phi^0 \leq 1.4$, we have $0 \leq \phi^{t+1} \leq \phi^t \leq \dots \leq \phi^0 \leq 1.4$ and

$$\frac{\phi^{t+1}}{\pi} \leq \left[\frac{\phi^t}{\pi} \right]^2$$

□

C. Accuracy of Regression Parameters and Mixing Weights in Noiseless Setting

Theorem B.5 (Proposition 14: Accuracy of EM Updates and Sub-optimality Angle). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$,*

$$\begin{aligned} \frac{\|\theta^t - \text{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} &\leq \frac{[\phi^{t-1}]^2}{2\pi} \\ \|\pi^t - \bar{\pi}^*\|_1 &= \frac{\phi^{t-1}}{\pi} \cdot \left\| \pi^* - \frac{\mathbf{1}}{2} \right\|_1 \end{aligned}$$

where $\phi^{t-1} := 2(\frac{\pi}{2} - \arccos |\rho^{t-1}|)$, $\rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \|\theta^*\|}$, and $\bar{\pi}^* := \frac{1}{2} + \text{sgn}(\rho^0)(\pi^* - \frac{1}{2})$, $\mathbf{1} := (1, 1)$.

Proof. Let's prove the first inequality by starting from the equations of Cycloid Trajectory parameterized by the angle ϕ^{t-1} .

$$\begin{aligned} \frac{\|\theta^t - \text{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} &= \frac{1}{\pi} \left\| [\phi^{t-1} - \sin \phi^{t-1}] \hat{e}_1 + [1 - \cos \phi^{t-1}] \hat{e}_2^0 \right\| = \frac{1}{\pi} \left\| \int_0^{\phi^{t-1}} ([\phi - \cos \phi] \hat{e}_1 + [\sin \phi] \hat{e}_2^0) d\phi \right\| \\ &\leq \frac{1}{\pi} \int_0^{\phi^{t-1}} \left\| [\phi - \cos \phi] \hat{e}_1 + [\sin \phi] \hat{e}_2^0 \right\| d\phi = \frac{1}{\pi} \int_0^{\phi^{t-1}} 2 \sin \frac{\phi}{2} d\phi = \frac{4}{\pi} (1 - \cos \frac{\phi^{t-1}}{2}) \\ &\leq \frac{[\phi^{t-1}]^2}{2\pi} \end{aligned}$$

In the above derivation, the first inequality is due to the fact that $\|\int_0^{\phi^t} \vec{v}(\phi) d\phi\| \leq \int_0^{\phi^t} \|\vec{v}(\phi)\| d\phi$ for vector $\vec{v}(\phi)$ and $\phi^{t-1} \geq 0$, and the second inequality follows from the inequality $1 - \cos x \leq \frac{x^2}{2}$.

The proof for the accuracy in mixing weights (the identity in the second line) is adapted from page 34, Appendix D of [1]. Using EM Updates in Noiseless Setting, and note that $\text{sgn}(\rho^{t-1}) = \text{sgn}(\rho^0)$, we obtain that equation.

$$\tanh(\nu^t) = \text{sgn}(\rho^0) \left(\frac{2}{\pi} \varphi^{t-1} \right) \cdot \tanh(\nu^*)$$

Since $\pi^t(1) = \frac{1+\tanh(\nu^t)}{2}$, $\pi^t(2) = \frac{1-\tanh(\nu^t)}{2}$ and $\bar{\pi}^*(1) = \frac{1+\text{sgn}(\rho^0)\tanh(\nu^*)}{2}$, $\bar{\pi}^*(2) = \frac{1-\text{sgn}(\rho^0)\tanh(\nu^*)}{2}$.

$$\|\pi^t - \bar{\pi}^*\|_1 = |\pi^t(1) - \bar{\pi}^*(1)| + |\pi^t(2) - \bar{\pi}^*(2)| = |\tanh(\nu^t) - \text{sgn}(\rho^0)\tanh(\nu^*)| = \left| 1 - \frac{2}{\pi} \varphi^{t-1} \right| \cdot \left\| \pi^* - \frac{1}{2} \right\|_1 = \frac{\phi^{t-1}}{\pi} \cdot \left\| \pi^* - \frac{1}{2} \right\|_1$$

In the above equation, we use such an identity $\tanh(\nu^*) = \left\| \pi^* - \frac{1}{2} \right\|_1$ and the fact that $\phi^{t-1} = \pi - 2\varphi^{t-1}$. $\square$

D. Convergence Guarantee of EM Updates at the Population Level in Noiseless Setting

Theorem B.6 (Theorem 15: Population Level Convergence, Theorem 4.1 in [1]). *In the noiseless setting, namely SNR $\eta := \|\theta^*\|/\sigma \rightarrow \infty$, if the initial sup-optimality angle cosine $\rho^0 := \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \|\theta^*\|} \neq 0$, then with the number of total iterations at most $T = \mathcal{O}(\log \frac{1}{|\rho^0|} \vee \log \log \frac{1}{\varepsilon})$, the error of EM update at the population level is bounded by: $\frac{\|\theta^{T+1} - \text{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} < \varepsilon$ and $\|\pi^{T+1} - \bar{\pi}^*\|_1 = \left\| \pi^* - \frac{1}{2} \right\|_1 \mathcal{O}(\sqrt{\varepsilon})$. where $\bar{\pi}^* := \frac{1}{2} + \text{sgn}(\rho^0)(\pi^* - \frac{1}{2})$, $\mathbf{1} := (1, 1)$.*

Proof. We provide a new straightforward proof for the convergence guarantee of EM updates at the population level in the noiseless setting, by invoking the results for accuracy in regression parameters and mixing weights, and applying the linear growth and quadratic convergence of the sub-optimality angles. The other old lengthy proof of this Theorem 4.1 of Population Level Convergence in [1] can be found on pages 34-35, Appendix D of [1].

Since when the sub-optimality angle $\phi^t > 1.4$ is large, the other sub-optimality angle $\varphi^t = \frac{\pi}{2} - \frac{\phi^t}{2} \leq \frac{\pi}{2} - 0.7 \approx 0.8708 < 1$ is also small enough. Suppose the initial sub-optimality angle $\varphi^0 < 1$, then after at most $T_1 := \lceil (\ln \tan 1 + \ln \frac{1}{|\rho^0|}) / \ln 2 \rceil_+ = \Theta(\log \frac{1}{|\rho^0|})$ EM iterations, the sub-optimality angle $\phi^t \leq 1.4, \forall t \geq T_1$ will be sufficiently small enough.

$$\begin{aligned} \tan \varphi^t &\geq \tan \varphi^{T_1} \geq 2^{T_1} \tan \varphi^0 \geq \tan 1 \cdot \frac{1}{|\rho^0|} \cdot \tan \varphi^0 = \tan 1 \cdot \frac{1}{\sin \varphi^0} \cdot \frac{\sin \varphi^0}{\cos \varphi^0} \geq \tan 1 \quad \forall t \geq T_1 \\ \implies \varphi^t &\geq 1 \implies \phi^t = 2 \left(\frac{\pi}{2} - \varphi^t \right) < 1.4 \quad \forall t \geq T_1 \end{aligned}$$

Otherwise, if the initial sub-optimality angle $\varphi^0 \geq 1$, then $T_1 = 0$ and we still have $\phi^t < 1.4, \forall t \geq T_1$. Then by the quadratic convergence of the sub-optimality angle ϕ^t when $\phi^t < 1.4$ is small enough, then after at most $T_2 := \lceil (\ln \ln \frac{\pi}{2\varepsilon} - \ln 2 - \ln \ln \frac{\pi}{1.4}) / \ln 2 \rceil_+ = \Theta(\log \log \frac{1}{\varepsilon})$, we have the following inequality:

$$\phi^{T_1+T_2} \leq \pi \left(\frac{\phi^{T_1}}{\pi} \right)^{2^{T_2}} < \pi \left(\frac{1.4}{\pi} \right)^{2^{T_2}} \leq \pi \left(\frac{1.4}{\pi} \right)^{\frac{\frac{1}{2} \ln \frac{\pi}{2\varepsilon}}{\ln \frac{\pi}{1.4}}} = \pi \left(\frac{\pi}{2\varepsilon} \right)^{-\frac{1}{2}} = \sqrt{2\pi\varepsilon}$$

Therefore, after at most $T := T_1 + T_2 = \Theta(\log \frac{1}{|\rho^0|} \vee \log \log \frac{1}{\varepsilon})$ EM iterations, we have $\phi^t < \sqrt{2\pi\varepsilon}$ for all $t \geq T$. By using the results for accuracy in regression parameters and mixing weights, we have the following bounds:

$$\frac{\|\theta^{T+1} - \text{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} \leq \frac{[\phi^T]^2}{2\pi} < \varepsilon, \quad \|\pi^{T+1} - \bar{\pi}^*\|_1 = \frac{\phi^T}{\pi} \cdot \left\| \pi^* - \frac{1}{2} \right\|_1 < \sqrt{\frac{2\varepsilon}{\pi}} \cdot \left\| \pi^* - \frac{1}{2} \right\|_1 = \mathcal{O}(\sqrt{\varepsilon}) \cdot \left\| \pi^* - \frac{1}{2} \right\|_1$$

where the last inequality of both bounds follows from $\phi^T < \sqrt{2\pi\varepsilon}$. $\square$

APPENDIX C
FINITE-SAMPLE LEVEL ANALYSIS IN NOISELESS SETTING

A. Statistical Error, Statistical Accuracy of EM Updates in Noiseless Setting

Theorem C.1 (Proposition 16: Projected Error of Easy EM Update for Regression Parameters, Proposition 5.2 in [1]). *In the noiseless setting, the projection on $\text{span}\{\theta, \theta^*\}$ for the statistical error of θ satisfies*

$$\frac{\|P_{\theta, \theta^*}[M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)]\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n}\right),$$

with probability at least $1 - \delta$, where $M_n(\theta, \nu)$, $M(\theta, \nu)$ are the EM update rules for θ at the Finite-sample level and the population level respectively, and the orthogonal projection matrix P_{θ, θ^*} satisfies $\text{span}(P_{\theta, \theta^*}) = \text{span}\{\theta, \theta^*\}$.

Proof. We provide a new straightforward proof here. In the noiseless setting, as $\text{SNR } \eta = \frac{\|\theta^*\|}{\sigma} \rightarrow \infty$, we have $y_i \rightarrow (-1)^{z_i+1}\langle x_i, \theta^* \rangle$, $\tanh(y_i \langle x_i, \theta \rangle / \sigma^2 + \nu) \rightarrow (-1)^{z_i+1} \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle)$ for the i -th data point $z_i, s_i = (x_i, y_i)$.

$$M_n^{\text{easy}}(\theta, \nu) = \frac{1}{n} \sum_{i=1}^n \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle) \langle x_i, \theta^* \rangle x_i$$

Without loss of generality, we assume $\dim \text{span}\{\theta, \theta^*\} = 2$ and therefore $\dim \text{span}\{\theta, \theta^*\}^\perp = d - 2$.

We introduce $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$, and define $\hat{e}_1 := \frac{\theta^*}{\|\theta^*\|}$, $\vec{e}_1 := \frac{\theta}{\|\theta\|}$, and $\hat{e}_2 := \frac{\theta - \langle \theta, \hat{e}_1 \rangle \hat{e}_1}{\|\theta - \langle \theta, \hat{e}_1 \rangle \hat{e}_1\|}$, $\vec{e}_2 := \frac{\theta^* - \langle \theta^*, \vec{e}_1 \rangle \vec{e}_1}{\|\theta^* - \langle \theta^*, \vec{e}_1 \rangle \vec{e}_1\|}$. Also, let $P_{\theta, \theta^*} = \vec{e}_1 \vec{e}_1^\top + \vec{e}_2 \vec{e}_2^\top$, which can be expressed with orthonormal basis matrices $U_{\theta, \theta^*} = (\text{sgn}(\rho) \vec{e}_1, \vec{e}_2)$ such that $P_{\theta, \theta^*} = U_{\theta, \theta^*} U_{\theta, \theta^*}^\top$. Also, we have $\varphi = \frac{\pi}{2} - \arccos |\rho|$, therefore $\rho = \text{sgn}(\rho) \sin \varphi$. By defining $g_i := \langle x_i, \hat{e}_1 \rangle$, $g'_i := \text{sgn}(\rho) \langle x_i, \vec{e}_1 \rangle$, then $g_i, g'_i \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[g_i g'_i] = \sin \varphi$, and $h'_i = \langle x_i, \vec{e}_2 \rangle \sim \mathcal{N}(0, 1)$ satisfies $g_i = \sin \varphi \cdot g'_i + \cos \varphi \cdot h'_i$, $h'_i \perp g'_i$.

$$U_{\theta, \theta^*}^\top \frac{M_n^{\text{easy}}(\theta, \nu)}{\|\theta^*\|} = \left(\frac{1}{n} \sum_{i=1}^n |g_i g'_i|, \text{sgn}(\rho) \frac{1}{n} \sum_{i=1}^n \text{sgn}(g_i g'_i) g_i h'_i \right)^\top$$

$$\begin{aligned} \frac{\|P_{\theta, \theta^*}[M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)]\|}{\|\theta^*\|} &= \left\| U_{\theta, \theta^*}^\top \frac{[M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)]}{\|\theta^*\|} \right\| = \left\| U_{\theta, \theta^*}^\top \frac{M_n^{\text{easy}}(\theta, \nu)}{\|\theta^*\|} - \mathbb{E} \left[U_{\theta, \theta^*}^\top \frac{M_n^{\text{easy}}(\theta, \nu)}{\|\theta^*\|} \right] \right\| \\ &= \sqrt{\left[\left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) |g_i g'_i| \right]^2 + \left[\left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) \text{sgn}(g_i g'_i) g_i h'_i \right]^2} \end{aligned}$$

where $|g_i g'_i|$, $\text{sgn}(g_i g'_i) g_i h'_i$ are i.i.d. Sub-Exponential random variables by Proposition 2.7.1 part (b) on page 33 of [52] and noting that for $\forall q \in \mathbb{Z}_+$,

$$\mathbb{E}[|g_i g'_i| - \mathbb{E}[|g_i g'_i|]]^q]^{1/q} \leq \mathbb{E}[|g_i g'_i|^q]^{1/q} + \mathbb{E}[|g_i g'_i|] \leq \mathbb{E} \left[\frac{g_i^2 + (g'_i)^2}{2} \right]^{1/q} + \mathbb{E} \left[\frac{g_i^2 + (g'_i)^2}{2} \right] \leq 2eq + 1 \leq 3eq,$$

$$\mathbb{E}[(\text{sgn}(g_i g'_i) g_i h'_i - \mathbb{E}[\text{sgn}(g_i g'_i) g_i h'_i])^q]^{1/q} \leq \mathbb{E}[|g_i h'_i|^q]^{1/q} + \mathbb{E}[\text{sgn}(g_i g'_i) g_i h'_i] \leq \mathbb{E}[|g_i h'_i|^q]^{1/q} + \mathbb{E}[|g_i h'_i|] \leq 2eq + 1 \leq 3eq,$$

where the first inequality follows from Minkowski's inequality, and the second last inequality is due to the facts for moment of Gaussian random variables $g_i, g'_i, h'_i \sim \mathcal{N}(0, 1)$. Consequently, by Bernstein's inequality (see Corollary 2.8.3 on page 38 of [52]) for Sub-Exponential random variables $|g_i g'_i|$, $\text{sgn}(g_i g'_i) g_i h'_i$, with probability at least $1 - \delta$, we have

$$\frac{\|P_{\theta, \theta^*}[M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)]\|}{\|\theta^*\|} \leq \left| \left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) |g_i g'_i| \right| + \left| \left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) \text{sgn}(g_i g'_i) g_i h'_i \right| \lesssim \sqrt{\frac{\log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n}$$

□

Theorem C.2 (Proposition 17: Statistical Error of EM Update for Regression Parameters, Proposition 5.3 in [1]). *In the noiseless setting, the statistical error of θ for finite-sample EM updates with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples satisfies*

$$\frac{\|M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}\right), \quad \frac{\|M_n(\theta, \nu) - M(\theta, \nu)\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}\right),$$

with probability at least $1 - \delta$, $M_n(\theta, \nu)$, $M(\theta, \nu)$ denote the EM update rules for θ at the Finite-sample level and the Population level.

Proof. Let the ensemble matrix of data samples be $\Sigma = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top$, and note that $M(\theta, \nu) = \mathbb{E}[M_n^{\text{easy}}(\theta, \nu)]$, by letting $\vec{v} = (M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)) / \|\theta^*\| = (M_n^{\text{easy}}(\theta, \nu) - \mathbb{E}[M_n^{\text{easy}}(\theta, \nu)]) / \|\theta^*\|$, we have

$$\frac{M_n(\theta, \nu) - M(\theta, \nu)}{\|\theta^*\|} = \Sigma^{-1} \vec{v} - \Sigma^{-1}(\Sigma - I_d) \frac{M(\theta, \nu)}{\|\theta^*\|}$$

Without loss of generality, we assume $\dim \text{span}\{\theta, \theta^*\} = 2$ and therefore $\dim \text{span}\{\theta, \theta^*\}^\perp = d-2$. Let $P_{\theta, \theta^*}^\perp = \sum_{j=1}^{d-2} \vec{e}_j^\perp (\vec{e}_j^\perp)^\top$ and $U_{\theta, \theta^*}^\perp = (\vec{e}_1^\perp, \dots, \vec{e}_{d-2}^\perp)$ with $\text{span}(U_{\theta, \theta^*}^\perp) = \text{span}\{\theta, \theta^*\}^\perp$. For the length of vector $\vec{v}$, noting that orthogonal projection matrix $P_{\theta, \theta^*}^\perp$ can be expressed with orthonormal basis matrix $U_{\theta, \theta^*}^\perp$ such that $P_{\theta, \theta^*}^\perp = U_{\theta, \theta^*}^\perp (U_{\theta, \theta^*}^\perp)^\top$, we have

$$\|\vec{v}\|^2 = \langle (P_{\theta, \theta^*} + P_{\theta, \theta^*}^\perp) \vec{v}, (P_{\theta, \theta^*} + P_{\theta, \theta^*}^\perp) \vec{v} \rangle = \langle \vec{v}, (P_{\theta, \theta^*} + P_{\theta, \theta^*}^\perp) \vec{v} \rangle = \langle \vec{v}, P_{\theta, \theta^*} \vec{v} \rangle + \langle \vec{v}, P_{\theta, \theta^*}^\perp \vec{v} \rangle = \|P_{\theta, \theta^*} \vec{v}\|^2 + \|(U_{\theta, \theta^*}^\perp)^\top \vec{v}\|^2$$

where the first term is bounded by $\|P_{\theta, \theta^*} \vec{v}\| \lesssim \sqrt{\frac{\log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n}$, and we rewrite the second term by introducing independent Gaussian variables $g_i := \langle x_i, \theta^* / \|\theta^*\| \rangle \sim \mathcal{N}(0, 1)$, $\vec{g}_i := \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle) (U_{\theta, \theta^*}^\perp)^\top x_i \sim \mathcal{N}(0, I_{d-2})$ with $g_i \perp \vec{g}_i$.

$$\|(U_{\theta, \theta^*}^\perp)^\top \vec{v}\|^2 = \frac{1}{n^2} \left\| \sum_{i=1}^n g_i \vec{g}_i \right\|^2 = \frac{1}{n^2} \left(\sum_{i=1}^n g_i^2 \right) \cdot \left\| \sum_{i=1}^n \frac{g_i}{\sqrt{\sum_{i'=1}^n (g_{i'}^2)}} \vec{g}_i \right\|^2 = \frac{1}{n^2} Z_1 \cdot Z_2$$

where $Z_1 := \sum_{i=1}^n g_i^2 \sim \chi^2(n)$ and the weighted sum $\sum_{i=1}^n (g_i / \sqrt{\sum_{i'=1}^n (g_{i'}^2)}) \vec{g}_i \sim \mathcal{N}(0, I_{d-2})$ by the rotational invariance of Gaussian distribution, therefore $Z_2 = \|\sum_{i=1}^n (g_i / \sqrt{\sum_{i'=1}^n (g_{i'}^2)}) \vec{g}_i\|^2 \sim \chi^2(d-2)$ with $Z_1 \perp Z_2$. By the concentration inequality for Chi-square distribution (see Lemma 1, page 1325 in [53]), given $n \geq d$, with probability at least $1 - \delta/2$,

$$\|\vec{v}\| \leq \|P_{\theta, \theta^*} \vec{v}\| + \frac{1}{n} \sqrt{Z_1 \cdot Z_2} \lesssim \sqrt{\frac{\log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n} + \frac{1}{n} \sqrt{\left(n \vee \log \frac{1}{\delta}\right) \left(d \vee \log \frac{1}{\delta}\right)} \asymp \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n}$$

From the assumption $n \gtrsim d \vee \log \frac{1}{\delta}$, by Example 6.2 (Operator norm bounds for the standard Gaussian ensemble) and Example 6.3 (Gaussian covariance estimation) on page 162 of [60], with probability at least $1 - \delta/2$, we have

$$\gamma_{\min}(\Sigma) \asymp 1, \quad \|\Sigma - I_d\|_2 \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$$

where $\gamma_{\min}$ represents the minimum eigenvalue, and $\|\cdot\|_2$ is the ℓ_2 -operator norm for matrix. Noting that $\sqrt{\frac{\log \frac{1}{\delta}}{n}} \gtrsim \frac{\log \frac{1}{\delta}}{n}$ when $n \gtrsim d \vee \log \frac{1}{\delta}$ and $\|M(\theta, \nu)\| / \|\theta^*\| \leq 1$ in the noiseless setting, and putting the above results together, we have

$$\frac{\|M_n(\theta, \nu) - M(\theta, \nu)\|}{\|\theta^*\|} \leq \frac{\|\vec{v}\|}{\gamma_{\min}(\Sigma)} + \frac{\|\Sigma - I_d\|_2}{\gamma_{\min}(\Sigma)} \frac{\|M(\theta, \nu)\|}{\|\theta^*\|} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$$

□

Theorem C.3 (Proposition 18: Statistical Error of EM Update for Mixing Weights). *In the noiseless setting, the statistical error of mixing weights for finite-sample EM updates satisfies*

$$|N_n(\theta, \nu) - N(\theta, \nu)| = \mathcal{O} \left(\frac{\log \frac{1}{\delta}/n}{\log \left(1 + \frac{\log \frac{1}{\delta}/n}{p}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$$

with probability at least $1 - \delta$, where $N_n(\theta, \nu), N(\theta, \nu)$ denote the EM update rules for imbalance $\tanh \nu$ of the mixing weights at the Finite-sample level and the Population level, and $p := \|\pi^* - \frac{1}{2}\|_{1, \frac{\phi}{2\pi}} + \min(\pi^*(1), \pi^*(2))$, $\phi = 2 \arccos |\rho|$, $\rho = \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$.

Proof. In the noiseless setting, as SNR $\eta \rightarrow \infty$, we have $y_i \rightarrow (-1)^{z_i+1} \langle x_i, \theta^* \rangle$ and therefore $\tanh(y_i \langle x_i, \theta \rangle / \sigma^2 + \nu) \rightarrow (-1)^{z_i+1} \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle)$ for the i -th data point $z_i, s_i = (x_i, y_i)$.

$$N_n(\theta, \nu) = \frac{1}{n} \sum_{i=1}^n (-1)^{z_i+1} \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle)$$

Let $\rho = \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$, and $\varphi = \frac{\pi}{2} - \arccos |\rho|$, $\phi = 2 \arccos |\rho|$, therefore $\rho = \text{sgn}(\rho) \sin \varphi$. By defining $g_i = \langle x_i, \theta^* / \|\theta^*\| \rangle$, $g'_i = \langle x_i, \text{sgn}(\rho) \theta / \|\theta\| \rangle$, and introducing $V_i = (1 + \text{sgn}(\nu^*)) (-1)^{z_i+1} \text{sgn}(g_i g'_i) / 2$, noting $N(\theta, \nu) = \mathbb{E}[N_n(\theta, \nu)]$, we have

$$|N_n(\theta, \nu) - N(\theta, \nu)| = 2 \left| \frac{1}{n} \sum_{i=1}^n (V_i - \mathbb{E}[V_i]) \right|$$

with the i.i.d. random variables $\{V_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(q)$ following Bernoulli distribution with parameter $q \geq \frac{1}{2}$ such that

$$q = \mathbb{E}[V_i] = \frac{1}{2} (1 + \text{sgn}(\nu^*) \mathbb{E}[(-1)^{z_i+1} \mathbb{E}[\text{sgn}(g_i g'_i)]] = \frac{1}{2} + \tanh |\nu^*| \cdot \frac{\varphi}{\pi}$$

by using the facts that $\mathbb{E}[(-1)^{z_i+1}] = \tanh \nu^*$ and $\mathbb{E}[\text{sgn}(g_i g'_i)] = \frac{2}{\pi} \varphi$ (see Lemma E.2). Then, we can express the $1 - q$ with $\phi = \pi - 2\varphi$ and $\|\pi - \frac{\mathbb{1}}{2}\|_1 = \tanh |\nu^*|$, $\min(\pi^*(1), \pi^*(2)) = \frac{1 - \tanh |\nu^*|}{2}$

$$1 - q = \frac{\pi - \tanh |\nu^*| \cdot 2\varphi}{2\pi} = \left\| \pi - \frac{\mathbb{1}}{2} \right\|_1 \cdot \frac{\phi}{2\pi} + \min(\pi^*(1), \pi^*(2))$$

By Lemma E.7 of concentration inequality for Bernoulli distribution, with probability at least $1 - \delta$,

$$|N_n(\theta, \nu) - N(\theta, \nu)| = 2 \left| \frac{1}{n} \sum_{i=1}^n (V_i - \mathbb{E}[V_i]) \right| \lesssim \min \left(\frac{\log \frac{1}{\delta}/n}{\log \left(1 + \frac{\log \frac{1}{\delta}/n}{1-q} \right)}, \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$$

We complete the proof by introducing $p := 1 - q = \|\pi^* - \frac{\mathbb{1}}{2}\|_1 \frac{\phi}{2\pi} + \min(\pi^*(1), \pi^*(2)) \asymp \tanh |\nu^*| \phi + (1 - \tanh |\nu^*|)$. $\square$

Lemma C.4 (Projected Statistical Accuracy of Easy EM Update for Regression Parameters). *In the noiseless setting, the projection on $\text{span}\{\theta, \theta^*\}$ for the statistical accuracy of regression parameters θ satisfies*

$$\frac{\|P_{\theta, \theta^*} [\text{sgn}(\rho) M_n^{\text{easy}}(\theta, \nu) - M_n^{\text{easy}}(\theta^*, \nu^*)]\|}{\|\theta^*\|} = \mathcal{O} \left(\phi^2 \vee \phi \frac{\log \frac{1}{\delta}}{n} \right)$$

with probability at least $1 - \delta$, where $M_n^{\text{easy}}(\theta, \nu)$ denotes the Easy EM update rule for regression parameters at the Finite-sample level, orthogonal projection matrix P_{θ, θ^*} satisfies $\text{span}(P_{\theta, \theta^*}) = \text{span}\{\theta, \theta^*\}$, and $\phi = 2 \arccos |\rho|$, $\rho = \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$.

Proof. Without loss of generality, we assume $\dim \text{span}\{\theta, \theta^*\} = 2$ and therefore $\dim \text{span}\{\theta, \theta^*\}^\perp = d - 2$. We introduce $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$, and define $\hat{e}_1 := \frac{\theta^*}{\|\theta^*\|}$, $\vec{e}_1 := \frac{\theta}{\|\theta\|}$, and $\hat{e}_2 := \frac{\theta - \langle \theta, \hat{e}_1 \rangle \hat{e}_1}{\|\theta - \langle \theta, \hat{e}_1 \rangle \hat{e}_1\|}$, $\vec{e}_2 := \frac{\theta^* - \langle \theta^*, \vec{e}_1 \rangle \vec{e}_1}{\|\theta^* - \langle \theta^*, \vec{e}_1 \rangle \vec{e}_1\|}$. Also, let $P_{\theta, \theta^*} = \hat{e}_1 \hat{e}_1^\top + \hat{e}_2 \hat{e}_2^\top$, which can be expressed with orthonormal basis matrices $U_{\theta, \theta^*} = (\hat{e}_1, \text{sgn}(\rho) \hat{e}_2)$ such that $P_{\theta, \theta^*} = U_{\theta, \theta^*} U_{\theta, \theta^*}^\top$. Also, we have $\phi = 2 \arccos |\rho|$, therefore $\rho = \text{sgn}(\rho) \cos \frac{\phi}{2}$. By defining $g_i := \langle x_i, \hat{e}_1 \rangle$, $g'_i := \text{sgn}(\rho) \langle x_i, \vec{e}_1 \rangle$, then $g_i, g'_i \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[g_i g'_i] = \cos \frac{\phi}{2}$, and $h_i = \langle x_i, \text{sgn}(\rho) \hat{e}_2 \rangle \sim \mathcal{N}(0, 1)$ satisfies $g'_i = \cos \frac{\phi}{2} \cdot g_i + \sin \frac{\phi}{2} \cdot h_i$, $h_i \perp g_i$.

$$-U_{\theta, \theta^*}^\top \frac{[\text{sgn}(\rho) M_n^{\text{easy}}(\theta, \nu) - M_n^{\text{easy}}(\theta^*, \nu^*)]}{\|\theta^*\|} = \left(\frac{1}{n} \sum_{i=1}^n [1 - \text{sgn}(g_i g'_i)] g_i^2, -\frac{1}{n} \sum_{i=1}^n [\text{sgn}(g_i g'_i) - 1] g_i h_i \right)^\top$$

Applying Proposition 14 of Errors of EM Updates and Sub-optimality Angle, and $\mathbb{E}[M_n^{\text{easy}}(\theta, \nu)] = M(\theta, \nu)$, $\mathbb{E}[M_n^{\text{easy}}(\theta^*, \nu^*)] = \theta^* \in \text{span}\{\theta, \theta^*\} = \text{span}(U_{\theta, \theta^*})$, then expectation $\mathbb{E}[U_{\theta, \theta^*}^\top \vec{v}]$ of $\vec{v} := [\text{sgn}(\rho) M_n^{\text{easy}}(\theta, \nu) - M_n^{\text{easy}}(\theta^*, \nu^*)] / \|\theta^*\|$ is bounded:

$$\|\mathbb{E}[U_{\theta, \theta^*}^\top \vec{v}]\| = \left\| U_{\theta, \theta^*}^\top \frac{\mathbb{E}[\text{sgn}(\rho) M_n^{\text{easy}}(\theta, \nu) - M_n^{\text{easy}}(\theta^*, \nu^*)]}{\|\theta^*\|} \right\| = \|\mathbb{E}[\vec{v}]\| = \frac{\|M(\theta, \nu) - \text{sgn}(\rho) \theta^*\|}{\|\theta^*\|} \leq \frac{\phi^2}{2\pi}$$

Applying $\sqrt{(s + s')^2 + (t + t')^2} \leq \sqrt{s^2 + t^2} + \sqrt{2} \max(s', t')_+$ for $s + s', t + t' \geq 0$, noting $P_{\theta, \theta^*} = U_{\theta, \theta^*} U_{\theta, \theta^*}^\top$:

$$\|P_{\theta, \theta^*} \vec{v}\| = \|U_{\theta, \theta^*}^\top \vec{v}\| \lesssim \phi^2 + \max \left(\left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) [1 - \text{sgn}(g_i g'_i)] g_i^2, \left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) [\text{sgn}(g_i g'_i) - 1] g_i h_i \right)_+$$

By introducing $\{\varrho_i\}_{i=1}^n = \{R_i^2/2\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(1)$ with $\{R_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} r \exp(r^2/2) \mathbb{1}_{r \geq 0}$ (the standard Rayleigh distribution), $\{V_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, 2\pi)$, and $U_i := 2V_i - \pi \bmod (2\pi) \sim \text{Unif}[0, 2\pi)$, we express $g_i = R_i \cos V_i$, $h_i = R_i \sin V_i$ and follows:

$$X_i := [1 - \text{sgn}(g_i g'_i)] g_i^2 = 2\varrho_i (1 - \cos U_i) \mathbb{1}_{U_i \in [0, \phi)} \geq 0, \quad Y_i := [\text{sgn}(g_i g'_i) - 1] g_i h_i = 2\varrho_i \sin U_i \mathbb{1}_{U_i \in [0, \phi)} \geq 0$$

Noting $\mathbb{E}[\varrho_i^q] = q!$ for $q \in \mathbb{Z}_+$, and $2(1 - \cos U_i) \mathbb{1}_{U_i \in [0, \phi)} \leq \phi^2 \mathbb{1}_{U_i \in [0, \phi)}$, $2 \sin U_i \leq 2\phi \mathbb{1}_{U_i \in [0, \phi)}$, then for $q \geq 3$:

$$\begin{aligned} \mathbb{E}[X_i^2] &\leq 2! [\phi^2]^2 \mathbb{E}[\mathbb{1}_{U_i \in [0, \phi)}] = \frac{\phi^5}{\pi}, \quad \mathbb{E}[X_i^q] \leq q! [\phi^2]^q \mathbb{E}[\mathbb{1}_{U_i \in [0, \phi)}] = \frac{q! \phi^{2q+1}}{2\pi} = \frac{q!}{2} \frac{\phi^5}{\pi} [\phi^2]^{q-2} \\ \mathbb{E}[Y_i^2] &\leq 2! [2\phi]^2 \mathbb{E}[\mathbb{1}_{U_i \in [0, \phi)}] = \frac{4\phi^3}{\pi}, \quad \mathbb{E}[Y_i^q] \leq q! [2\phi]^q \mathbb{E}[\mathbb{1}_{U_i \in [0, \phi)}] = \frac{q! 2^q \phi^{q+1}}{2\pi} = \frac{q!}{2} \frac{4\phi^3}{\pi} [2\phi]^{q-2} \end{aligned}$$

By Bernstein's inequality (see Theorem 2.10 on page 37 of [61]) with probability at least $1 - \delta$, noting $\phi^2 \lesssim \phi$, we have:

$$\max \left(\left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) X_i, \left(\frac{1}{n} \sum_{i=1}^n -\mathbb{E} \right) Y_i \right)_+ \lesssim \max \left(\phi^{\frac{5}{2}} \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \phi^2 \frac{\log \frac{1}{\delta}}{n}, \phi^{\frac{3}{2}} \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \phi \frac{\log \frac{1}{\delta}}{n} \right) \asymp \phi^{\frac{3}{2}} \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \phi \frac{\log \frac{1}{\delta}}{n}$$

Therefore, we complete the proof by using $a + \sqrt{ab} + b \asymp a + b \asymp a \vee b$ for $a, b \geq 0$. $\square$

Theorem C.5 (Proposition 19: Statistical Accuracy of EM Updates for Regression Parameters and Mixing Weights). *In the noiseless setting, the finite-sample EM with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples achieves the statistical accuracy of regression parameters and mixing weights:*

$$\frac{\|M_n(\theta, \nu) - \text{sgn}(\rho)\theta^*\|}{\|\theta^*\|} = \mathcal{O}\left(\phi^2 \vee \phi^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right)$$

$$|N_n(\theta, \nu) - \text{sgn}(\rho) \tanh \nu^*| = \mathcal{O}\left(\phi \left\|\pi^* - \frac{\mathbf{1}}{2}\right\|_1 \vee \left[\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right]\right)$$

with probability at least $1 - \delta$, where $M_n(\theta, \nu)$ denotes the EM update rule for regression parameters at the Finite-sample level, and $\phi := 2 \arccos |\rho|$, $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \|\theta^*\|}$ and $p := \frac{\phi}{2\pi} \|\pi^* - \frac{\mathbf{1}}{2}\|_1 + \min(\pi^*(1), \pi^*(2))$.

Proof. Noting $\Sigma M_n(\theta, \nu) = M_n^{\text{easy}}(\theta, \nu) = \frac{1}{n} \sum_{i=1}^n \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle) \langle x_i, \theta^* \rangle x_i$ in the noiseless setting, with the ensemble matrix of data samples $\Sigma = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top$, and letting $\vec{v} := [\text{sgn}(\rho) M_n^{\text{easy}}(\theta, \nu) - M_n^{\text{easy}}(\theta^*, \nu^*)] / \|\theta^*\|$, then:

$$\frac{M_n(\theta, \nu) - \text{sgn}(\rho)\theta^*}{\|\theta^*\|} = \text{sgn}(\rho) \Sigma^{-1} \cdot \vec{v}$$

Let $P_{\theta, \theta^*}^\perp = \sum_{j=1}^{d-2} \hat{e}_j^\perp (\hat{e}_j^\perp)^\top$ and $U_{\theta, \theta^*}^\perp = (\hat{e}_1^\perp, \dots, \hat{e}_{d-2}^\perp)$ with $\text{span}(U_{\theta, \theta^*}^\perp) = \text{span}\{\theta, \theta^*\}^\perp$. For the length of vector $\vec{v}$, noting that orthogonal projection matrix $P_{\theta, \theta^*}^\perp$ can be expressed with orthonormal basis matrix $U_{\theta, \theta^*}^\perp$ such that $P_{\theta, \theta^*}^\perp = U_{\theta, \theta^*}^\perp (U_{\theta, \theta^*}^\perp)^\top$.

$$\|\vec{v}\|^2 = \langle (P_{\theta, \theta^*} + P_{\theta, \theta^*}^\perp) \vec{v}, (P_{\theta, \theta^*} + P_{\theta, \theta^*}^\perp) \vec{v} \rangle = \langle \vec{v}, (P_{\theta, \theta^*} + P_{\theta, \theta^*}^\perp) \vec{v} \rangle = \langle \vec{v}, P_{\theta, \theta^*} \vec{v} \rangle + \langle \vec{v}, P_{\theta, \theta^*}^\perp \vec{v} \rangle = \|P_{\theta, \theta^*} \vec{v}\|^2 + \|(U_{\theta, \theta^*}^\perp)^\top \vec{v}\|^2$$

where the first term is bounded by $\|P_{\theta, \theta^*} \vec{v}\| \lesssim \phi^2 \vee \phi \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, and we rewrite the second term by introducing Gaussian r.v. $g_i := \langle x_i, \theta^* / \|\theta^*\| \rangle$, $g'_i := \langle x_i, \text{sgn}(\rho)\theta / \|\theta\| \rangle \sim \mathcal{N}(0, 1)$, $\vec{g}_i := \text{sgn}(\langle x_i, \theta^* \rangle \langle x_i, \theta \rangle) (U_{\theta, \theta^*}^\perp)^\top x_i \sim \mathcal{N}(0, I_{d-2})$ with $g_i, g'_i \perp \vec{g}_i$.

$$\|(U_{\theta, \theta^*}^\perp)^\top \vec{v}\|^2 = \frac{1}{n^2} \left\| \sum_{i=1}^n [1 - \text{sgn}(g'_i g_i)] g_i \vec{g}_i \right\|^2 = \frac{2}{n^2} \left(\sum_{i=1}^n [1 - \text{sgn}(g'_i g_i)] g_i^2 \right) \cdot \left\| \sum_{i=1}^n \frac{[1 - \text{sgn}(g'_i g_i)] g_i}{\sqrt{\sum_{i'=1}^n 2[1 - \text{sgn}(g'_{i'} g_{i'})] (g_{i'})^2}} \vec{g}_i \right\|^2$$

where the weighted sum $\sum_{i=1}^n ([1 - \text{sgn}(g'_i g_i)] g_i / \sqrt{\sum_{i'=1}^n 2[1 - \text{sgn}(g'_{i'} g_{i'})] (g_{i'})^2}) \vec{g}_i \sim \mathcal{N}(0, I_{d-2})$ by the rotational invariance of Gaussian distribution, therefore $\|\sum_{i=1}^n ([1 - \text{sgn}(g'_i g_i)] g_i / \sqrt{\sum_{i'=1}^n 2[1 - \text{sgn}(g'_{i'} g_{i'})] (g_{i'})^2}) \vec{g}_i\|^2 \sim \chi^2(d-2)$. By the concentration inequality for Chi-square distribution (see Lemma 1, page 1325 in [53]), then:

$$\frac{1}{n} \left\| \sum_{i=1}^n \frac{[1 - \text{sgn}(g'_i g_i)] g_i}{\sqrt{\sum_{i'=1}^n 2[1 - \text{sgn}(g'_{i'} g_{i'})] (g_{i'})^2}} \vec{g}_i \right\|^2 \lesssim \frac{d \vee \log \frac{1}{\delta}}{n}$$

Letting $\{\varrho_i\}_{i=1}^n = \{R_i^2/2\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(1)$ with $\{R_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} r \exp(r^2/2) \mathbf{1}_{r \geq 0}$ (the standard Rayleigh distribution), $\{V_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, 2\pi)$, and $U_i := 2V_i - \pi \pmod{2\pi} \sim \text{Unif}[0, 2\pi)$, we express $g_i = R_i \cos V_i$, $h_i = R_i \sin V_i$, $g'_i = \cos \frac{\phi}{2} \cdot g_i + \sin \frac{\phi}{2} \cdot h_i$.

$$X_i := [1 - \text{sgn}(g_i g'_i)] g_i^2 = 2\varrho_i (1 - \cos U_i) \mathbf{1}_{U_i \in [0, \phi)} \geq 0, \quad \mathbb{E}[X_i] \leq \mathbb{E}[\varrho_i] \mathbb{E}[\phi^2 \mathbf{1}_{U_i \in [0, \phi)}] = \frac{\phi^3}{2\pi}$$

By the concentration inequality for X_i obtained in proof of previous Lemma, $(\frac{1}{n} \sum_{i=1}^n - \mathbb{E}) X_i \lesssim \phi^{\frac{5}{2}} \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \phi^2 \frac{\log \frac{1}{\delta}}{n}$, then:

$$\frac{1}{n} \sum_{i=1}^n [1 - \text{sgn}(g'_i g_i)] g_i^2 = \frac{1}{n} \sum_{i=1}^n X_i \lesssim \phi^3 + \phi^{\frac{5}{2}} \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \phi^2 \frac{\log \frac{1}{\delta}}{n} \asymp \phi^2 \left(\phi \vee \frac{\log \frac{1}{\delta}}{n} \right)$$

By combining the above inequalities together, we bound $\vec{v}$ as follows with $n \gtrsim d \vee \log \frac{1}{\delta}$ and probability at least $1 - \delta/4$.

$$\|\vec{v}\| \leq \|P_{\theta, \theta^*} \vec{v}\| + \|(U_{\theta, \theta^*}^\perp)^\top \vec{v}\| \lesssim \phi^2 \vee \phi \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \sqrt{\phi^2 \left(\phi \vee \frac{\log \frac{1}{\delta}}{n} \right) \frac{d \vee \log \frac{1}{\delta}}{n}} \asymp \phi^2 \vee \phi^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi \sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

From the assumption $n \gtrsim d \vee \log \frac{1}{\delta}$, by Example 6.2 (Operator norm bounds for the standard Gaussian ensemble) and Example 6.3 (Gaussian covariance estimation) on page 162 of [60], with probability at least $1 - \delta/4$, we have $\gamma_{\min}(\Sigma) \asymp 1$, where $\gamma_{\min}$ represents the minimum eigenvalue.

$$\frac{\|M_n(\theta, \nu) - \text{sgn}(\rho)\theta^*\|}{\|\theta^*\|} \leq \frac{\|\vec{v}\|}{\gamma_{\min}(\Sigma)} \asymp \phi^2 \vee \phi^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi \sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

For the mixing weights, by using the Triangle inequality, Proposition for error of mixing weights in population level, and the Proposition for statistical error of mixing weights in finite-sample level, we have the following with probability at least $1 - \delta/2$.

$$\begin{aligned} |N_n(\theta, \nu) - \text{sgn}(\rho) \tanh \nu^*| &\leq |N(\theta, \nu) - \text{sgn}(\rho) \tanh \nu^*| + |N_n(\theta, \nu) - N(\theta, \nu)| \\ &\lesssim \phi \left\| \pi^* - \frac{\mathbf{1}}{2} \right\|_1 + \min \left(\frac{\log \frac{1}{\delta}/n}{\log \left(1 + \frac{\log \frac{1}{\delta}/n}{p} \right)}, \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right) \end{aligned}$$

where $p := \|\pi^* - \frac{\mathbf{1}}{2}\|_1 \frac{\phi}{2\pi} + \min(\pi^*(1), \pi^*(2)) \asymp \tanh |\nu^*| \phi + (1 - \tanh |\nu^*|)$. $\square$

B. Initialization of Easy EM Updates and Perturbation in Sub-Optimality Angle

Theorem C.6 (Proposition 20: Initialization with Easy EM, Proposition 5.4 in [1]). *In the noiseless setting, suppose we run the sample-splitting finite-sample Easy EM with $n' \asymp \frac{n}{\log \frac{1}{\delta}}$ fresh samples for each iteration, then after at most $T_0 = \mathcal{O}(\log \frac{1}{\delta})$ iterations, it satisfies $\varphi^{T_0} \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$ with probability at least $1 - \delta$.*

Proof. See the detailed proof on pages 41-44, Appendix E of [1] based on Berry-Esseen bound for the central limit theorem (see Theorem 1.1 of [62]). The proposition is restated here for completeness and to keep the paper self-contained. $\square$

Lemma C.7 (Perturbation in Sub-Optimality Angle). *For $\vartheta := \sin \varphi \hat{e}_1 + \cos \varphi \hat{e}_2$, where $\{\hat{e}_1, \hat{e}_2\}$ is an orthonormal basis for the subspace $\text{span}\{\hat{e}_1, \hat{e}_2\} \subset \mathbb{R}^d$, and $\varphi \in (0, \frac{\pi}{2})$; with a perturbation vector $\varrho \in \mathbb{R}^d$ with length $\|\varrho\| = r \in (0, \sin \varphi)$; then the angle $\varphi' := \arcsin \frac{|\langle \vartheta + \varrho, \hat{e}_1 \rangle|}{\|\vartheta + \varrho\|}$, satisfies $\varphi' \geq \varphi - \arcsin r$*

Proof. Note that with $\|\varrho\| = r \in (0, \sin \varphi)$, $\langle \vartheta + \varrho, \hat{e}_1 \rangle = \langle \vartheta, \hat{e}_1 \rangle + \langle \varrho, \hat{e}_1 \rangle \geq \sin \varphi - \|\varrho\| = \sin \varphi - r > 0$, thus

$$\sin \varphi' = \frac{\langle \vartheta + \varrho, \hat{e}_1 \rangle}{\|\vartheta + \varrho\|} > 0$$

Express the perturbation vector by $\varrho = -r' \cos(\varphi - \Delta) \hat{e}_1 + r' \sin(\varphi - \Delta) \hat{e}_2 + \sqrt{r^2 - [r']^2} \hat{e}$, where $r' \in [0, r]$, $\Delta \in (-\pi, \pi]$ and $\hat{e} \in \text{span}\{\hat{e}_1, \hat{e}_2\}^\perp$, $\|\hat{e}\| = 1$

$$\langle \vartheta + \varrho, \hat{e}_1 \rangle = \langle \vartheta, \hat{e}_1 \rangle + \langle \varrho, \hat{e}_1 \rangle = \sin \varphi - r' \cos(\varphi - \Delta)$$

$$\|\vartheta + \varrho\| = \left\| [\sin \varphi - r' \cos(\varphi - \Delta)] \hat{e}_1 + [\cos \varphi + r' \sin(\varphi - \Delta)] \hat{e}_2 + \sqrt{r^2 - [r']^2} \hat{e} \right\| = \sqrt{[1 + r^2] - 2r' \sin \Delta}$$

Hence, let $p := \frac{r' |\sin \Delta|}{r} \in [0, 1]$, $-r' \cos \Delta \geq -\sqrt{[r']^2 - [rp]^2} \geq -\sqrt{r^2 - [rp]^2} = -r \sqrt{1 - p^2}$ and $p = \frac{r'}{r} |\sin \Delta| \leq 1$

$$\sin \varphi' = \frac{[1 - r' \sin \Delta] \sin \varphi - r' \cos \Delta \cos \varphi}{\sqrt{[1 + r^2] - 2r' \sin \Delta}} \geq \frac{[1 - rp] \sin \varphi - r \sqrt{1 - p^2} \cos \varphi}{\sqrt{[1 + r^2] - 2rp}} := \psi(p)$$

For $\psi(p) := \frac{[1 - rp] \sin \varphi - r \sqrt{1 - p^2} \cos \varphi}{\sqrt{[1 + r^2] - 2rp}}$, by $\cos(\varphi + [\frac{\pi}{2} - \arcsin p]) \leq \cos \varphi$, then $\cos \varphi - r \cos(\varphi + [\frac{\pi}{2} - \arcsin p]) > 0$

$$\frac{d}{dp} \log \psi = \frac{r \cdot \{ \cos \varphi - r \cos(\varphi + [\frac{\pi}{2} - \arcsin p]) \} (p - r)}{\sqrt{1 - p^2} \{ [1 - rp] \sin \varphi - r \sqrt{1 - p^2} \cos \varphi \} \cdot \{ [1 + r^2] - 2rp \}}$$

therefore, $\frac{d}{dp} \log \psi < 0, \forall p \in (0, r)$; $\frac{d}{dp} \log \psi > 0, \forall p \in (r, 1)$, hence $\psi(p) \geq \psi(p) |_{p=r}$ for $p \in [0, 1]$

$$\sin \varphi' \geq \psi(p) \geq \psi(p) |_{p=r} \sqrt{1 - r^2} \sin \varphi - r \cos \varphi = \sin(\varphi - \arcsin r)$$

Note that $r \in (0, \sin \varphi)$, that is $\frac{\pi}{2} > \varphi > \varphi - \arcsin r > 0$, and we show that

$$\varphi' \geq \varphi - \arcsin r$$

$\square$

C. Convergence Guarantees for Finite-Sample EM Updates in Noiseless Setting

Theorem C.8 (Proposition 21: Convergence of Sub-Optimality Angle). *In the noiseless setting, suppose $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, given a positive number $\varepsilon \lesssim \frac{\log \frac{1}{\delta}}{n}$, we run Easy finite-sample EM for $T_1 = \mathcal{O}(\log \frac{d}{\log \frac{1}{\delta}})$ iterations followed by the standard finite-sample EM for at most $T' = \mathcal{O}(\lceil \log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}} \rceil \vee \log \lceil \log \frac{n}{\ln \frac{1}{\delta}} / \log \frac{n}{d} \rceil \vee \lceil \log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}} \rceil)$ iterations*

with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples, then it satisfies

$$\phi^T \leq \varepsilon,$$

with probability at least $1 - T\delta$, where $T := T_1 + T'$, $\varphi^0 := \frac{\pi}{2} - \arccos \left| \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|} \right|$ and $\varphi^T := \frac{\pi}{2} - \arccos \left| \frac{\langle \theta^T, \theta^* \rangle}{\|\theta^T\| \cdot \|\theta^*\|} \right|$.

Proof. We provide a new concise and rigorous proof for the Proposition for convergence of angle, which is reproduced here for completeness. The other old lengthy proof of this Proposition for convergence of angle can be found on pages 45-50, Appendix E of [1]. We divide the proof into four stages:

(1) In the first stage, we show that after running the Easy EM for at most $T_1 = \mathcal{O} \left(\log \frac{d}{\log \frac{1}{\delta}} \right)$ iterations at the finite-sample level, given the initial sub-optimality angle $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, the sub-optimality angle satisfies $\varphi^{T_1} \gtrsim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$.

(2) In the second stage, we show that after running the standard EM for at most $T_2 = \mathcal{O} \left(\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}} \right)$ iterations at the finite-sample level, given that $\varphi^{T_1} \gtrsim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$, the sub-optimality angle satisfies $\phi^{T_1+T_2} \lesssim 1$.

(3) In the third stage, we show that after running the standard EM for at most $T_3 = \mathcal{O} \left(\log \left[\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}} \right] \right)$ iterations at the finite-sample level, given that $\phi^{T_1+T_2} \lesssim 1$, the sub-optimality angle satisfies $\phi^{T_1+T_2+T_3} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$.

(4) In the fourth stage, we show that after running the standard EM for at most $T_4 = \mathcal{O} \left(\log \left[\log \frac{n}{\ln \frac{1}{\delta}} / \log \frac{n}{d} \right] \vee \left[\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}} \right] \right)$ iterations at the finite-sample level, given that $\phi^{T_1+T_2+T_3} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$, the sub-optimality angle satisfies $\phi^{T_1+T_2+T_3+T_4} \leq \varepsilon$.

We denote $\bar{e}_1 = \theta^* / \|\theta^*\|$ as the unit vector in the direction of θ^* , also define $\varphi^t = \frac{\pi}{2} - \arccos \left| \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \cdot \|\theta^*\|} \right|$, $\phi^t = 2 \arccos \left| \frac{\langle \theta^t, \theta^* \rangle}{\|\theta^t\| \cdot \|\theta^*\|} \right|$ as the sub-optimality angles. Also, we denote $\bar{\theta}^t = M(\theta^{t-1}, \nu^{t-1})$ and $\bar{\varphi}^t := \frac{\pi}{2} - \arccos \left| \frac{\langle \bar{\theta}^t, \theta^* \rangle}{\|\bar{\theta}^t\| \cdot \|\theta^*\|} \right|$, $\bar{\phi}^t = 2 \arccos \left| \frac{\langle \bar{\theta}^t, \theta^* \rangle}{\|\bar{\theta}^t\| \cdot \|\theta^*\|} \right|$ as the sub-optimality angles of $\bar{\theta}^t$ for the analysis of population level. Further, we denote $c_{\text{proj}}, c_{\text{stat}}$ as the universal constants for the projected error and the statistical error such that $\|P_{\theta, \theta^*}[M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)]\| / \|\theta^*\| \leq c_{\text{proj}} \times \sqrt{\frac{\log \frac{1}{\delta}}{n}}$ and $\|M_n^{\text{easy}}(\theta, \nu) - M(\theta, \nu)\| / \|\theta^*\|, \|M_n(\theta, \nu) - M(\theta, \nu)\| / \|\theta^*\| \leq c_{\text{stat}} \times \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$ with $n \gtrsim d \vee \log \frac{1}{\delta}$ large enough sample size.

Stage 1: run Easy EM from $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$ to $\varphi^{T_1} \gtrsim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$

Let's select sample number $n \gtrsim d \vee \log \frac{1}{\delta}$ large enough such that assumption $\varphi^t < C_1 \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} < \frac{1}{20}$ holds for all $t \in [0, T_1]$, and the statistical error of $\|\theta^t - \bar{\theta}^t\| / \|\theta^*\| \leq c_{\text{stat}} \times \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} < \frac{2}{\pi} \times \frac{1}{20}$ is also small enough. then we can establish the lower bound for the projection of $\bar{\theta}^{t+1} = M(\theta^t, \nu^t)$ onto $\bar{e}_1$, which is denoted by

$$\frac{\langle \bar{\theta}^{t+1}, \hat{e}_1 \rangle}{\langle \theta^t, \hat{e}_1 \rangle} = \frac{\|M(\theta^t, \nu^t)\|}{\|\bar{\theta}^t + (\theta^t - \bar{\theta}^t)\|} \times \frac{\tan \bar{\varphi}^{t+1}}{\tan \varphi^t} \times \frac{\cos \bar{\varphi}^{t+1}}{\cos \varphi^t} \geq \frac{2/\pi}{(2/\pi)(\frac{1}{20} + \cos \frac{1}{20}) + (2/\pi) \times \frac{1}{20}} \times \frac{1 + \sqrt{5}}{2} \times \frac{1}{\tan \frac{1}{20} + 1} > \frac{7}{5}$$

where the first inequality is by the fact the distance from the point on the cycloid trajectory to the origin is increasing as the sub-optimality angle φ^t is increasing, and its minimum value is $2/\pi$, and the second inequality is due to the Proposition for the linear growth of $\tan \varphi^t$, and the last inequality is from the proof of the Proposition of the recurrence relation. Without loss of generality, we assume $\langle \theta^0, \hat{e}_1 \rangle > 0$, then by the assumption $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, we have $\langle \theta^t, \hat{e}_1 \rangle > 0$ for all t , then

$$\sin \varphi^{t+1} = \frac{\langle \bar{\theta}^{t+1}, \hat{e}_1 \rangle - \langle \bar{\theta}^{t+1} - \theta^{t+1}, \hat{e}_1 \rangle}{\|\bar{\theta}^{t+1} + (\theta^{t+1} - \bar{\theta}^{t+1})\|} \geq \frac{\frac{7}{5} \langle \theta^t, \hat{e}_1 \rangle / \|\theta^*\| - c_{\text{proj}} \times \sqrt{\frac{\log \frac{1}{\delta}}{n}}}{\|\bar{\theta}^{t+1}\| / \|\theta^*\| + \frac{2}{\pi} \times \frac{1}{20}} \geq \frac{\frac{7}{5} \times \frac{19}{20}}{\frac{1}{10} + \cos \frac{1}{20}} \sin \varphi^t - \frac{c_{\text{proj}}}{\frac{1}{10} + \cos \frac{1}{20}} \times \sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

where the second inequality is due the following fact and noting that $\sin \varphi^t = \frac{\langle \theta^t, \hat{e}_1 \rangle}{\|\theta^t\|}$:

$$\frac{2}{\pi} \leq \|\bar{\theta}^{t+1}\| / \|\theta^*\| \leq \frac{2}{\pi} \times \left(\frac{1}{20} + \cos \frac{1}{20} \right), \quad \|\theta^t\| / \|\theta^*\| \geq \|\bar{\theta}^t\| / \|\theta^*\| - \|\theta^t - \bar{\theta}^t\| / \|\theta^*\| \geq \frac{2}{\pi} \times \left(1 - \frac{1}{20} \right)$$

Suppose $\varphi^0 \geq 45c_{\text{proj}} \sqrt{\frac{\log \frac{1}{\delta}}{n}}$ by $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, then $\sin \varphi^0 - 30c_{\text{proj}} \sqrt{\frac{\log \frac{1}{\delta}}{n}} \geq 10c_{\text{proj}} \sqrt{\frac{\log \frac{1}{\delta}}{n}}$, the linear growth is shown.

$$\sin \varphi^{t+1} \geq \left(1 + \frac{1}{30} \right) \sin \varphi^t - c_{\text{proj}} \times \sqrt{\frac{\log \frac{1}{\delta}}{n}} \implies \sin \varphi^{t+1} - 30c_{\text{proj}} \times \sqrt{\frac{\log \frac{1}{\delta}}{n}} \geq \left(1 + \frac{1}{30} \right) \left(\sin \varphi^t - 30c_{\text{proj}} \times \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$$

Since the population EM update only depends on the sub-optimality angle φ^t , the above inequality for $t > 0$ also holds for $t = 0$ without loss of generality. Therefore, by the induction, we have shown that after running the Easy EM for at most $T_1 \leq \left\lceil \frac{\ln(C_1/(10c_{\text{proj}})) + \frac{1}{2} \max(0, \ln(d/\log \frac{1}{\delta}))}{\ln(1+1/30)} \right\rceil_+ \asymp \log \frac{d}{\log \frac{1}{\delta}}$ iterations, we have $\varphi^{T_1} \geq \sin \varphi^{T_1} \geq C_1 \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$ with some universal constant $C_1 > 0$.

Stage 2: run Standard EM from $\varphi^{T_1} \gtrsim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$ **to** $\phi^{T_1+T_2} \lesssim 1$

Given that $\varphi^{T_1} \geq C_1 \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$, for some positive constant $C_1 > 0$, here we let $C_1 = 20c_{\text{stat}}$, we continue to run the standard EM with the initial sub-optimality angle $\varphi^{T_1} \geq 20c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} > c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$.

For the ease of analysis, we introduce the following notations: $r := \frac{\pi}{2} c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} < \frac{1}{20}$ and $s := \frac{r}{\sqrt{1-r^2}} < (1 + \frac{1}{100})r$, namely $r = \frac{s}{\sqrt{1+s^2}}$. Then by the previous established Lemma for single iteration of perturbation and using the fact that $\|\theta^{t+1} - \bar{\theta}^{t+1}\| \leq c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$ and $\|\bar{\theta}^{t+1}\| \geq \frac{2}{\pi}$, we have shown that

$$\varphi^{t+1} \geq \bar{\varphi}^{t+1} - \arcsin \left(\frac{\|\theta^{t+1} - \bar{\theta}^{t+1}\|}{\|\bar{\theta}^{t+1}\|} \right) \geq \bar{\varphi}^{t+1} - \arcsin \left(\frac{c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}}{2/\pi} \right) = \bar{\varphi}^{t+1} - \arcsin r$$

Applying the linear growth of $\tan \varphi$ in population level in the proven Proposition, we have $\tan \bar{\varphi}^{t+1} \geq 2 \tan \varphi^t$. and noting that $\tan(x - y) = \frac{\tan x - \tan y}{1 + \tan x \tan y}$, $\tan(\arcsin r) = s$,

$$\tan \varphi^{t+1} \geq \tan(\bar{\varphi}^{t+1} - \arcsin r) \geq \frac{2 \tan \varphi^t - s}{1 + 2 \tan \varphi^t \cdot s} = (1 + \frac{1}{10}) \tan \varphi^t - \frac{s}{1 + 2s \cdot \tan \varphi^t} + \frac{\tan \varphi^t (2 - 1.1 - 1.1s \tan \varphi^t)}{1 + 2s \cdot \tan \varphi^t}$$

In particular, when $\varphi^t \leq 1$, we have $2 - 1.1 - 2.2s \tan \varphi^t > 0$ for $s < (1 + \frac{1}{100})r < (1 + \frac{1}{100})\frac{1}{20}$, therefore,

$$\tan \varphi^{t+1} \geq (1 + \frac{1}{10}) \tan \varphi^t - s \geq (1 + \frac{1}{10}) \tan \varphi^t - (1 + \frac{1}{100})r \geq (1 + \frac{1}{10}) \tan \varphi^t - \frac{16}{10} c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$$

By the initial assumption $\varphi^{T_1} \geq 20c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} > c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$, as long as $\varphi^t \leq 1$, we have

$$\tan \varphi^{t+1} - 16c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} \geq (1 + \frac{1}{10}) \left(\tan \varphi^t - 16c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} \right) > 0$$

Therefore, by the induction, we have shown that after running the standard EM for at most $T_2 \leq \left\lceil \frac{\ln(\frac{\tan 1}{4c_{\text{stat}}}) + \frac{1}{2} \min(\ln \frac{n}{d}, \ln \frac{n}{\log \frac{1}{\delta}})}{\ln(1+1/10)} \right\rceil_+$ $\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}$ iterations, we have $\varphi^{T_1+T_2} \geq 1$, and therefore $\phi^{T_1+T_2} = \pi - 2\varphi^{T_1+T_2} < 1.2$.

Stage 3: run Standard EM from $\phi^{T_1+T_2} \lesssim 1$ **to** $\phi^{T_1+T_2+T_3} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$

Given that $\phi^{T_1+T_2} < 1.2 < 1.4$, we continue to run the standard EM with quadratic convergence rate. Recall the established inequality $\varphi^{t+1} \geq \bar{\varphi}^{t+1} - \arcsin r$, then by the relation $\phi^{t+1} = \pi - 2\varphi^{t+1}$, $\bar{\phi}^{t+1} = \pi - 2\bar{\varphi}^{t+1}$ and noting that $\arcsin r \leq \frac{\pi}{2}r$.

$$\phi^{t+1} \leq \bar{\phi}^{t+1} + 2 \arcsin r \leq \bar{\phi}^{t+1} + \pi r$$

Applying Proposition for the quadratic convergence rate of ϕ^t in population level, namely $\bar{\phi}^{t+1}/\pi \leq [\phi^t/\pi]^2$ when $\phi^t \leq 1.4$.

$$\frac{\phi^{t+1}}{\pi} \leq \left[\frac{\phi^t}{\pi} \right]^2 + r$$

In one iteration, $\phi^{T_1+T_2+1}/\pi \leq (1.2/\pi)^2 + r < (1.2/\pi)^2 + \frac{1}{20} < \frac{1}{5}$, then we have $\phi^t < \frac{\pi}{5} < \frac{\pi}{4}$ in the following iterations.

$$\left[\frac{\phi^t}{\pi} - 2r \right] \leq \left[\frac{\phi^t}{\pi} \right]^2 - r = \left[\frac{\phi^t}{\pi} - 2r \right]^2 - 4r^2 - \frac{4r}{\pi} \left(\frac{\pi}{4} - \phi^t \right) < \left[\frac{\phi^t}{\pi} - 2r \right]^2$$

Therefore, by the above inequality, we have shown that after running the standard EM for at most $T_3 \leq 1 + \left\lceil \frac{\ln(\frac{\ln(1/r)}{\ln 5})}{\ln 2} \right\rceil = 1 + \left\lceil \frac{\ln[\frac{1}{2} \min(\ln(n/d), \ln(n/\log \frac{1}{\delta}))] - \ln(\frac{\pi}{2} c_{\text{stat}}) - \ln \ln 5}{\ln 2} \right\rceil_+$ $\log[\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}]$ iterations, we have $\phi^{T_1+T_2+T_3} \leq 3r = \frac{3\pi}{2} c_{\text{stat}} \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$.

Stage 4: run Standard EM from $\phi^{T_1+T_2+T_3} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}}$ **to** $\phi^{T_1+T_2+T_3+T_4} \lesssim \frac{\log \frac{1}{\delta}}{n}$

By the previous established Proposition of statistical accuracy, we have the following inequality:

$$\phi^{t+1} \leq \pi \sin \frac{\phi^{t+1}}{2} \leq \pi \frac{\|\theta^{t+1} - \text{sgn}(\rho^{t+1})\theta^*\|}{\|\theta^*\|} \lesssim [\phi^t]^2 \vee [\phi^t]^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi^t \sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

With the assumption that $n \gtrsim d \vee \log \frac{1}{\delta}$ and the fact that $\phi^{T_1+T_2+T_3} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} \lesssim 1$, then for $t \geq T_1 + T_2 + T_3 + 1$, we have

$$\phi^t \leq \phi^{T_1+T_2+T_3+1} \lesssim \sqrt{\frac{d \vee \log \frac{1}{\delta}}{n}} \phi^{T_1+T_2+T_3} \lesssim \frac{d \vee \log \frac{1}{\delta}}{n}, \quad \phi^t \lesssim [\phi^t]^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi^t \sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

Under the assumption that $n \gtrsim d \vee \log \frac{1}{\delta}$, if (i) $n \lesssim d^2 / \log \frac{1}{\delta}$ (therefore, $d \vee \log \frac{1}{\delta} \asymp d$ must hold), then only when $\phi^t \gtrsim \log \frac{1}{\delta} / d$:

$$\phi^{t+1} \lesssim [\phi^t]^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi^t \sqrt{\frac{\log \frac{1}{\delta}}{n}} \lesssim [\phi^t]^{\frac{3}{2}} \sqrt{\frac{d}{n}} \implies \phi^{[T_1+T_2+T_3+1]+\Theta\left(\left\lceil \ln\left(\ln \frac{n}{\log \frac{1}{\delta}} / (2 \ln \frac{n}{d})\right) / \ln \frac{3}{2}\right\rceil\right)} \lesssim \frac{\log \frac{1}{\delta}}{d}$$

then, for $t \geq T_1 + T_2 + T_3 + 1 + \Theta\left(\left\lceil \ln\left(\ln \frac{n}{\log \frac{1}{\delta}} / (2 \ln \frac{n}{d})\right) / \ln \frac{3}{2}\right\rceil\right)$, we have

$$\phi^{t+1} \lesssim \phi^t \sqrt{\frac{\log \frac{1}{\delta}}{n}} \implies \phi^{[T_1+T_2+T_3+1]+\Theta\left(\left\lceil \ln\left(\ln \frac{n}{\log \frac{1}{\delta}} / (2 \ln \frac{n}{d})\right) / \ln \frac{3}{2}\right\rceil\right)+\Theta\left(\left\lceil 2[\ln \frac{1}{\varepsilon} - \ln \frac{d}{\log \frac{1}{\delta}}] / \ln \frac{n}{\log \frac{1}{\delta}}\right\rceil\right)} \leq \varepsilon$$

Otherwise, if (ii) $n \gtrsim d^2 / \log \frac{1}{\delta}$, then for $t \geq T_1 + T_2 + T_3 + 1$, we have

$$\phi^{t+1} \lesssim \phi^t \sqrt{\frac{\log \frac{1}{\delta}}{n}} \implies \phi^{[T_1+T_2+T_3+1]+\Theta\left(\left\lceil 2[\ln \frac{1}{\varepsilon} - \ln \frac{n}{\log \frac{1}{\delta}} \wedge \ln \frac{n}{d}] / \ln \frac{n}{\log \frac{1}{\delta}}\right\rceil\right)} \leq \varepsilon$$

Combining the above two case (i) and (ii), we have shown that after running the standard EM for at most $T_4 \lesssim \log[\log \frac{n}{\log \frac{1}{\delta}} / \log \frac{n}{d}] \vee [\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}}]$ iterations, we have $\phi^{T_1+T_2+T_3+T_4} \leq \varepsilon$.

Summary: number of iterations Therefore, after running Easy EM for at most $T_1 = \mathcal{O}\left(\log \frac{d}{\log \frac{1}{\delta}}\right)$ iterations, and then running Standard EM for at most $T' := T_2+T_3+T_4 = \mathcal{O}\left(\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}} + \log[\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}] + \log[\log \frac{n}{\log \frac{1}{\delta}} / \log \frac{n}{d}] \vee [\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}}]\right) = \mathcal{O}\left([\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}] \vee \log[\log \frac{n}{\log \frac{1}{\delta}} / \log \frac{n}{d}] \vee [\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}}]\right)$ iterations, we have $\phi^T = \phi^{T_1+T_2+T_3+T_4} \leq \varepsilon$ with probability at least $1 - T\delta$, where $T := T_1 + T' = T_1 + T_2 + T_3 + T_4$. $\square$

Theorem C.9 (Theorem 22: Finite-Sample Level Convergence). *In the noiseless setting, suppose any initial mixing weights π^0 and any initial regression parameters $\theta^0 \in \mathbb{R}^d$ ensuring that $\varphi^0 \gtrsim \sqrt{\frac{\log \frac{1}{\delta}}{n}}$. Given a positive number $\varepsilon \lesssim \frac{\log \frac{1}{\delta}}{n}$, we run finite-sample Easy EM for at most $T_1 = \mathcal{O}\left(\log \frac{d}{\log \frac{1}{\delta}}\right)$ iterations followed by the finite-sample standard EM for at most $T' = \mathcal{O}\left([\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}] \vee \log[\log \frac{n}{\log \frac{1}{\delta}} / \log \frac{n}{d}] \vee [\log \frac{1}{\varepsilon} / \log \frac{n}{\log \frac{1}{\delta}}]\right)$ iterations with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples, then*

$$\frac{\|\theta^{T+1} - \text{sgn}(\rho^{T+1})\theta^*\|}{\|\theta^*\|} = \mathcal{O}\left(\varepsilon \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right), \quad \|\pi^{T+1} - \bar{\pi}^*\|_1 = \mathcal{O}\left(\varepsilon \left\|\frac{1}{2} - \pi^*\right\|_1 \vee \left[\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p(\varepsilon, \pi^*)}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right]\right),$$

with probability at least $1 - T\delta$, where $T := T_1 + T'$, $\varphi^0 := \frac{\pi}{2} - \arccos\left|\frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|}\right|$, $\rho^{T+1} := \frac{\langle \theta^{T+1}, \theta^* \rangle}{\|\theta^{T+1}\| \cdot \|\theta^*\|}$, $\bar{\pi}^* := \frac{1}{2} + \text{sgn}(\rho^{T+1})(\pi^* - \frac{1}{2})$, and $p(\varepsilon, \pi^*) := \varepsilon \|\pi^* - \frac{1}{2}\|_1 + \min(\pi^*(1), \pi^*(2))$.

Proof. By Proposition for convergence of angle, we have shown that $\phi^T \leq \varepsilon \lesssim \frac{\log \frac{1}{\delta}}{n}$ after T iterations with $n \gtrsim d \vee \log \frac{1}{\delta}$ samples. Then, by using Proposition for statistical accuracy of EM updates, and noting that $p^T := \frac{\phi^T}{2\pi} \|\pi^* - \frac{1}{2}\|_1 + \min(\pi^*(1), \pi^*(2)) \leq \varepsilon \|\pi^* - \frac{1}{2}\|_1 + \min(\pi^*(1), \pi^*(2)) \equiv p(\varepsilon, \pi^*)$:

$$\begin{aligned} \frac{\|\theta^{T+1} - \text{sgn}(\rho^{T+1})\theta^*\|}{\|\theta^*\|} &= \frac{\|M_n(\theta^T, \nu^T) - \text{sgn}(\rho^T)\theta^*\|}{\|\theta^*\|} \lesssim [\phi^T]^2 \vee [\phi^T]^{\frac{3}{2}} \sqrt{\frac{d}{n}} \vee \phi^T \sqrt{\frac{\log \frac{1}{\delta}}{n}} \lesssim \varepsilon \sqrt{\frac{\log \frac{1}{\delta}}{n}} \\ \|\pi^{T+1} - \bar{\pi}^*\|_1 &= |N_n(\theta^T, \nu^T) - \text{sgn}(\rho^T) \tanh \nu^*| \lesssim \phi^T \left\|\pi^* - \frac{1}{2}\right\|_1 \vee \left[\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p^T}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right] \\ &\leq \varepsilon \left\|\pi^* - \frac{1}{2}\right\|_1 \vee \left[\frac{\log \frac{1}{\delta}/n}{\log\left(1 + \frac{\log \frac{1}{\delta}/n}{p(\varepsilon, \pi^*)}\right)} \wedge \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right] \end{aligned}$$

$\square$

APPENDIX D

DERIVATIONS FOR EM UPDATE RULES

The detailed derivations for the EM update rules are adapted from pages 20-25, Appendix B of [1], and included here in this Appendix Chapter for completeness.

A. Negative Log-Likelihood, Surrogate Function, and Q Function

Lemma D.1. *The negative expected log-likelihood $f(\theta, \pi) := -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\log p(s | \theta, \pi)]$ for the mixture model of $s := (x, y)$, $z \in [M]$ with the mixing weights $\pi^* \in \mathbb{R}^M$ and regression parameters θ is as follows.*

$$\begin{aligned} -f(\theta, \pi) &= -\text{KL}_s[p(s | \theta^*, \pi^*) || p(s | \theta, \pi)] - \mathcal{H}_s[p(s | \theta^*, \pi^*)] \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\ln p(s | \theta, \pi)] \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \ln p(s, z | \theta, \pi) + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathcal{H}_z[q_s(z)] + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \text{KL}_z[q_s(z) || p(z | s; \theta, \pi)] \end{aligned}$$

where $\text{KL}_s, \mathcal{H}_s$ are KL divergence and Shannon's entropy wrt. $s = (x, y)$;
 $\text{KL}_z, \mathcal{H}_z, \text{softmax}_z$ are KL divergence, Shannon's entropy and softmax wrt. $z \in \mathcal{Z} = [M]$;
 $\{q_s(z) | s \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}\}$ is a family of distributions wrt. $z \in \mathcal{Z} = [M]$, namely $\sum_{z \in \mathcal{Z}} q_s(z) = 1$.

Proof. Note that $p(s | \theta, \pi) = \frac{p(s, z | \theta, \pi)}{p(z | s; \theta, \pi)}$, we obtain the following expression.

$$\begin{aligned} -f(\theta, \pi) &= -\text{KL}_s[p(s | \theta^*, \pi^*) || p(s | \theta, \pi)] - \mathcal{H}_s[p(s | \theta^*, \pi^*)] \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\log p(s | \theta, \pi)] \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \left[\sum_{z \in \mathcal{Z}} q_s(z) \log p(s | \theta, \pi) \right] \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \left[\sum_{z \in \mathcal{Z}} q_s(z) \ln \left(\frac{p(s, z | \theta, \pi)}{q_s(z)} \cdot \frac{q_s(z)}{p(z | s; \theta, \pi)} \right) \right] \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \ln p(s, z | \theta, \pi) + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathcal{H}_z[q_s(z)] + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \text{KL}_z[q_s(z) || p(z | s; \theta, \pi)] \end{aligned}$$

□

Lemma D.2. *The surrogate function g^t of $f(\theta, \pi) := -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\ln p(s | \theta, \pi)]$ at $(t-1)$ -th iteration $(\theta^{t-1}, \pi^{t-1})$ be expressed as follows.*

$$-g^t(\theta, \pi) = \left\{ \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \ln p(s, z | \theta, \pi) + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathcal{H}_z[q_s(z)] \right\}_{q_s(z)=p(z|s;\theta^{t-1}, \pi^{t-1})}$$

that is $g^t(\theta, \pi) \geq f(\theta, \pi)$, and $g^t(\theta, \pi) |_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = f(\theta, \pi) |_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})}$,
 $\nabla_{\theta} g^t(\theta, \pi) |_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = \nabla_{\theta} f(\theta, \pi) |_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})}$; $\nabla_{\pi} g^t(\theta, \pi) |_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = \nabla_{\pi} f(\theta, \pi) |_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})}$;
 where $\text{KL}_s, \mathcal{H}_s$ are KL divergence and Shannon's entropy wrt. $s = (x, y)$;
 $\text{KL}_z, \mathcal{H}_z, \text{softmax}_z$ are KL divergence, Shannon's entropy and softmax wrt. $z \in \mathcal{Z} = [M]$

Proof. Let $r^t := g^t - f$, note that $r^t = \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \text{KL}_z[q_s(z) || p(z | s; \theta, \pi)]_{q_s(z)=p(z|s;\theta^{t-1}, \pi^{t-1})} \geq 0$, and

$$\begin{aligned} r^t(\theta^{t-1}, \pi^{t-1}) &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \text{KL}_z[q_s(z) || p(z | s; \theta^{t-1}, \pi^{t-1})]_{q_s(z)=p(z|s;\theta^{t-1}, \pi^{t-1})} = 0 \\ \left[\frac{dq_s(z) \ln \frac{q_s(z)}{p(z|s;\theta, \pi)}}{dp(z | s; \theta, \pi)} \right]_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} &= -\frac{q_s(z)}{p(z | s; \theta, \pi)}_{q_s(z)=p(z|s;\theta^{t-1}, \pi^{t-1}), (\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = -1 \end{aligned}$$

Hence, the gradients of r^t wrt. (θ, π) at $(t-1)$ -th iteration $(\theta^{t-1}, \pi^{t-1})$ are all zeros by the chain rule.

$$\nabla_{\theta} \text{KL}_z[q_s(z) || p(z | s; \theta, \pi)]_{q_s(z)=p(z|s;\theta^{t-1}, \pi^{t-1}), (\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = - \left[\nabla_{\theta} \sum_{z \in \mathcal{Z}} p(z | s; \theta, \pi) \right]_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = \vec{0}$$

$$\nabla_{\pi} r^t(\theta, \pi)_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \nabla_{\theta} \text{KL}_z[q_s(z) || p(z | s; \theta, \pi)]_{q_s(z)=p(z|s;\theta^{t-1}, \pi^{t-1}), (\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = \vec{0}$$

The same proof can be applied to $\nabla_{\pi} r^t(\theta, \pi)_{(\theta, \pi)=(\theta^{t-1}, \pi^{t-1})} = \vec{0}$.

□

Lemma D.3. Assuming $(z; \pi) \perp\!\!\!\perp \theta$, and $\pi \perp\!\!\!\perp s \mid z$, and $x \perp\!\!\!\perp (z; \theta, \pi)$; then

$$\begin{aligned} Q(\theta, \pi \mid \theta^{t-1}, \pi^{t-1}) &:= [\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \ln p(s, z \mid \theta, \pi)]_{q_s(z) \leftarrow p(z|s; \theta^{t-1}, \pi^{t-1})} \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \ln p(y \mid x, z; \theta) \\ &\quad + \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \ln p(x) \\ &\quad - \text{KL}_z [\pi^t(z) \parallel \pi(z)] \\ &\quad - \mathcal{H}_z [\pi^t(z)] \\ p(z \mid s; \theta, \pi) &= \text{softmax}_z (\ln \pi(z) + \ln p(y \mid x, z; \theta)) \end{aligned}$$

where $\text{KL}_s, \mathcal{H}_s$ are KL divergence and Shannon's entropy wrt. $s = (x, y)$;
 $\text{KL}_z, \mathcal{H}_z, \text{softmax}_z$ are KL divergence, Shannon's entropy and softmax wrt. $z \in \mathcal{Z} = [M]$;
 $q_s(z) \leftarrow p(z \mid s; \theta^{t-1}, \pi^{t-1})$ and $\pi^t = \{\pi(z)\}_{z \in \mathcal{Z}}, \pi^t(z) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} q_s(z)$.

Proof.

- $(z; \pi) \perp\!\!\!\perp \theta$ are independent:
 therefore $p(z \mid \theta, \pi) = p(z \mid \pi) = \pi(z)$
- $\pi \perp\!\!\!\perp (s; \theta) \mid z$ are conditional independent given z :
 then $p(\pi \mid z) = p(\pi \mid z; \theta) = p(\pi \mid z, s; \theta)$, it implies $\frac{p(s, z; \theta, \pi)}{p(z; \theta, \pi)} = \frac{p(s, z; \theta)}{p(z; \theta)}$;
 hence $p(s \mid z; \theta, \pi) = p(s \mid z; \theta)$, $p(s, z \mid \theta, \pi) = p(s \mid z; \theta, \pi) p(z \mid \theta, \pi) = p(s \mid z; \theta) \pi(z)$
- $x \perp\!\!\!\perp (z; \theta, \pi)$ are independent:
 then $p(x \mid z; \theta) = p(x)$
 hence $p(s \mid z; \theta) = p(y \mid x, z; \theta) \cdot p(x \mid z; \theta) = p(y \mid x, z; \theta) \cdot p(x)$
 therefore $p(s \mid \theta, \pi) = \sum_{z \in \mathcal{Z}} p(s, z \mid \theta, \pi) = p(x) \cdot \sum_{z \in \mathcal{Z}} \pi(z) p(y \mid x, z; \theta)$
 and $p(z \mid s; \theta, \pi) = \frac{p(s, z; \theta, \pi)}{p(s \mid \theta, \pi)} = \frac{\pi(z) \cdot p(y \mid x, z; \theta)}{\sum_{z' \in \mathcal{Z}} \pi(z') \cdot p(y \mid x, z'; \theta)} = \text{softmax}_z (\log \pi(z) + \log p(y \mid x, z; \theta))$

With the above assumptions, we obtain that $p(s, z \mid \theta, \pi) = p(y \mid x, z; \theta) \cdot p(x) \cdot \pi(z)$, further prove this Lemma. $\square$

B. Negative Log-Likelihood, Surrogate Function for General MLR and 2MLR

Lemma D.4. For MLR $y = \langle x, \theta_z^* \rangle + \varepsilon, z \in \mathcal{Z} = [M]$, $\theta := \{\theta_z\}_{z \in \mathcal{Z}}, \pi := \{\pi(z)\}_{z \in \mathcal{Z}}$, with assumptions: $(z; \pi) \perp\!\!\!\perp \theta$, and $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$, and $x \perp\!\!\!\perp (z; \theta, \pi)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned} f(\theta, \pi) &= -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \sum_{z \in \mathcal{Z}} \exp \left[-\frac{\|y - \langle \theta_z, x \rangle\|^2}{2\sigma^2} + \log \pi(z) \right] - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c \\ g^t(\theta, \pi) &= (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \|y - \langle \theta_z, x \rangle\|^2 \\ &\quad + \text{KL}_z [\pi^t(z) \parallel \pi(z)] \\ &\quad + \mathcal{H}_z [\pi^t(z)] - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c \\ &\quad - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathcal{H}_z [q_s(z)] \end{aligned}$$

where $c = -\frac{1}{2} \log (2\pi\sigma^2)$ and $q_s(z) \leftarrow p(z \mid s; \theta^{t-1}, \pi^{t-1}) = \text{softmax}_z \left(-\frac{\|y - \langle \theta_z^{t-1}, x \rangle\|^2}{2\sigma^2} + \log \pi^{t-1}(z) \right)$, and $\pi^t = \{\pi(z)\}_{z \in \mathcal{Z}}, \pi^t(z) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} q_s(z)$.

Proof. Since $(z; \pi) \perp\!\!\!\perp \theta$, and $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$, and $x \perp\!\!\!\perp (z; \theta, \pi)$, then implies $\pi \perp\!\!\!\perp s \mid z$ because of $p(\pi \mid z, s) = p(\pi \mid z, x, y) = p(\pi \mid z, x, \varepsilon) = p(\pi \mid z) = \pi(z)$. Hence, we can apply the previous Lemma.

Furthermore, $p(y, x, z; \theta) = p(\varepsilon, x, z; \theta) \Big|_{\frac{\partial \varepsilon}{\partial y}} = p(\varepsilon) \cdot p(x, z; \theta)$

$$p(y \mid x, z; \theta) = p(\varepsilon) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left(-\frac{\|\varepsilon\|^2}{2\sigma^2} \right) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left(-\frac{\|y - \langle x, \theta_z \rangle\|^2}{2\sigma^2} \right) = \mathcal{N}(\langle x, \theta_z \rangle, \sigma^2)$$

Hence, we obtain $\log p(y \mid x, z; \theta) = -\frac{\|y - \langle x, \theta_z \rangle\|^2}{2\sigma^2} + c$, where $c = -\frac{1}{2} \log (2\pi\sigma^2)$.

Subsequently, note that $g^t(\theta, \pi) = Q(\theta, \pi \mid \theta^{t-1}, \pi^{t-1}) - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathcal{H}_z [q_s(z)]$, we prove the expression for g^t by substituting $\log p(y \mid x, z; \theta)$ in the previous Lemma.

As shown in the proof of the previous Lemma, $p(s, z \mid \theta, \pi) = p(y \mid x, z; \theta) \cdot p(x) \cdot \pi(z)$.

Hence $p(s \mid \theta, \pi) = \sum_{z \in \mathcal{Z}} p(s, z \mid \theta, \pi) = p(x) \cdot \sum_{z \in \mathcal{Z}} p(y \mid x, z; \theta) \cdot \pi(z)$, we prove the expression for f by substituting $\log p(y \mid x, z; \theta)$ in the previous Lemma. $\square$

Theorem D.5. For 2MLR, $y = (-1)^{z+1}\langle x, \theta^* \rangle + \varepsilon$, $z \in \mathcal{Z} = \{1, 2\}$, $\theta_1 = \theta$, $\theta_2 = -\theta$, $\pi := \{\pi(z)\}_{z \in \mathcal{Z}}$, with assumptions: $(z; \pi) \perp\!\!\!\perp \theta$, and $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$, and $x \perp\!\!\!\perp (z; \theta, \pi)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Then the negative expected log-likelihood $f(\theta, \pi) := -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}[\ln p(s | \theta, \pi)]$ and the surrogate function $g^t(\theta, \pi)$ can be expressed as follows.

$$\begin{aligned} f &= (2\sigma^2)^{-1} \langle \theta, \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^\top \cdot \theta \rangle + \ln \cosh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \ln \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \\ &+ (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} y^2 - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \ln p(x) - c \\ g^t &= (2\sigma^2)^{-1} \langle \theta, \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^\top \cdot \theta \rangle + \log \cosh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})} \\ &- \left\langle \nabla_\theta \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})}, \theta - \theta^{t-1} \right\rangle \\ &- \left\langle \nabla_\nu \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})}, \nu - \nu^{t-1} \right\rangle \\ &+ (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} y^2 - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c \end{aligned}$$

where $\nu := \frac{\log \pi(1) - \log \pi(2)}{2}$, $c = -\frac{1}{2} \log(2\pi\sigma^2)$.

Proof. Note that $\text{sigmoid}(2t) + \text{sigmoid}(-2t) = 1$ and $\text{sigmoid}(2t) - \text{sigmoid}(-2t) = \tanh(t)$ for $\forall t$,
 $q_s(z) \leftarrow p(z | s; \theta^{t-1}, \pi^{t-1}) = \text{softmax}_z(-\frac{\|y - \langle \theta_z^{t-1}, x \rangle\|^2}{2\sigma^2} + \log \pi^{t-1}(z)) = \text{sigmoid}(2(-1)^{z+1} \left[\frac{y \langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1} \right])$,
 $\nu^{t-1} := \frac{\log \pi^{t-1}(1) - \log \pi^{t-1}(2)}{2}$.

$$\begin{aligned} &(2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \|y - \langle \theta_z, x \rangle\|^2 \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathbb{E}_{z \sim q_s(z)} \frac{\|(-1)^{z+1} y - \langle x, \theta \rangle\|^2}{2\sigma^2} \\ &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \frac{y^2 + \langle x, \theta \rangle^2}{2\sigma^2} - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \frac{y \langle x, \theta \rangle}{\sigma^2} \tanh \left(\frac{y \langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1} \right) \\ &= (2\sigma^2)^{-1} \langle \theta, \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^\top \cdot \theta \rangle - \left\langle \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh \left(\frac{y \langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1} \right) \frac{yx}{\sigma^2}, \theta \right\rangle + (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} y^2 \end{aligned}$$

Consider the other terms for g^t , note $\pi^t(z) := \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} q_s(z)$, and $\pi(z) = \text{sigmoid}(2(-1)^{z+1}\nu)$. Note that $\log 2 + \log \cosh \nu = -\left[\frac{\log \pi(1) + \log \pi(2)}{2}\right]$ and $\text{sigmoid}(2t) - \text{sigmoid}(-2t) = \tanh(t)$ for $\forall t$.

$$\begin{aligned} &\text{KL}_z[\pi^t(z) \| \pi(z)] + \mathcal{H}_z[\pi^t(z)] \\ &= -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \sum_{z \in \mathcal{Z}} q_s(z) \log \pi(z) \\ &= \log 2 + \log \cosh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh \left(\frac{y \langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1} \right) \cdot \nu \end{aligned}$$

To sum up, we obtain the following.

$$\begin{aligned} g^t &= (2\sigma^2)^{-1} \langle \theta, \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^\top \cdot \theta \rangle - \left\langle \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh \left(\frac{y \langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1} \right) \frac{yx}{\sigma^2}, \theta \right\rangle \\ &+ \log 2 + \log \cosh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh \left(\frac{y \langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1} \right) \cdot \nu \\ &- \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \mathcal{H}_z[q_s(z)] \\ &+ (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} y^2 - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c \end{aligned}$$

For the negative expectation of log-likelihood f , we show the following.

$$\begin{aligned}
f &= -\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \sum_{z \in \mathcal{Z}} \exp \left[-\frac{\|y - \langle \theta_z, x \rangle\|^2}{2\sigma^2} + \log \pi(z) \right] - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c \\
&= (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} [y^2 + \langle x, \theta \rangle^2] - \frac{\log \pi(1) + \log \pi(2)}{2} \\
&\quad - \log 2 - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c \\
&= (2\sigma^2)^{-1} \langle \theta, \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^\top \cdot \theta \rangle + \log \cosh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \\
&\quad + (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} y^2 - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c
\end{aligned}$$

Note $g^t = f$ at $(\theta, \nu) = (\theta^{t-1}, \nu^{t-1})$, by comparing the expressions f, g^t , and use $\frac{d \log \cosh(t)}{dt} = \tanh(t)$.

$$\begin{aligned}
g^t &= (2\sigma^2)^{-1} \langle \theta, \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^\top \cdot \theta \rangle + \log \cosh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})} \\
&\quad - \left\langle \nabla_\theta \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})}, \theta - \theta^{t-1} \right\rangle \\
&\quad - \left\langle \nabla_\nu \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log \cosh \left(\frac{y \langle x, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})}, \nu - \nu^{t-1} \right\rangle \\
&\quad + (2\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} y^2 - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \log p(x) - c
\end{aligned}$$

□

Theorem D.6. For 2MLR, $y = (-1)^{z+1} \langle x, \theta^* \rangle + \varepsilon$, $z \in \mathcal{Z} = \{1, 2\}$, $\theta_1 = \theta, \theta_2 = -\theta, \pi := \{\pi(z)\}_{z \in \mathcal{Z}}$, with assumptions: $(z; \pi) \perp\!\!\!\perp \theta$, and $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$, and $x \perp\!\!\!\perp (z; \theta, \pi)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Then the negative Maximum Likelihood Estimat (MLE) $f_n(\theta, \pi) := -\frac{1}{n} \sum_{i=1}^n [\log p(s_i | \theta, \pi)]$ and the surrogate function $g_n^t(\theta, \pi)$ for the dataset $\mathcal{S} := \{s_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. samples can be expressed as follows.

$$\begin{aligned}
f_n &= (2\sigma^2)^{-1} \langle \theta, \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \cdot \theta \rangle + \log \cosh \nu - \frac{1}{n} \sum_{i=1}^n \log \cosh \left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu \right) \\
&\quad + (2\sigma^2)^{-1} \frac{1}{n} \sum_{i=1}^n y_i^2 - \frac{1}{n} \sum_{i=1}^n \log p(x_i) - c \\
g_n^t &= (2\sigma^2)^{-1} \langle \theta, \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \cdot \theta \rangle + \log \cosh \nu - \frac{1}{n} \sum_{i=1}^n \log \cosh \left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})} \\
&\quad - \left\langle \nabla_\theta \frac{1}{n} \sum_{i=1}^n \log \cosh \left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})}, \theta - \theta^{t-1} \right\rangle \\
&\quad - \left\langle \nabla_\nu \frac{1}{n} \sum_{i=1}^n \log \cosh \left(\frac{y_i \langle x_i, \theta \rangle}{\sigma^2} + \nu \right) \Big|_{(\theta, \nu) \leftarrow (\theta^{t-1}, \nu^{t-1})}, \nu - \nu^{t-1} \right\rangle \\
&\quad + (2\sigma^2)^{-1} \frac{1}{n} \sum_{i=1}^n y_i^2 - \frac{1}{n} \sum_{i=1}^n \log p(x_i) - c
\end{aligned}$$

where $\nu := \frac{\log \pi(1) - \log \pi(2)}{2}$, $c = -\frac{1}{2} \log(2\pi\sigma^2)$.

Proof. This is proved by substituting $\frac{1}{n} \sum_{i=1}^n s_i := (x_i, y_i)$ for $\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}$, $s := (x, y)$ in the previous Theorem. □

C. Derivations of EM Update Rules for 2MLR

Theorem D.7. (Derivation of Population EM Update Rules: Eq. (6), (7)) For 2MLR, $y = (-1)^{z+1}\langle x, \theta^* \rangle + \varepsilon$, $z \in \mathcal{Z} = \{1, 2\}$, $\theta_1 = \theta$, $\theta_2 = -\theta$, $\pi := \{\pi(z)\}_{z \in \mathcal{Z}}$, with assumptions: $(z; \pi) \perp\!\!\!\perp \theta$, and $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$, and $x \perp\!\!\!\perp (z; \theta, \pi)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $x \sim \mathcal{N}(0, I_d)$.

The EM update rules $M(\theta^{t-1}, \nu^{t-1})$, $N(\theta^{t-1}, \nu^{t-1})$ for θ , $\tanh(\nu)$ at the population level, namely the minimizer of the surrogate g^t / the maximizer of Q , are the following.

$$\begin{aligned} M(\theta^{t-1}, \nu^{t-1}) &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) yx \\ N(\theta^{t-1}, \nu^{t-1}) &= \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) \end{aligned}$$

Proof. Take the gradients of g^t wrt. θ, ν , we obtain the following.

$$\begin{aligned} \nabla_{\theta} g^t &= (\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^{\top} \cdot \theta - (\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) yx \\ \nabla_{\nu} g^t &= \tanh \nu - \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) \end{aligned}$$

Furthermore, the Hessian of g^t wrt. θ, ν are positive-definite, we show that the solution to $\nabla_{\theta} g^t = 0$, $\nabla_{\nu} g^t = 0$ must be the minimizer of g^t . Since the Hessian of g^t at the solution is positive-definite, $\nabla_{\theta}(\nabla_{\nu} g^t) = \vec{0}$ and

$$\begin{aligned} \nabla_{\theta}^2 g^t &= (\sigma^2)^{-1} \mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^{\top} \\ \nabla_{\nu}^2 g^t &= \cosh^{-2} \nu \end{aligned}$$

Note that $\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)} x x^{\top} = I_d$ for $x \sim \mathcal{N}(0, I_d)$, we derive the expressions for EM update rules. □

Theorem D.8. (Derivation of Finite-Sample EM Update Rules: Eq. (8)) For 2MLR, $y = (-1)^{z+1}\langle x, \theta^* \rangle + \varepsilon$, $z \in \mathcal{Z} = \{1, 2\}$, $\theta_1 = \theta$, $\theta_2 = -\theta$, $\pi := \{\pi(z)\}_{z \in \mathcal{Z}}$, with assumptions: $(z; \pi) \perp\!\!\!\perp \theta$, and $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$, and $x \perp\!\!\!\perp (z; \theta, \pi)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $x \sim \mathcal{N}(0, I_d)$.

The EM update rules $M_n(\theta^{t-1}, \nu^{t-1})$, $N_n(\theta^{t-1}, \nu^{t-1})$ for θ , $\tanh(\nu)$ at the finite-sample level, namely the minimizer of the surrogate g_n^t , are the following.

$$\begin{aligned} M_n(\theta^{t-1}, \nu^{t-1}) &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^{\top} \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^n \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) yx \\ N_n(\theta^{t-1}, \nu^{t-1}) &= \frac{1}{n} \sum_{i=1}^n \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) \end{aligned}$$

Proof. This is proved by substituting $\frac{1}{n} \sum_{i=1}^n s_i := (x_i, y_i)$ for $\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}$, $s := (x, y)$ in the previous Theorem, but note that $\frac{1}{n} \sum_{i=1}^n x_i x_i^{\top} \neq I_d$. □

APPENDIX E
HELPER LEMMAS USED IN THE PROOFS OF RESULTS

A. Asymptotic Analysis of Integrals

Lemma E.1. Suppose $0 \leq a < \frac{\pi}{2}$ and $0 < b \lesssim \cos^2 a$, then

$$\begin{aligned} \int_0^{2\pi} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} &= \mathcal{O}\left(\frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right) \\ \int_0^{2\pi} \frac{b^2 \operatorname{sgn}(\sin(x+a)) |\cos x + \cos a| dx}{(b + |\sin x + \sin a|)^2} &= \mathcal{O}\left(\tan a \cdot \frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right) \end{aligned}$$

Proof. Consider the two roots of $\sin x + \sin a = 0$: $r_1 = \pi + a, r_2 = 2\pi - a$, split integrals into two regions, select $\delta = \cos a$:

$$\text{Near roots} \quad E = E_1 \cup E_2, E_i := \{(r_i + u) \bmod 2\pi \mid |u| < \delta\}, \forall i \in \{1, 2\}$$

$$\text{Away from roots} \quad \bar{E} = \bar{E}_1 \cup \bar{E}_2, \bar{E}_1 = \{x \mid r_1 + \delta \leq x \leq r_2 - \delta\}, \bar{E}_2 = \{x \mid 0 \leq x \leq r_1 - \delta \text{ or } r_2 + \delta \leq x \leq 2\pi\}$$

(i) Consider the region E near the roots:

Note $|\cos r_i| = \cos a$ and by Taylor's theorem, apply $|u| < \delta$ and $\delta = \cos a$:

$$\frac{1}{2} \cos a \cdot |u| \leq \cos a \cdot |u| - \frac{u^2}{2} \leq |\sin(r_i + u) + \sin a| \leq \cos a \cdot |u| + \frac{u^2}{2} \leq \frac{3}{2} \cos a \cdot |u|, \forall i \in \{1, 2\}$$

For the sign $\operatorname{sgn}(\sin(x+a))$,

$$\operatorname{sgn}(\sin(r_1 + u + a)) = \operatorname{sgn}(\sin(\pi + a + u + a)) = -\operatorname{sgn}(\sin(2a + u))$$

For the sign $-\operatorname{sgn}(\sin x + \sin a)$, note $-\frac{\pi}{2} < -1 \leq a - \cos a = a - \delta \leq a + u \leq a + \delta = a + \cos a \leq \frac{\pi}{2}$ for $|u| < \delta$,

$$-\operatorname{sgn}(\sin(r_1 + u) + \sin a) = -\operatorname{sgn}(\sin(\pi + a + u) + \sin a) = \operatorname{sgn}(\sin(a + u) - \sin a) = \operatorname{sgn}(u)$$

$$-\operatorname{sgn}(\sin(r_2 + u) + \sin a) = -\operatorname{sgn}(\sin(2\pi - a + u) + \sin a) = \operatorname{sgn}(\sin(a - u) - \sin a) = -\operatorname{sgn}(u)$$

(ii) Consider the region $\bar{E}$ away from the roots:

By convexity of $\sin(t)$ on $[\pi, 2\pi]$, note $r_1 = \pi + a, \delta = \cos a, |\cos(r_1 + \frac{\delta}{2})| = \cos(a + \frac{\cos a}{2}) \geq \frac{1}{2} \cos a$, for $\forall x \in \bar{E}$,

$$|\sin x + \sin a| = |\sin x - \sin r_1| \geq \sin r_1 - \sin(r_1 + \delta) \geq \delta \times \left| \cos\left(r_1 + \frac{\delta}{2}\right) \right| \geq \cos a \times \cos\left(a + \frac{\cos a}{2}\right) \geq \frac{1}{2} \cos^2 a$$

For the measure $|\bar{E}_1|$ of the set $\bar{E}_1 = \{x \mid r_1 + \delta \leq x \leq r_2 - \delta\}$, note that $\cos a \leq \frac{\pi}{2} - a \leq \frac{\pi}{2} \cos a$,

$$0 \leq |\bar{E}_1| = (r_2 - r_1) - 2\delta = (\pi - 2a) - 2\cos a \leq (\pi - 2) \cos a$$

Regarding the differences of values of $\sin$, letting $r(a) := \frac{\pi}{2} - a$ for brevity, then by Taylor's theorem

$$\begin{aligned} \sin(a + r(a)t) - \sin a &= \cos a \cdot r(a)t - \frac{\sin \theta}{2} r^2(a)t^2 \quad t \in [0, 1], \theta \in [a, a + r(a)] = \left[a, \frac{\pi}{2}\right] \\ \sin a - \sin(a - r(a)t) &= \cos a \cdot r(a)t + \frac{\sin \theta'}{2} r^2(a)t^2 \quad t \in [0, 1], \theta' \in [a - r(a), a] = \left[2a - \frac{\pi}{2}, a\right] \end{aligned}$$

By $0 \leq t^2 \leq t \leq 1, 0 \leq \sin a \leq \sin \theta \leq 1, \cos a \leq r(a) \leq \frac{\pi}{2} \cos a$ and $1 - 2\cos^2 a = \sin(2a - \frac{\pi}{2}) \leq \sin \theta' \leq \sin a$, note $\sin$ is concave on $[a, \frac{\pi}{2}]$, and $1 + \frac{1-2\cos^2 a}{2} \cdot \frac{r(a)}{\cos a} t \geq 1 - \frac{\pi}{4}$, then

$$\begin{aligned} \frac{\cos^2 a}{2} t &= \left(\frac{1 - \sin^2 a}{2}\right) t \leq (1 - \sin a)t \leq \sin(a + r(a)t) - \sin a \leq \cos a \cdot r(a)t \\ \left(1 - \frac{\pi}{4}\right) \cos a \cdot r(a)t &\leq \left(1 + \frac{1 - 2\cos^2 a}{2} \cdot \frac{r(a)}{\cos a} t\right) \cos a \cdot r(a)t \leq \sin a - \sin(a - r(a)t) \leq \left(\cos a \cdot r(a) + \frac{r^2(a)}{2}\right) t \end{aligned}$$

By applying $\cos a \leq r(a) \leq \frac{\pi}{2} \cos a$ again, thus

$$\begin{aligned} \frac{1}{\pi} \cos a \cdot r(a)t &\leq \sin(a + r(a)t) - \sin a \leq \cos a \cdot r(a)t \\ \left(1 - \frac{\pi}{4}\right) \cos a \cdot r(a)t &\leq \sin a - \sin(a - r(a)t) \leq \left(1 + \frac{\pi}{4}\right) \cos a \cdot r(a)t \end{aligned}$$

Namely, for $\forall u \in [0, r(a)] = [0, \frac{\pi}{2} - a]$,

$$\begin{aligned} \left(1 - \frac{\pi}{4}\right) \cos a \cdot u &\leq \frac{1}{\pi} \cos a \cdot u \leq \sin(a + u) - \sin a \leq \cos a \cdot u \leq \left(1 + \frac{\pi}{4}\right) \cos a \cdot u \\ \left(1 - \frac{\pi}{4}\right) \cos a \cdot u &\leq \sin a - \sin(a - u) \leq \left(1 + \frac{\pi}{4}\right) \cos a \cdot u \end{aligned}$$

(1) For the integral $\int_0^{2\pi} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2}$

In the region $E = E_1 \cup E_2$ near the roots, by using $\frac{1}{2} \cos a \cdot |u| \leq |\sin(r_i + u) + \sin a| \leq \frac{3}{2} \cos a \cdot |u|$ for $|u| < \delta$,

$$b^2 \int_{|u| < \delta} \frac{\frac{1}{2} \cos a |u| du}{\left(b + \frac{3}{2} \cos a |u|\right)^2} \leq \int_{E_i} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} \leq b^2 \int_{|u| < \delta} \frac{\frac{3}{2} \cos a |u| du}{\left(b + \frac{1}{2} \cos a |u|\right)^2} \quad \forall i \in \{1, 2\}$$

Let $t = \frac{u}{\delta} = \frac{u}{\cos a}$, and note that $\int_E = \int_{E_1} + \int_{E_2}$, then

$$2 \times \frac{\frac{1}{2}}{\left(\frac{3}{2}\right)^2} \times \frac{b^2}{\cos a} \int_0^1 \frac{t dt}{\left(\frac{\frac{3}{2}b}{\cos^2 a} + t\right)^2} \leq \int_E \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} \leq 2 \times \frac{\frac{3}{2}}{\left(\frac{1}{2}\right)^2} \times \frac{b^2}{\cos a} \int_0^1 \frac{t dt}{\left(\frac{2b}{\cos^2 a} + t\right)^2}$$

Note that $\int_0^1 \frac{t dt}{(C+t)^2} = \ln(1+C^{-1}) - \frac{1}{1+C} = \Theta(\log C^{-1})$ for $0 < C \lesssim 1$, and under the assumption $0 < b \lesssim \cos^2 a$,

$$\int_E \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} = \Theta\left(\frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right)$$

In the region $\bar{E} = \bar{E}_1 \cup \bar{E}_2$ away from the roots, note that $0 \leq |\bar{E}_1| \leq (\pi - 2) \cos a$, $|\sin x + \sin a| \geq \frac{\cos^2 a}{2}, \forall x \in \bar{E}$,

$$0 \leq \int_{\bar{E}_1} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} \leq |\bar{E}_1| \times \sup_{x \in \bar{E}_1} \frac{b^2}{|\sin x + \sin a|} \leq (\pi - 2) \cos a \times \frac{2b^2}{\cos^2 a} = \Theta\left(\frac{b^2}{\cos a}\right)$$

Under the assumption $0 < b \lesssim \cos^2 a$, then $2|\sin x + \sin a| \geq \cos^2 a \gtrsim b, \forall x \in \bar{E}$, and note $\sin x + \sin a > 0, \forall x \in \bar{E}_2$,

$$\frac{1}{9} \cdot b^2 \times \int_{-a+\delta}^{\pi+a-\delta} \frac{dx}{\sin x + \sin a} \leq \int_{\bar{E}_2} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} \leq b^2 \times \int_{-a+\delta}^{\pi+a-\delta} \frac{dx}{\sin x + \sin a}$$

Note that $\int_{-a+\delta}^{\pi+a-\delta} \frac{dx}{\sin x + \sin a} = \frac{2}{\cos a} \ln\left(\frac{\cos(a-\frac{\delta}{2})}{\sin(\frac{\delta}{2})}\right) = \Theta\left(\frac{1}{\cos a}\right)$, since $\frac{\cos(a-\frac{\delta}{2})}{\sin(\frac{\delta}{2})} \in [1.8, 3]$ for $\delta = \cos a, 0 \leq a < \frac{\pi}{2}$,

$$\int_{\bar{E}_2} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} = \Theta\left(\frac{b^2}{\cos a}\right)$$

Note that $\int_{\bar{E}} = \int_{\bar{E}_1} + \int_{\bar{E}_2}$, then

$$\int_{\bar{E}} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} = \Theta\left(\frac{b^2}{\cos a}\right)$$

Combine these two regions E and $\bar{E}$, $[0, 2\pi] = E \cup \bar{E}$

$$\int_0^{2\pi} \frac{b^2 |\sin x + \sin a| dx}{(b + |\sin x + \sin a|)^2} = \Theta\left(\frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right) + \Theta\left(\frac{b^2}{\cos a}\right) = \Theta\left(\frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right)$$

(2) For the integral $\int_0^{2\pi} \frac{b^2 \operatorname{sgn}(\sin(x+a)) |\cos x + \cos a| dx}{(b + |\sin x + \sin a|)^2}$

$$\begin{aligned} & \int_0^{2\pi} \frac{b^2 \operatorname{sgn}(\sin(x+a)) |\cos x + \cos a| dx}{(b + |\sin x + \sin a|)^2} = \int_0^{2\pi} \frac{b^2 (-1)^{\mathbf{1}_{0 \leq u \leq 2a}} (-\cos(a+u) + \cos a) du}{(b + |-\sin(a+u) + \sin a|)^2} \\ &= \left\{ -\int_0^{\frac{\pi}{2}-a} \frac{b^2 (-\cos(a+u) + \cos a) du}{(b + \sin(a+u) - \sin a)^2} + \int_{-(\frac{\pi}{2}-a)}^0 \frac{b^2 (-\cos(a+u) + \cos a) du}{(b - \sin(a+u) + \sin a)^2} \right\} \\ &+ \left\{ -\int_{\frac{\pi}{2}-a}^{2(\frac{\pi}{2}-a)} \frac{b^2 (-\cos(a+u) + \cos a) du}{(b + \sin(a+u) - \sin a)^2} + \int_{2(\frac{\pi}{2}-a)}^{3(\frac{\pi}{2}-a)} \frac{b^2 (-\cos(a+u) + \cos a) du}{(b - \sin(a+u) + \sin a)^2} \right\} + \int_{3(\frac{\pi}{2}-a)}^{\frac{3\pi}{2}+a} \frac{b^2 (-\cos(a+u) + \cos a) du}{(b - \sin(a+u) + \sin a)^2} \\ &= 2b^2 \cos a \int_0^{\frac{\pi}{2}-a} \left[-\frac{1}{(b + \sin(a+u) - \sin a)^2} + \frac{1}{(b - \sin(a-u) + \sin a)^2} \right] du + 2b^2 \cos a \int_0^{2a} \frac{du}{(b + \cos u + \sin a)^2} \\ &= -2b^2 \cos a \times 2 \sin a \int_0^{\frac{\pi}{2}-a} \frac{4 \sin^2 \frac{u}{2} (b + \cos a \sin u)}{(b + \sin(a+u) - \sin a)^2 (b - \sin(a-u) + \sin a)^2} du + 2b^2 \cos a \int_0^{2a} \frac{du}{(b + \cos u + \sin a)^2} \end{aligned}$$

Note that $\max\{\sin(a+u) - \sin a, \sin a - \sin(a-u)\} \leq (1 + \frac{\pi}{4}) \cos a \cdot u$, and $(1 - \frac{\pi}{4}) \cos a \cdot u \leq \max\{\sin(a+u) - \sin a, \sin a - \sin(a-u)\}$, for $r(a) = \frac{\pi}{2} - a$, we have $\cos a \leq r(a) \leq \frac{\pi}{2} \cos a$, $\frac{2}{\pi}u \leq \sin u \leq u$, $\frac{8}{\pi^2}u^2 \leq 4 \sin^2 \frac{u}{2} \leq u^2$, then we can establish such upper/lower bounds for the integral:

$$\begin{aligned}
& \int_0^{\frac{\pi}{2}-a} \frac{4 \sin^2 \frac{u}{2} (b + \cos a \sin u)}{(b + \sin(a+u) - \sin a)^2 (b - \sin(a-u) + \sin a)^2} du \leq \int_0^{\frac{\pi}{2}-a} \frac{(b + \cos a \cdot u) u^2 du}{(b + (1 - \frac{\pi}{4}) \cos a \cdot u)^4} \\
& \leq \frac{1}{1 - \frac{\pi}{4}} \int_0^{r(a)} \frac{u^2 du}{(b + (1 - \frac{\pi}{4}) \cos a \cdot u)^3} = \frac{1}{1 - \frac{\pi}{4}} \int_0^1 \frac{t^2 dt}{\left(\frac{b}{r(a)} + (1 - \frac{\pi}{4}) \cos a \cdot t\right)^3} \\
& \leq \frac{1}{(1 - \frac{\pi}{4})^4} \times \frac{1}{\cos^3 a} \int_0^1 \frac{t^2 dt}{\left(\frac{b/\cos^2 a}{(1 - \frac{\pi}{4}) \cdot \frac{\pi}{2}} + t\right)^3} \\
& \int_0^{\frac{\pi}{2}-a} \frac{4 \sin^2 \frac{u}{2} (b + \cos a \sin u)}{(b + \sin(a+u) - \sin a)^2 (b - \sin(a-u) + \sin a)^2} du \geq \frac{8}{\pi^2} \int_0^{\frac{\pi}{2}-a} \frac{(b + \frac{2}{\pi} \cos a \cdot u) u^2 du}{(b + (1 + \frac{\pi}{4}) \cos a \cdot u)^4} \\
& \geq \frac{16}{\pi^3 (1 + \frac{\pi}{4})} \int_0^{r(a)} \frac{u^2 du}{(b + (1 + \frac{\pi}{4}) \cos a \cdot u)^3} = \frac{16}{\pi^3 (1 + \frac{\pi}{4})} \int_0^1 \frac{t^2 dt}{\left(\frac{b}{r(a)} + (1 + \frac{\pi}{4}) \cos a \cdot t\right)^3} \\
& \geq \frac{16}{\pi^3 (1 + \frac{\pi}{4})^4} \times \frac{1}{\cos^3 a} \int_0^1 \frac{t^2 dt}{\left(\frac{b/\cos^2 a}{1 + \frac{\pi}{4}} + t\right)^3}
\end{aligned}$$

Note $\int_0^1 \frac{t^2 dt}{(C+t)^3} = \ln(1 + C^{-1}) - \frac{3+2C}{2(1+C)^2} = \Theta(\log C^{-1})$ for $0 < C \lesssim 1$, and under the assumption $0 < b \lesssim \cos^2 a$,

$$2b^2 \cos a \times 2 \sin a \int_0^{\frac{\pi}{2}-a} \frac{4 \sin^2 \frac{u}{2} (b + \cos a \sin u)}{(b + \sin(a+u) - \sin a)^2 (b - \sin(a-u) + \sin a)^2} du = \Theta\left(\tan a \cdot \frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right)$$

For the other integral, note that

$$\begin{aligned}
\cos u - \cos 2a & \geq \min\left\{\frac{1 - \cos 2a}{2a}, \sin 2a\right\} (2a - u) = \min\left\{\frac{\sin a}{a}, 2 \cos a\right\} \sin a \cdot (2a - u) \geq \cos a \cdot \sin a \cdot (2a - u) \\
\cos u + \sin a & \geq \frac{1 + 2 \sin a}{1 + \sin a} \cos^2 a + 2a \sin a \times \cos a \cdot \left(1 - \frac{u}{2a}\right) \geq \cos^2 a + 2 \sin^2 a \times \cos a \cdot \left(1 - \frac{u}{2a}\right)
\end{aligned}$$

By substitution $t = 1 - \frac{u}{2a} \in [0, 1]$, and $a \leq \frac{\pi}{2} \sin a$ for $a \in [0, \frac{\pi}{2}]$

$$\begin{aligned}
\int_0^{2a} \frac{du}{(b + \cos u + \sin a)^2} & \leq 2a \int_0^1 \frac{dt}{(\cos^2 a + 2 \sin^2 a \times \cos a \cdot t)^2} = \frac{2a}{\cos^4 a} \int_0^1 \frac{dt}{\left(1 + \frac{2 \sin^2 a}{\cos a} t\right)^2} = \frac{2a}{\cos^4 a} \cdot \frac{1}{1 + \frac{2 \sin^2 a}{\cos a}} \\
& = \frac{2a}{\cos^3 a} \cdot \frac{1}{1 + (1 - \cos a)(1 + 2 \cos a)} \leq \frac{2a}{\cos^3 a} \leq \pi \times \tan a \cdot \frac{1}{\cos^2 a}
\end{aligned}$$

Thus, we have

$$2b^2 \cos a \int_0^{2a} \frac{du}{(b + \cos u + \sin a)^2} = \mathcal{O}\left(\tan a \cdot \frac{b^2}{\cos a}\right)$$

Combine these two integrals

$$\begin{aligned}
\int_0^{2\pi} \frac{b^2 \operatorname{sgn}(\sin(x+a)) |\cos x + \cos a| dx}{(b + |\sin x + \sin a|)^2} & = \Theta\left(\tan a \cdot \frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right) - \mathcal{O}\left(\tan a \cdot \frac{b^2}{\cos a}\right) \\
& = \mathcal{O}\left(\tan a \cdot \frac{b^2}{\cos a} \log\left(\frac{\cos^2 a}{b}\right)\right)
\end{aligned}$$

□

B. Identities and Inequalities for Useful Expectations

Lemma E.2. Suppose standard Gaussian variables $g, g' \sim \mathcal{N}(0, 1)$ have the correlation coefficient $\mathbb{E}[gg'] = \sin \varphi$ with $\varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, then

$$\mathbb{E}[\text{sgn}(gg')] = \frac{2}{\pi}\varphi, \quad \mathbb{E}[|gg'|] = \frac{2}{\pi}[\varphi \sin \varphi + \cos \varphi], \quad \mathbb{E}[g^2 \text{sgn}(gg')] = \frac{2}{\pi}[\varphi + \sin \varphi \cos \varphi].$$

Proof. Let's prove the first identity (Grothendieck's identity) at the beginning. Since we can express g, g' in terms of two independent variables R, U with $R \sim r \exp(-\frac{r^2}{2})\mathbb{I}_{r \geq 0}$ (standard Rayleigh distribution with $\mathbb{E}[R^2] = 2$) and $U \sim \text{Unif}[0, 4\pi]$,

$$g = R \cos((U - \varphi)/2), \quad g' = R \sin((U + \varphi)/2)$$

Noting that $\text{sgn}(gg') = \text{sgn}(\sin(U) + \sin \varphi)$ for $R \neq 0$, we have

$$\mathbb{E}[\text{sgn}(gg')] = \mathbb{E}[\text{sgn}(\sin(U) + \sin \varphi)] = \frac{(+1) \cdot (2\pi + 4\varphi) + (-1) \cdot (2\pi - 4\varphi)}{4\pi} = \frac{2}{\pi}\varphi$$

We start the proof of the second identity by applying Price's theorem (see [63]),

$$\frac{1}{\cos \varphi} \cdot \frac{d\mathbb{E}[|gg'|]}{d\varphi} = \frac{d\mathbb{E}[|gg'|]}{d \sin \varphi} = \mathbb{E}\left[\frac{\partial^2 |gg'|}{\partial g \partial g'}\right] = \mathbb{E}[\text{sgn}(gg')] = \frac{2}{\pi}\varphi$$

Note that when $\varphi = -\frac{\pi}{2}$, $\mathbb{E}[|gg'|]_{\varphi=-\frac{\pi}{2}} = \mathbb{E}[g^2] = 1$, we have

$$\mathbb{E}[|gg'|] = \mathbb{E}[|gg'|]_{\varphi=-\frac{\pi}{2}} + \frac{2}{\pi} \int_{-\frac{\pi}{2}}^{\varphi} \varphi' \cos \varphi' d\varphi' = \frac{2}{\pi}[\varphi \sin \varphi + \cos \varphi]$$

To prove the third identity, we start by expressing $g' = \sin \varphi \cdot g + \cos \varphi \cdot h$ for some $h \sim \mathcal{N}(0, 1)$ with $h \perp g$, and applying Stein's lemma (see Lemma 2.1 of [62]), and noting that $\frac{\partial}{\partial g} \{ |g| \text{sgn}(\sin \varphi \cdot g + \cos \varphi \cdot h) \} = \text{sgn}(gg') + 2 \sin \varphi \cdot |g| \delta(g')$,

$$\mathbb{E}[g^2 \text{sgn}(gg')] = \mathbb{E}_h \mathbb{E}_g [g \cdot |g| \text{sgn}(\sin \varphi \cdot g + \cos \varphi \cdot h)] = \mathbb{E}_h \mathbb{E}_g \left[\frac{\partial |g| \text{sgn}(\sin \varphi \cdot g + \cos \varphi \cdot h)}{\partial g} \right] = \mathbb{E}[\text{sgn}(gg')] + 2 \sin \varphi \mathbb{E}[|g| \delta(g')]$$

where the former term is the first identity $\mathbb{E}[\text{sgn}(gg')] = \frac{2}{\pi}\varphi$, and the latter term can be evaluated by the symmetricity of g, g' ,

$$\mathbb{E}[|g| \delta(g')] = \mathbb{E}[|g'| \delta(g)] = \mathbb{E}_h \mathbb{E}_g [|\sin \varphi \cdot g + \cos \varphi \cdot h| \delta(g)] = \mathbb{E}_h \mathbb{E}_g [|\cos \varphi \cdot h| \delta(g)] = \cos \varphi \mathbb{E}_h [|h|] \mathbb{E}_g [\delta(g)] = \frac{1}{\pi} \cos \varphi$$

Thus, by combining these two terms, we obtain the third identity, and the proof is complete. $\square$

Lemma E.3. Suppose a random variable X has a density $X \sim f_X(x) = K_0(|x|)/\pi$ involving modified Bessel function K_0 of the second kind of order 0, then:

$$\mathbb{E}[|X|] = \frac{2}{\pi}, \quad \mathbb{E}[X^{2n}] = [(2n-1)!!]^2 = [(2n-1) \times \cdots \times 3 \times 1]^2 \quad \forall n \in \mathbb{Z}_+$$

Proof. Since the product $X = gg'$ of two independent standard Gaussian variables $g, g' \sim \mathcal{N}(0, 1)$, $\mathbb{E}[gg'] = 0$ has a density $X \sim f_X(x) = K_0(|x|)/\pi$ involving modified Bessel function K_0 of the second kind of order 0 (see page 50, Section 4.4 Bessel Function Distributions, Chapter 12 Continuous Distributions (General) of [49]), we have

$$\begin{aligned} \mathbb{E}[|X|] &= \mathbb{E}[|gg'|] = \mathbb{E}[|g|] \mathbb{E}[|g'|] = \mathbb{E}[|g|]^2 = \frac{2}{\pi} \\ \mathbb{E}[X^{2n}] &= \mathbb{E}[(gg')^{2n}] = \mathbb{E}[g^{2n}] \mathbb{E}[g'^{2n}] = \mathbb{E}[g^{2n}]^2 = [(2n-1)!!]^2 \end{aligned}$$

due to the well-known results that $\mathbb{E}[g^{2n}] = (2n-1)!!$, $\mathbb{E}[|g|] = \sqrt{\frac{2}{\pi}}$ for $g \sim \mathcal{N}(0, 1)$ (see equations (13.11), (13.14) on page 89 and page 91, Section 3 moments and other properties, Chapter 13 Normal Distributions of [49]). $\square$

Lemma E.4. Suppose $A > 0$ and a random variable X has a density $X \sim f_X(x) = K_0(|x|)/\pi$ involving modified Bessel function K_0 of the second kind of order 0, then:

$$\frac{\sqrt{12A^2+1}-1}{6A} < \mathbb{E}[\tanh(AX)X] < \frac{\sqrt{4A^2+1}-1}{2A}$$

Proof. Let $f(A) := \mathbb{E}[\tanh(AX)X]$, and $\underline{B}(A) := \frac{\sqrt{12A^2+1}-1}{6A}$, $\overline{B}(A) := \frac{\sqrt{4A^2+1}-1}{2A}$. For $A > 0$, we have:

$$0 < f(A) < \mathbb{E}[|X|] = \frac{2}{\pi}, \quad 0 < \underline{B}(A) < \lim_{A \rightarrow \infty} \underline{B}(A) = \frac{1}{\sqrt{3}}, \quad 0 < \overline{B}(A) < \lim_{A \rightarrow \infty} \overline{B}(A) = 1$$

And $f(A), \underline{B}(A), \overline{B}(A)$ are all increasing and concave with respect to $A > 0$, due to the fact that:

$$f'(A) = \mathbb{E}[(1 - \tanh^2(AX))X^2] > 0, f''(A) = -2\mathbb{E}\left[\frac{\tanh(AX)X^3}{\cosh^2(AX)}\right] < 0; \quad \underline{B}'(A) > 0, \underline{B}''(A) < 0; \overline{B}'(A) > 0, \overline{B}''(A) < 0$$

By Taylor expansions of $\tanh(t) = t - \frac{t^3}{3} + \frac{2t^5}{15} + \dots$ and $\underline{B}(A), \overline{B}(A)$ and using $\mathbb{E}[X^{2n}] = [(2n-1)!!]^2$ in Lemma E.3:

$$A - 3A^3 < f(A) < A - 3A^3 + 30A^5, \quad A - 3A^3 < \underline{B}(A) < A - 3A^3 + 18A^5, \quad A - A^3 < \overline{B}(A) < A - A^3 + 2A^5$$

When $0 < A \leq 1/4$ is small enough, we have shown:

$$f(A) < A - 3A^3 + 30A^5 = A - A^3 - 2A^3(1 - 15A^2) < A - A^3 < \overline{B}(A)$$

Similarly, we show $\underline{B}(A) < f(A)$ when A is sufficiently small by comparing the series approximations of $f(A), \underline{B}(A)$ at $A \rightarrow 0_+$, and we extent it to the range of $0 < A \leq 1/4$ by separating $\underline{B}(A)$ and $f(A)$ with piecewise linear functions based on the numerical evaluations of $f(A), f'(A), \underline{B}(A), \underline{B}'(A)$ at finite points and noting $f(A), \underline{B}(A)$ are concave.

Let's show the inequalities $\underline{B}(A) < f(A)$ and $f(A) < \overline{B}(A)$ for $1/4 \leq A \leq 5/4$ by numerical evaluations at some points $A_1 = \frac{1}{4}, A_2 = \frac{1}{3}, A_3 = \frac{1}{2}, A_4 = \frac{3}{4}, A_5 = \frac{5}{4}$.

$$\underline{B}(A) \leq \underline{B}(A_i) + \underline{B}'(A_i)(A - A_i) < f(A_i) + \frac{f(A_{i+1}) - f(A_i)}{A_{i+1} - A_i}(A - A_i) \leq f(A)$$

$$f(A) \leq f(A_i) + f'(A_i)(A - A_i) < \overline{B}(A_i) + \frac{\overline{B}(A_{i+1}) - \overline{B}(A_i)}{A_{i+1} - A_i}(A - A_i) \leq \overline{B}(A) \quad A \in [A_i, A_{i+1}], i \in \{1, 2, 3, 4\}$$

The above inequalities hold due to that $f(A), \underline{B}(A), \overline{B}(A)$ are concave and the numerical evaluations show $\underline{B}(A_i) < f(A_i) < \overline{B}(A_i), \underline{B}(A_i) + \underline{B}'(A_i)(A_{i+1} - A_i) < f(A_{i+1}), f(A_i) + f'(A_i)(A_{i+1} - A_i) < \overline{B}(A_{i+1})$ for $i \in \{1, 2, 3, 4\}$. Therefore, we have shown $\underline{B}(A) < f(A) < \overline{B}(A)$ for $0 < A \leq 5/4$ by combining these two cases. For the case of $A > 5/4$, we have shown:

$$\underline{B}(A) < \underline{B}(2.33) \approx 0.5102 < 0.5111 \approx f(5/4) < f(A) \quad \forall A \in (5/4, 2.33)$$

$$\underline{B}(A) < \lim_{A \rightarrow \infty} \underline{B}(A) = \frac{1}{\sqrt{3}} \approx 0.5774 < 0.5776 \approx f(2.33) \leq f(A) \quad \forall A \in [2.33, \infty)$$

$$f(A) \leq \lim_{A \rightarrow \infty} f(A) = \mathbb{E}[|X|] = \frac{2}{\pi} \approx 0.6366 < 0.6770 \approx \overline{B}(5/4) < \overline{B}(A) \quad \forall A \in (5/4, \infty)$$

by the fact that $f(A), \underline{B}(A), \overline{B}(A)$ are increasing and $\mathbb{E}[|X|] = \frac{2}{\pi}$ in Lemma E.3, then the proof is complete. $\square$

C. Limit Behavior and Non-Increasing Property of Some Helper Functions

Lemma E.5. Suppose these three conditions for function $F(A, k) : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ hold:

- 1) $F(A, k)$ is continuous in k for any $A > 0$, its unique root $k = k^*(A)$ of $F(A, k) = 0$ is a interior point of the compact interval $I := [a, b], 0 < a < b < \infty$.
- 2) $F(A, k)$ converges uniformly to $F_{A_0}(k) := \lim_{A \rightarrow A_0} F(A, k)$ on compact interval $I = [a, b]$ as $A \rightarrow A_0$, where A_0 is a positive number or $0_+, \infty$.
- 3) The unique root $k = k_{A_0}^*$ of $F_{A_0}(k) = 0$ is a interior point of the compact interval $I = [a, b]$; the derivative $\frac{dF_{A_0}(k)}{dk} \big|_{k=k_{A_0}^*} \neq 0$, and $\frac{dF_{A_0}(k)}{dk}$ is continuous in k on $I = [a, b]$.

Then, we have:

$$\lim_{A \rightarrow A_0} k^*(A) = k_{A_0}^*.$$

Proof. To prove that $\lim_{A \rightarrow A_0} k^*(A) = k_{A_0}^*$, we must show that for any $\varepsilon > 0$, if A_0 is finite (A_0 is a positive number or 0_+), there exists a $\delta > 0$ such that $|k^*(A) - k_{A_0}^*| < \varepsilon$ for all $A > 0$ with $|A - A_0| < \delta$; and if $A_0 = \infty$, there exists an $N > 0$ such that $|k^*(A) - k_{A_0}^*| < \varepsilon$ for all $A > N$.

By the condition 3, for the root $k = k_{A_0}^* \in (a, b)$ of $F_{A_0}(k) = 0$, there exists a $0 < \delta_{A_0} < \min(\varepsilon, k_{A_0}^* - a, b - k_{A_0}^*)$ such that

$$F_{A_0}(k_{A_0}^* + \Delta k) \cdot F_{A_0}(k_{A_0}^* - \Delta k) = (F_{A_0}(k_{A_0}^* + \Delta k) - F_{A_0}(k_{A_0}^*))(F_{A_0}(k_{A_0}^* - \Delta k) - F_{A_0}(k_{A_0}^*)) < 0$$

for any $\Delta k \in (0, \delta_{A_0})$. Let's select $\Delta k = \delta_{A_0}/2$, then we have:

$$F_{A_0}\left(k_{A_0}^* + \frac{\delta_{A_0}}{2}\right) \cdot F_{A_0}\left(k_{A_0}^* - \frac{\delta_{A_0}}{2}\right) < 0.$$

Without loss of generality, we may assume $F_{A_0}\left(k_{A_0}^* + \frac{\delta_{A_0}}{2}\right) > 0, F_{A_0}\left(k_{A_0}^* - \frac{\delta_{A_0}}{2}\right) < 0$, and select:

$$m_{A_0} := \min\left(\left|F_{A_0}\left(k_{A_0}^* + \frac{\delta_{A_0}}{2}\right)\right|, \left|F_{A_0}\left(k_{A_0}^* - \frac{\delta_{A_0}}{2}\right)\right|\right) > 0$$

By the condition 2, since $k_{A_0}^* + \frac{\delta_{A_0}}{2}, k_{A_0}^* - \frac{\delta_{A_0}}{2} \in [a, b] = I$, there exists a $\delta > 0$ for all $A > 0$ with $|A - A_0| < \delta$ if A_0 is finite; or there exists an $N > 0$ for all $A > N$ if $A_0 = \infty$ such that:

$$\max \left(\left| F(A, k_{A_0}^* + \frac{\delta_{A_0}}{2}) - F_{A_0}(k_{A_0}^* + \frac{\delta_{A_0}}{2}) \right|, \left| F(A, k_{A_0}^* - \frac{\delta_{A_0}}{2}) - F_{A_0}(k_{A_0}^* - \frac{\delta_{A_0}}{2}) \right| \right) \leq \sup_{k \in I} |F_{A_0}(k) - F_{A_0}(k_{A_0}^*)| < \frac{m_{A_0}}{2}.$$

Therefore, the signs of function values will not change as long as A is close to A_0 .

$$\begin{aligned} F(A, k_{A_0}^* + \frac{\delta_{A_0}}{2}) &\geq F_{A_0}(k_{A_0}^* + \frac{\delta_{A_0}}{2}) - \left| F(A, k_{A_0}^* + \frac{\delta_{A_0}}{2}) - F_{A_0}(k_{A_0}^* + \frac{\delta_{A_0}}{2}) \right| > m_{A_0} - \frac{m_{A_0}}{2} = \frac{m_{A_0}}{2} > 0 \\ F(A, k_{A_0}^* - \frac{\delta_{A_0}}{2}) &\leq F_{A_0}(k_{A_0}^* - \frac{\delta_{A_0}}{2}) + \left| F(A, k_{A_0}^* - \frac{\delta_{A_0}}{2}) - F_{A_0}(k_{A_0}^* - \frac{\delta_{A_0}}{2}) \right| < -m_{A_0} + \frac{m_{A_0}}{2} = -\frac{m_{A_0}}{2} < 0 \end{aligned}$$

By condition 1, the application of the intermediate value theorem, under the condition that $F(A, k)$ is continuous in k for any $A > 0$ and the uniqueness of the root $k = k^*(A)$ of $F(A, k) = 0$, ensures that:

$$k^*(A) \in \left(k_{A_0}^* - \frac{\delta_{A_0}}{2}, k_{A_0}^* + \frac{\delta_{A_0}}{2} \right) \subset (a, b) \subset I$$

Then, we have:

$$|k^*(A) - k_{A_0}^*| < \max \left(\left| (k_{A_0}^* - \frac{\delta_{A_0}}{2}) - k_{A_0}^* \right|, \left| (k_{A_0}^* + \frac{\delta_{A_0}}{2}) - k_{A_0}^* \right| \right) = \frac{\delta_{A_0}}{2} < \min(\varepsilon, k_{A_0}^* - a, b - k_{A_0}^*)/2 < \varepsilon$$

for all $A > 0$ with $|A - A_0| < \delta$ if A_0 is finite (A_0 is a positive number or 0_+); or for all $A > N$ if $A_0 = \infty$. Therefore, we have shown that $\lim_{A \rightarrow A_0} k^*(A) = k_{A_0}^*$, and the proof is complete. $\square$

Lemma E.6. Suppose $s \in [0, 1]$ and $\delta \geq 0$, then the function $H(s, \delta)$ is non-decreasing with respect to $s \in [0, 1]$ given fixed $\delta \geq 0$, and $H(s, \delta)$ is defined as:

$$H(s, \delta) = \left(s \arcsin \frac{s}{\sqrt{1+\delta}} + \sqrt{1-s^2+\delta} \right)^2 + \frac{(1+\delta)(1-s^2)}{1-s^2+\delta}$$

Proof. When $\delta = 0$, we have $H(s, 0) = (s \arcsin s + \sqrt{1-s^2})^2$, which is non-decreasing with respect to $s \in [0, 1]$, due to the fact that $\partial_s H(s, 0) = 2(s \arcsin s + \sqrt{1-s^2}) \cdot \arcsin s > 0, \forall s \in (0, 1)$. When $\delta > 0$, we introduce following notations to show $\partial_s H(s, \delta) > 0, \forall s \in (0, 1)$:

$$r := 1 - s^2, y := \arcsin \frac{s}{\sqrt{1+\delta}}, q := \sqrt{1-s^2+\delta}, K(s) := sy + q$$

Then we have $H(s, \delta) = K(s)^2 + \frac{(1+\delta)r}{r+\delta}$ and

$$y'(s) = 1/q, q' = -s/q, K'(s) = y \implies \partial_s H(s, \delta) = 2K(s)K'(s) + \partial_s \left(\frac{(1+\delta)r}{r+\delta} \right) = 2 \left[K(s)y - \frac{s(1+\delta)\delta}{q^4} \right]$$

By substitution $s = \sqrt{1+\delta} \sin y, q = \sqrt{1+\delta} \cos y, K(s) = \sqrt{1+\delta}(y \sin y + \cos y)$ for some $y \in (0, 1)$, we have

$$K(s)y - \frac{s(1+\delta)\delta}{q^4} = \frac{\sin y}{\sqrt{1+\delta} \cos^4 y} \Phi(y) \geq \frac{\sin y}{\sqrt{1+\delta} \cos^4 y} \Phi(0_+) > 0$$

where the last inequality is due to the fact that $\Phi(y)$ is non-decreasing with respect to $y \in (0, \frac{\pi}{2})$ given fixed δ , and $\Phi(0_+) = 1$, and $\Phi(s)$ is defined as:

$$\Phi(y) = (1+\delta)y(y + \cot y) \cos^4 y - \delta \implies \partial_y \Phi(y) = \frac{(1+\delta) \cos^3 y (1 + 4y \sin y)}{2 \sin^2 y} \cos(2y) (\tanh(2y) - 2y) \geq 0$$

Therefore, $\partial_s H(s, \delta) > 0, \forall s \in (0, 1)$ and $H(s, \delta)$ is non-decreasing with respect to $s \in [0, 1]$ given fixed $\delta \geq 0$. $\square$

D. Probability Concentration Inequality for Bernoulli Random Variables

Lemma E.7. Let $q := \max(p, 1-p)$ for $V_i \stackrel{\text{iid}}{\sim} \text{Bern}(p), \forall i \in [n] := \{1, 2, \dots, n\}$, then for $t \in \mathbb{R}_{\geq 0}$

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right| \geq t \right) \leq 2 \exp(-2nt^2)$$

for $t \in [e(1-q), q] \neq \emptyset$

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right| \geq t \right) \leq 2 \exp \left(-n \left\{ \frac{t}{q} \left[\ln \frac{t}{(1-q)} - 1 \right] + \frac{t^2}{2q^2} \right\} \right)$$

for $t \in (q, \infty)$, we have $\mathbb{P} \left(\left| \frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right| \geq t \right) = 0$. Moreover, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right| \lesssim \min \left(\frac{\log \frac{1}{\delta}/n}{\log \left(1 + \frac{\log \frac{1}{\delta}/n}{1-q} \right)}, \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$$

Proof. Let's denote $q := \max(p, 1-p)$, $V' := V_i - \mathbb{E}[V_i]$, $i \in [n]$, thus $\mathbb{E}[V'^2] = \text{Var}[V_i] = p(1-p) = q(1-q)$, $|V'| \leq q$. With Chenorff bound and let $\psi(\lambda) := \mathbb{E}[\exp(\lambda(V_i - \mathbb{E}[V_i]))] = -\lambda p + \ln(1 + p(\exp(\lambda) - 1)) \leq \frac{\lambda^2}{8}$

$$\log \mathbb{P} \left[\sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \geq nt \right] \leq \inf_{\lambda > 0} \left\{ \log \mathbb{E} \left[\exp \left(\lambda \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right) \right] - \lambda nt \right\} \leq \inf_{\lambda > 0} \left\{ n \left[\frac{\lambda^2}{8} - \lambda t \right] \right\} = -2nt^2$$

Hence, one side of the first probability inequality is proved, the other side is proved by replacing V_i with $1 - V_i$.

Let's focus on next concentration inequality. We begin with bounding $\psi(\lambda)$, note that $2 \sinh(x) > \exp(x) - x - 1$, $\forall x > 0$.

$$\psi(\lambda) = 1 + \sum_{k \geq 2} \frac{\mathbb{E}[V'^k]}{k!} \lambda^k \leq 1 + \sum_{k \geq 2} \frac{\mathbb{E}[V'^2] \cdot q^{k-2}}{k!} \lambda^k \leq 1 + \frac{(1-q)}{q} \{ \exp(q\lambda) - q\lambda - 1 \} \leq 1 + 2 \frac{(1-q)}{q} \sinh(q\lambda)$$

Let $\mu := 2 \frac{(1-q)}{q} \sinh(q\lambda)$, then $\lambda = \frac{1}{q} \text{arcsinh} \left(\frac{q}{2(1-q)} \mu \right)$, and $\mu' := \frac{q}{(1-q)} \mu$, $\gamma := \frac{(1-q)}{q} \in (0, 1]$, $\tau := \frac{t}{(1-q)}$.

$$\begin{aligned} \log \mathbb{P} \left(\frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \geq t \right) &\leq n \inf_{\lambda > 0} [\ln \psi(\lambda) - \lambda t] \leq n \inf_{\lambda > 0} \left\{ \ln \left[1 + 2 \frac{(1-q)}{q} \sinh(q\lambda) \right] - \lambda t \right\} \\ &= n\gamma \inf_{\mu' > 0} \left\{ \frac{1}{\gamma} \ln(1 + \gamma\mu') - \tau \text{arcsinh} \left(\frac{\mu'}{2} \right) \right\} \leq n\gamma \inf_{\mu' > 0} \left\{ \frac{1}{\gamma} \ln(1 + \gamma\mu') - \tau \ln(\mu') \right\} \end{aligned}$$

If $t > q$, then $\tau := \frac{t}{(1-q)} \in [\frac{1}{\gamma}, \infty)$, $n\gamma \inf_{\mu' > 0} \left\{ \frac{1}{\gamma} \ln(1 + \gamma\mu') - \tau \ln(\mu') \right\} = -\infty$. If $t \in [e(1-q), q]$, $\tau := \frac{t}{(1-q)} \in [e, \frac{1}{\gamma}]$.

$$\begin{aligned} &n\gamma \inf_{\mu' > 0} \left\{ \frac{1}{\gamma} \ln(1 + \gamma\mu') - \tau \log(\mu') \right\} \\ &= n\gamma \left\{ \frac{1}{\gamma} \ln(1 + \gamma\mu') - \tau \ln(\mu') \right\}_{\mu' = \frac{\tau}{1-\gamma\tau}} = n\gamma \left\{ -\tau \ln \tau + \frac{1}{\gamma} [-(1 - \gamma\tau) \ln(1 - \gamma\tau)] \right\} \\ &\leq n\gamma \left\{ -\tau \ln \tau + \frac{\gamma\tau}{\gamma} \left(1 - \frac{\gamma\tau}{2} \right) \right\} \leq n\gamma \left\{ -\tau [\ln \tau - 1] - \frac{\gamma}{2} \tau^2 \right\} = n \left\{ -\frac{t}{q} \left[\ln \frac{t}{(1-q)} - 1 \right] - \frac{t^2}{2q^2} \right\} \end{aligned}$$

Therefore, one side of the other probability inequality is proved, similarly, the other side is proved by replacing V_i with $1 - V_i$.

Also, we show the probability is 0 when $t \geq q$.

By the first concentration inequality, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right| \lesssim \sqrt{\frac{\log(1/\delta)}{n}}$$

By the second concentration inequality and letting $t = 2 \frac{\ln(2/\delta)}{n} / \ln(1 + \frac{\ln(2/\delta)}{n \times e(1-q)}) \geq 2e(1-q)$, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i \in [n]} (V_i - \mathbb{E}[V_i]) \right| \leq 2 \frac{\ln(2/\delta)}{n} / \ln \left(1 + \frac{\ln(2/\delta)}{n \times e(1-q)} \right) \asymp \frac{\log \frac{1}{\delta}/n}{\log \left(1 + \frac{\log \frac{1}{\delta}/n}{1-q} \right)}$$

Since the failure probability is less than δ , by using $2x/\sqrt{1+x} > \ln(1+x)$ for $x = \frac{\ln(2/\delta)}{n \times e(1-q)} > 0$,

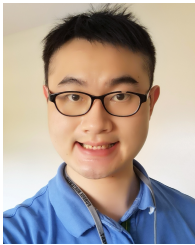
$$2 \exp \left(-nt \log \frac{t}{e(1-q)} \right)_{t=2 \frac{\ln(2/\delta)}{n} / \ln(1 + \frac{\ln(2/\delta)}{n \times e(1-q)})} < 2 \exp \left(-n \frac{\ln(2/\delta)}{n} \right) = \delta$$

The proof is completed by combining the bounds for $t \in \mathbb{R}_{\geq 0}$ and $t \in [e(1-q), q]$. $\square$

REFERENCES

- [1] Z. Luo and A. Hashemi, "Unveiling the cycloid trajectory of EM iterations in mixed linear regression," in *International Conference on Machine Learning*. PMLR, 2024, pp. 33 329–33 382. 1, 2, 8, 9, 10, 11, 12, 13, 17, 23, 25, 26, 28, 29, 33, 34, 37
- [2] E. M. Beale and R. J. Little, "Missing values in multivariate analysis," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 37, no. 1, pp. 129–145, 1975. 1
- [3] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999. 1
- [4] W. Kong, R. Somani, S. Kakade, and S. Oh, "Robust meta-learning for mixed linear regression with small batches," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4683–4696. 1
- [5] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh, "Meta-learning for mixed linear regression," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 13–18 Jul 2020, pp. 5394–5404. 1
- [6] Y. Shen and S. Sanghavi, "Iterative least trimmed squares for mixed linear regression," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. 1
- [7] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," in *Conference On Learning Theory*. PMLR, 2018, pp. 1125–1144. 1
- [8] 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)*, vol. 39, no. 1, pp. 1–22, 1977. 2
- [9] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of statistics*, pp. 95–103, 1983. 2
- [10] R. D. De Veaux, "Mixtures of linear regressions," *Computational Statistics & Data Analysis*, vol. 8, no. 3, pp. 227–245, 1989. 2
- [11] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994. 2
- [12] M. I. Jordan and L. Xu, "Convergence results for the EM approach to mixtures of experts architectures," *Neural networks*, vol. 8, no. 9, pp. 1409–1431, 1995. 2
- [13] M. Wedel and W. S. DeSarbo, "A mixture likelihood approach for generalized linear models," *Journal of classification*, vol. 12, pp. 21–55, 1995. 2
- [14] C. Cai, S. Sanghavi, and H. Vikalo, "Structured low-rank matrix factorization for haplotype assembly," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 647–657, 2016. 2
- [15] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015. 2
- [16] A. Sankaranarayanan, H. Vikalo, and F. Baccelli, "Comhapdet: a spatial community detection algorithm for haplotype assembly," *BMC genomics*, vol. 21, pp. 1–14, 2020. 2
- [17] J. M. Klusowski, D. Yang, and W. D. Brinda, "Estimating the coefficients of a mixture of two linear regressions by expectation maximization," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3515–3524, 2019. 2, 3, 12
- [18] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," *arXiv preprint arXiv:1312.7006*, 2013. 2
- [19] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017. 2, 3, 11
- [20] A. Reiszadeh, K. Gattmiry, and A. Ozdaglar, "Em for mixture of linear regression with clustered data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 2341–2349. 2
- [21] 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*. PMLR, 2019, pp. 2055–2110. 2, 3, 6, 7
- [22] 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*. PMLR, 2021, pp. 1405–1413. 2, 3, 10, 11
- [23] N. Weinberger and G. Bresler, "The EM algorithm is adaptively-optimal for unbalanced symmetric gaussian mixtures," *Journal of Machine Learning Research*, vol. 23, no. 103, pp. 1–79, 2022. 3
- [24] S. Dasgupta and L. Schulman, "A probabilistic analysis of EM for mixtures of separated, spherical Gaussians," *Journal of Machine Learning Research*, vol. 8, no. 7, pp. 203–226, 2007. 3
- [25] R. Zhao, Y. Li, and Y. Sun, "Statistical convergence of the EM algorithm on Gaussian mixture models," *Electronic Journal of Statistics*, vol. 14, pp. 632–660, 2020. 3
- [26] 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," *Advances in neural information processing systems*, vol. 29, 2016. 3
- [27] Y. Chen and X. Xi, "Likelihood landscape and local minima structures of Gaussian mixture models," *arXiv preprint arXiv:2009.13040*, 2020. 3
- [28] W. Qian, Y. Zhang, and Y. Chen, "Structures of spurious local minima in k-means," *IEEE Transactions on Information Theory*, vol. 68, no. 1, pp. 395–422, 2022. 3
- [29] Y. Zhang, Q. Qu, and J. Wright, "From symmetry to geometry: Tractable nonconvex problems," *arXiv preprint arXiv:2007.06753*, 2020. 3
- [30] J. Hong, W. Qian, Y. Chen, and Y. Zhang, "A geometric approach to k-means," *arXiv preprint arXiv:2201.04822*, 2022. 3
- [31] A. Katsevich and A. S. Bandeira, "Likelihood maximization and moment matching in low SNR Gaussian mixture models," *Communications on Pure and Applied Mathematics*, vol. 76, no. 4, pp. 788–842, 2023. 3
- [32] J. M. Klusowski and W. Brinda, "Statistical guarantees for estimating the centers of a two-component Gaussian mixture by EM," *arXiv preprint arXiv:1608.02280*, 2016. 3
- [33] J. Xu, D. J. Hsu, and A. Maleki, "Global analysis of expectation maximization for mixtures of two Gaussians," *Advances in Neural Information Processing Systems*, vol. 29, 2016. 3
- [34] C. Daskalakis, C. Tzamos, and M. Zampetakis, "Ten steps of EM suffice for mixtures of two Gaussians," in *Proceedings of the 2017 Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, S. Kale and O. Shamir, Eds., vol. 65. PMLR, 07–10 Jul 2017, pp. 704–710. 3
- [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*, vol. 4, no. 3, 2021. 3
- [36] W. Qian, Y. Zhang, and Y. Chen, "Global convergence of least squares EM for demixing two log-concave densities," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 3
- [37] W. Qian, *Local Minima in Mixture Problems and Their Algorithmic Implications*. Cornell University, 2020. 3
- [38] M. Ndaoud, "Sharp optimal recovery in the two-component Gaussian mixture model," *arXiv preprint arXiv:1812.08078*, 2018. 3
- [39] J. Kwon *et al.*, "Statistical learning with latent variables: mixture models and reinforcement learning," Ph.D. dissertation, The University of Texas at Austin, 2022. 3
- [40] 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*, vol. 64, no. 3, pp. 1738–1766, 2018. 3
- [41] Y. Xu and A. Zeevi, "Towards optimal problem dependent generalization error bounds in statistical learning theory," *arXiv preprint arXiv:2011.06186*, 2020. 3
- [42] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating minimization for mixed linear regression," in *International Conference on Machine Learning*. PMLR, 2014, pp. 613–621. 3

- [43] —, “Solving a mixture of many random linear equations by tensor decomposition and alternating minimization,” *arXiv preprint arXiv:1608.05749*, 2016. 3
- [44] A. Ghosh and R. Kannan, “Alternating minimization converges super-linearly for mixed linear regression,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 26–28 Aug 2020, pp. 1093–1103. 3, 12
- [45] J. Kwon and C. Caramanis, “EM converges for a mixture of many linear regressions,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1727–1736. 3
- [46] R. Dwivedi, N. Ho, K. Khamaru, M. Wainwright, M. Jordan, and B. Yu, “Sharp analysis of expectation-maximization for weakly identifiable models,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1866–1876. 3
- [47] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, “Singularity, misspecification and the convergence rate of EM,” *The Annals of Statistics*, vol. 48, no. 6, pp. 3161 – 3182, 2020. 3
- [48] Z. Luo and A. Hashemi, “Characterizing evolution in expectation-maximization estimates for overspecified mixed linear regression,” *arXiv preprint arXiv:2508.10154*, 2025. 3
- [49] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Distributions in Statistics: Continuous Univariate Distributions, Volume 1, Second Edition*. Houghton-Mifflin, 1970. 5, 18, 19, 45
- [50] E. Lehman, *Mathematics for computer science*. Samurai Media Limited, 2017. 5
- [51] J. Kwon, W. Qian, Y. Chen, C. Caramanis, D. Davis, and N. Ho, “Global optimality of the em algorithm for mixtures of two-component linear regressions,” *IEEE Transactions on Information Theory*, 2024. 6, 7, 12
- [52] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47. 11, 29
- [53] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” *Annals of statistics*, pp. 1302–1338, 2000. 11, 30, 32
- [54] T. Hastie, R. Tibshirani, J. Friedman *et al.*, “The elements of statistical learning,” 2009. 13
- [55] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014. 20
- [56] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2019. 20, 21
- [57] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976. 21
- [58] F. W. Olver, D. W. Lozier, R. Boisvert, and C. W. Clark, *NIST handbook of mathematical functions*. Cambridge university press, 2010. 22
- [59] J. W. Harris and H. Stöcker, *Handbook of mathematics and computational science*. Springer Science & Business Media, 1998. 26
- [60] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48. 30, 32
- [61] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. 31
- [62] N. Ross, “Fundamentals of Stein’s method,” *Probability Surveys*, vol. 8, no. none, pp. 210 – 293, 2011. 33, 45
- [63] R. Price, “A useful theorem for nonlinear devices having gaussian inputs,” *IRE Transactions on Information Theory*, vol. 4, no. 2, pp. 69–72, 1958. 45



Zhankun Luo (Graduate Student Member, IEEE) received the B.S. degree in telecommunication engineering from Beijing Institute of Technology, China, in 2019, and the M.S.E. degree in electrical and computer engineering in 2021 from Purdue University Northwest, Hammond, IN, USA. He is currently pursuing the Ph.D. degree with the Elmore Family School of Electrical and Computer Engineering, Purdue University, USA. His research interests include machine learning, statistical learning, generative diffusion models, and optimization for distributed and decentralized algorithms.



Abolfazl Hashemi (Member, IEEE) received the B.Sc. degree in electrical engineering from Sharif University of Technology, Iran, in 2014, and the M.S.E. and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin, USA, in 2016 and 2020, respectively. From 2020 to 2021, he was a Postdoctoral Scholar with the Oden Institute for Computational Engineering and Sciences, University of Texas at Austin. Since 2021, he has been an Assistant Professor with the Elmore Family School of Electrical and Computer Engineering, Purdue University. His research interests include large-scale optimization. He was the 2019 Schmidt Science Fellows Award nominee from UT Austin, the recipient of the Iranian National Elite Foundation Fellowship, and the Best Student Paper Award finalist at the 2018 American Control Conference.