Unveiling the Cycloid Trajectory of EM Iterations in Mixed Linear Regression

Zhankun Luo [1] **Abolfazl Hashemi** [1]

Abstract

We study the trajectory of iterations and the convergence rates of the Expectation-Maximization (EM) algorithm for two-component Mixed Linear Regression (2MLR). The fundamental goal of MLR is to learn the regression models from unlabeled observations. The EM algorithm finds extensive applications in solving the mixture of linear regressions. Recent results have established the super-linear convergence of EM for 2MLR in the noiseless and high SNR settings under some assumptions and its global convergence rate with random initialization has been affirmed. However, the exponent of convergence has not been theoretically estimated and the geometric properties of the trajectory of EM iterations are not well-understood. In this paper, first, using Bessel functions we provide explicit closed-form expressions for the EM updates under all SNR regimes. Then, in the noiseless setting, we completely characterize the behavior of EM iterations by deriving a recurrence relation at the population level and notably show that all the iterations lie on a certain cycloid. Based on this new trajectory-based analysis, we exhibit the theoretical estimate for the exponent of super-linear convergence and further improve the statistical error bound at the finite-sample level. Our analysis provides a new framework for studying the behavior of EM for Mixed Linear Regression.

1. Introduction

A mixture model of parameterized distributions, such as the Mixture of Linear Regression (MLR) and Gaussian Mixture Model (GMM), is remarkably powerful for modeling intricate relationships in practice. It is highly suitable to address the challenges arising from data with corruptions, missing values, and latent variables (Beale & Little, 1975). In this paper, we focus on the symmetrical two-component mixed linear regression (2MLR) that can be expressed as follows:

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

where ε is the addictive noise, $s = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ is a pair of the covariate and response random variable, $z \in \{1, 2\} \sim \mathcal{CAT}(\pi^*)$ represents the latent variable, namely the label of data, and θ^*, π^* are the true values for the regression parameters and the mixing weights, respectively.

Maximum Likelihood Estimation (MLE) provides a systematic framework to study such models. However, computing the Maximum Likelihood Estimate (MLE) for high-dimensional data is intractable due to its non-convexity and numerous spurious local maxima. Various approaches have been proposed to handle this intractable problem. Tipping & Bishop (1999) adopted PCA by connecting the inherent geometric property and probabilistic interpretation with Gaussian covariates and errors. Kong et al. (2020a;b) employed a meta-learning approach to learn the parameters of MLR with small batches. Shen & Sanghavi (2019) proposed an iterative variant of the least trimmed squares to handle MLR with corruptions. Another competitor is the moment-based method, in conjunction with the gradient descent algorithm (Li & Liang, 2018). Moreover, the Expectation Maximization (EM) method stands out for its computational efficiency and ease of practical implementation. In the context of (1), EM estimates the regression parameters and the mixing weights from observations. It operates in two steps: E-step computes the expected log-likelihood using the current parameter estimate; M-step updates the parameters to maximize the expected log-likelihood compute in the E-step. These steps serve to maximize the lower bound on the MLE objective iteratively until convergence.

Dempster et al. (1977) presented the modern EM algorithm and demonstrated its likelihood to be monotonically increasing with EM updates. Theoretically, Wu (1983) established the global convergence of a unimodal likelihood under some regularity conditions. Empirically, EM showed success in the MLR problem (De Veaux, 1989; Jordan & Jacobs, 1994; Jordan & Xu, 1995). Additionally, Wedel &

[1]School of Electrical and Computer Engineering, Purdue University, IN, USA. Correspondence to: Zhankun Luo <luo333@purdue.edu>, Abolfazl Hashemi <abolfazl@purdue.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

DeSarbo (1995) introduced a framework of EM for MLR involving unknown latent variables.

Related Works. Both MLR and GMM can be viewed as instances of subspace clustering, thus sharing similarities in the analysis of EM. Dasgupta & Schulman (2007) showed parameterized well-separated spherical Gaussians can be learned to near-optimal precision using a variant of EM. Furthermore, Zhao et al. (2020) illustrated the linear global convergence of EM with well-separated spherical Gaussians and initialization within a ball around the truth. For GMM with $k \geq 3$ components, Jin et al. (2016) demonstrated the EM with a random initialization is frequently trapped in local minima with high probability, and local maxima can exhibit arbitrarily inferior likelihood than that of any global maximum. Chen & Xi (2020) and Qian et al. (2022) characterized the only types of local minima for EM and k-means (EM with hard labels) in GMM under a separation condition. By leveraging the characterized structures of local minimum, a general framework was proposed to escape local minima (Zhang et al., 2020; Hong et al., 2022), unifying variants of k-means from a geometric perspective. Katsevich & Bandeira (2023) revealed the link between EM and the moment method for GMM via an asymptotic expansion of log-likelihood in low Signal-to-Noise Ratio (SNR). So far, the specific case of GMM with $k = 2$ components (2GMM) has been studied intensively. The global convergence of EM with random initialization for spherical 2GMM was established in (Klusowski & Brinda, 2016; Xu et al., 2016; Daskalakis et al., 2017). Wu & Zhou (2021) elaborated the convergence result in all SNR regimes, while constraining the initialization within a very small radius. Qian et al. (2019); Qian (2020) extended this convergence result from the spherical Gaussian to rotation-invariant log-concave densities. Ndaoud (2018) examined the phase transition threshold of SNR for the exact recovery of 2GMM. Similarly, EM for MLR with two components (2MLR) with random initialization converges globally. Balakrishnan et al. (2017) firstly proved the global convergence of EM for 2MLR with valid initialization within a ball around the truth. Klusowski et al. (2019) extended the convergence result in the high SNR regime for the case where the cosine angle between the initial parameters and the truth is sufficiently large. Kwon et al. (2022; 2019) confirmed that EM for 2MLR converges from a random initialization with high probability. Chen et al. (2018) bounded the statistical error of EM for 2MLR in different SNR regimes. Xu & Zeevi (2020) illustrated the generalization error bounds of log-likelihood of the first-order EM for 2MLR. Kwon et al. (2021) further studied the statistical error and the convergence rate of EM for 2MLR under all regimes of SNR. Yi et al. (2014; 2016) considered Alternating Minimization (AM), an EM variant with hard labels, for 2MLR in the setting of no noise. Accordingly, Ghosh & Kannan (2020) demonstrated a super-linear convergence rate of AM for 2MLR in the noiseless setting within a specific convergence region. Kwon et al. (2021) generalized the noiseless setting to the high SNR regime while retaining the super-linear convergence. Kwon & Caramanis (2020) provided a convergence analysis of EM for MLR with multiple components, covering the most general scenarios.

Previous works on the convergence analysis of EM for 2MLR have overlooked the existence of unbalanced mixing weights and assumed a balanced setting. Dwivedi et al. (2020a;b) accounted for unbalanced mixing weights and revealed a sharp contrast in statistical error and convergence rate between unbalanced and balanced cases, for the special case of no separation of parameters. In the previous convergence analysis of EM, the location scale of GMM and the noise variance of MLR are fixed. In light of this, Ren et al. (2022) proposed an EM variant for 2GMM with the unknown location scale to speed up the convergence of EM. Chandrasekher et al. (2021) devised a tool that demonstrated noteworthy potential, employing Gordon state evolution update (Thrampoulidis et al., 2014; 2015) to accurately estimate both the statistical error and the convergence rate at the finite-sample level.

Contributions. In this paper, we propose a framework that offers explicit closed-form expressions with Bessel functions (Chapter 10 of (Olver et al., 2010)) for the EM updates of 2MLR, enabling the analysis of convergence rate across all SNR regimes. Moreover, our framework includes both scenarios where the mixing weights are balanced and unbalanced. More specifically, we focus on EM updates in the noiseless setting, and present the following contributions:

- We derive the recurrence relation, and further show the cycloid trajectory of EM iterations at the population level.
- We establish the super-linear convergence without restrictions in previous works (Ghosh & Kannan, 2020; Kwon et al., 2021).
- We conduct a finer analysis for the statistical errors in regression parameters, and explore how the error in mixing weights is influenced by the angle formed by the EM iteration and the true regression parameters, the true mixing weights.

2. Problem Setup

Notation. Consider the 2MLR model in (1). Let n denote the number of samples $\mathcal{S} := \{x_i, y_i\}_{i=1}^n$ used for each EM update, $\{z_i\}_{i=1}^n$ be the values of latent variable for these samples. Further, σ^2 denotes the noise variance, $\eta := \frac{\|\theta^*\|}{\sigma}$ is the signal-to-noise ratio (SNR), and $\bar{\theta} := \frac{\theta}{\sigma}, \bar{\theta}^* := \frac{\theta^*}{\sigma}$

are the normalized parameters. $f(\nu) * g(\nu)$ stands for a convolution of $f(\nu)$ and $g(\nu)$, and $a \vee b, a \wedge b$ refer to the the least upper bound $\max(a, b)$ and greatest lower bound $\min(a, b)$ of a, b respectively. The symbol $\perp\!\!\!\perp$ indicates that the random variables are independent. K_0, K_1 are the modified Bessel functions of the second kind with parameters 0,1 respectively (Chapter 10 of (Olver et al., 2010)).

Assumptions.

Assumption 2.1. The latent variable and the mixing weights $(z; \pi)$ are independent of the regression parameters θ, namely $(z; \pi) \perp\!\!\!\perp \theta$.

Assumption 2.2. The additive noise ε is independent of the covariate random variable, latent variable, the regression parameters, and the mixing weights, that is $\varepsilon \perp\!\!\!\perp (x, z; \theta, \pi)$.

Assumption 2.3. The covariate random variable x is independent of the latent variable, the regression parameters, and the mixing weights, namely $x \perp\!\!\!\perp (z; \theta, \pi)$.

Assumption 2.4. Both the covariate x and the noise ε are Gaussians, $x \sim \mathcal{N}(0, I_d), \varepsilon \sim \mathcal{N}(0, \sigma^2)$, where I_d is a d by d identity matrix.

We leverage the assumption of the Gaussianity of the covariate, which is standard in this line of work (see, e.g., Assumption 1 in (Ghosh & Kannan, 2020) and Section 2.1 in (Kwon et al., 2021)). The above standard assumptions are necessary to derive the forthcoming results.

EM Updates. Balakrishnan et al. (2017) considered the following population EM update for 2MLR given the balanced mixing weights $\pi = \pi^* = \{\frac{1}{2}, \frac{1}{2}\}$, where $\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} := \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} \mathbb{E}_{y|x\sim\pi^*(1)\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)+\pi^*(2)\mathcal{N}(-\langle x,\theta^*\rangle,\sigma^2)}$.

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

To extend the EM update for both balanced and unbalanced mixing weights, we introduce

$$\nu := \frac{\log \pi(1) - \log \pi(2)}{2}, \quad \pi = \{\pi(1), \pi(2)\}, \tag{2}$$

that is $\tanh(\nu) = \pi(1) - \pi(2)$. Thus, the population EM update rule for regression parameters θ becomes

$$M(\theta, \nu) := \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \tanh\left(\frac{y\langle x, \theta\rangle}{\sigma^2} + \nu\right) yx, \tag{3}$$

while the corresponding EM update rule for $\tanh(\nu)$ is (see the derivations in the supplementary, Appendix B)

$$N(\theta, \nu) := \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \tanh\left(\frac{y\langle x, \theta\rangle}{\sigma^2} + \nu\right). \tag{4}$$

Subsequently, the finite-sample EM update rules are

$$\begin{aligned} M_n(\theta, \nu) &= \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1} \\ &\quad \left(\frac{1}{n}\sum_{i=1}^n \tanh\left(\frac{y_i\langle x_i, \theta\rangle}{\sigma^2} + \nu\right) y_i x_i\right) \\ N_n(\theta, \nu) &= \frac{1}{n}\sum_{i=1}^n \tanh\left(\frac{y_i\langle x_i, \theta\rangle}{\sigma^2} + \nu\right). \end{aligned} \tag{5}$$

For the ease of theoretical analysis, we will use the easy EM method as discussed in Section 5 with the following update

$$M_n^{\text{easy}}(\theta, \nu) = \frac{1}{n}\sum_{i=1}^n \tanh\left(\frac{y_i\langle x_i, \theta\rangle}{\sigma^2} + \nu\right) y_i x_i. \tag{6}$$

Trajectory-Relevant Quantities. Our trajectory-based analysis further utilizes certain angles described next. We denote the cosine of the angle between the estimate for parameters θ and the true value θ^* by $\rho := \frac{\langle\theta,\theta^*\rangle}{\|\theta\|\cdot\|\theta^*\|}$. Furthermore, $\varphi := \frac{\pi}{2} - \arccos|\rho| \in [0, \frac{\pi}{2}), \phi := 2\arccos|\rho| \in (0, \pi]$ are defined accordingly. As apparent from these definitions, we refer to ρ as the sub-optimality cosine and φ and ϕ as *sub-optimality angles*, respectively.

Further, let $\hat{e}_1 := \frac{\theta^*}{\|\theta^*\|}, \vec{e}_1 := \frac{\theta}{\|\theta\|}$ be the direction unit vectors of θ^*, θ, and $\hat{e}_2 := \frac{\theta - \hat{e}_1\hat{e}_1^\top\theta}{\|\theta - \hat{e}_1\hat{e}_1^\top\theta\|}, \vec{e}_2 := \frac{\theta^* - \vec{e}_1\vec{e}_1^\top\theta^*}{\|\theta^* - \vec{e}_1\vec{e}_1^\top\theta^*\|}$ be the unit vectors on the plane $\text{span}\{\theta^*, \theta\}$ which are perpendicular to $\hat{e}_1, \vec{e}_1$ respectively. The superscript t stands for the t-th EM iteration. For instance, θ^t, π^t denote the t-th iteration for regression parameters and mixing weights. These vectors will again simplify the ensuing trajectory-based discussion (see Fig. 1 for visualization).

3. Population EM Updates

In this section we derive closed-form expressions for the update of population EM we introduced in Eq. (3), (4). Let $U(\theta, \nu) := \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \log\cosh(\frac{y\langle x,\theta\rangle}{\sigma^2} + \nu)$. Then, Eq. (3), (4) can be written as the following relations.

$$M(\theta, \nu) = \sigma^2\nabla_\theta U(\theta, \nu), \quad N(\theta, \nu) = \nabla_\nu U(\theta, \nu). \tag{7}$$

Therefore, we need to derive a closed-form expression for $U(\theta, \nu)$ which appears in the population update of both θ and ν. In the supplementary materials (Appendix A), we provide the explicit expression of this expectation by introducing K_0, the modified Bessel function of the second kind with parameter 0 (Chapter 10 of (Olver et al., 2010)). Subsequently, we derive the closed-form expressions for EM update rules at the population level in Theorem 3.1 below (see Appendices A, C).

Theorem 3.1. (EM Updates across All SNR) Let $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$, $\bar{\theta} := \frac{\theta}{\sigma}$, $\bar{\theta}^* := \frac{\theta^*}{\sigma}$, then the EM update rules for $\theta, \tanh(\nu)$ at Population level are

$$M(\theta, \nu) = \left[-\frac{\sigma}{\pi} \cdot \frac{\|\bar{\theta}^*\|^2}{\|\bar{\theta}\|^2} \cdot \frac{\sqrt{1-\rho^2}}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{3}{2}}} \cosh^{-1}(\nu^*) \right]$$
$$\left\{ \tanh(\nu) * \nu \left[\alpha(\nu) \left(\frac{\frac{\bar{\theta}}{\|\bar{\theta}\|}}{\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2} + \hat{e}_2 \right) + \beta(\nu)\vec{e}_2 \right] \right\},$$
$$N(\theta, \nu) = \frac{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{1}{2}}}{\pi \|\bar{\theta}\| \cosh(\nu^*)} \int_{\mathbb{R}} \mathrm{d}\nu' \tanh(\nu - \nu')$$
$$K_0 \left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2} \cdot \left| \frac{\nu'}{\|\bar{\theta}\|} \right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} \right) \cosh \left(\frac{\rho \|\bar{\theta}^*\| \left(\frac{\nu'}{\|\bar{\theta}\|} \right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^* \right)$$

where these coefficients are defined as

$$\alpha(\nu) := \cosh \left(\frac{\rho \|\bar{\theta}^*\| \left(\frac{\nu}{\|\bar{\theta}\|} \right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^* \right)$$
$$K_0 \left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2} \cdot \left| \frac{\nu}{\|\bar{\theta}\|} \right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} \right),$$
$$\beta(\nu) := \mathrm{sgn}(\nu) \frac{\sqrt{1+\|\bar{\theta}^*\|^2}}{\|\bar{\theta}^*\|} \sinh \left(\frac{\rho \|\bar{\theta}^*\| \left(\frac{\nu}{\|\bar{\theta}\|} \right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^* \right)$$
$$K_1 \left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2} \cdot \left| \frac{\nu}{\|\bar{\theta}\|} \right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} \right).$$

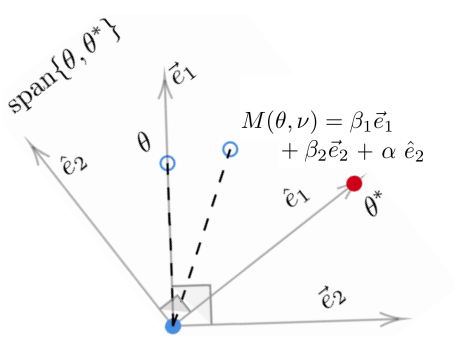


Figure 1: The EM update $M(\theta, \nu)$ for regression parameters lies on span$\{\theta, \theta^*\}$.

Note that in Theorem 3.1 we have $\{\bar{\theta}, \hat{e}_2, \vec{e}_2\} \in$ span$\{\theta, \theta^*\}$, which in turn implies the population EM update for the regression parameters satisfies $M(\theta, \nu) \in$ span$\{\theta, \theta^*\}$. We visualize these vectors and their relations in Fig. 1.

Special cases. For the special case of no separation SNR $:= \frac{\|\theta^*\|}{\sigma} \to 0$ in Theorem 3.1, the population EM update rules can be simplified (see Appendix C, proof of Corollary 3.2).

Corollary 3.2. (EM Updates for No Separation Case) For the special case of no separation of parameters, namely SNR $:= \frac{\|\theta^*\|}{\sigma} \to 0$, the EM update rules for $\theta^t, \tanh(\nu^t)$ at Population level are $\bar{\theta}^t = \frac{\bar{\theta}^0}{\|\bar{\theta}^0\|} \cdot \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\|\bar{\theta}^{t-1}\| x - \nu^{t-1}) x K_0(|x|) \mathrm{d}x$ and $\tanh(\nu^t) = \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\nu^{t-1} - \|\bar{\theta}^{t-1}\| x) K_0(|x|) \mathrm{d}x$, where $\bar{\theta}^t := \frac{\theta^t}{\sigma}$ and $\bar{\theta}^0 := \frac{\theta^0}{\sigma}$.

The second special case, which will be our main focus moving forward, arises by letting SNR $:= \frac{\|\theta^*\|}{\sigma} \to \infty$ in Theorem 3.1. In doing so, we obtain the EM update rules at the population level in the noiseless setting described below (see Appendix C, proof of Corollary 3.3).

Corollary 3.3. (EM Updates in Noiseless Setting) In the noiseless setting, namely SNR $:= \frac{\|\theta^*\|}{\sigma} \to \infty$, the EM update rules for $\theta^t, \tanh(\nu^t)$ at the Population level are

$$\frac{\theta^t}{\|\theta^*\|} = \frac{2}{\pi} \left[\mathrm{sgn}(\rho^{t-1}) \varphi^{t-1} \frac{\theta^*}{\|\theta^*\|} + \cos \varphi^{t-1} \frac{\theta^{t-1}}{\|\theta^{t-1}\|} \right]$$
$$\tanh(\nu^t) = \mathrm{sgn}(\rho^{t-1}) \left(\frac{2}{\pi} \varphi^{t-1} \right) \cdot \tanh(\nu^*),$$

where $\rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \|\theta^*\|}$, $\varphi^{t-1} := \frac{\pi}{2} - \arccos |\rho^{t-1}|$.

4. Population Level Analysis

In this section, we focus on the properties of EM update rules (Corollary 3.3) in the noiseless setting at the population level and present one of our main theoretical results on the convergence rate and the estimation of error of the population EM through a trajectory-based analysis. In particular, we show that the error of mixing weights (in ℓ_1 norm) is proportional to the angle between the estimated regression parameters and the true parameters and establish a consistent quadratic convergence rate, which, interestingly, is independent of mixing weights.

Theorem 4.1. (Population Level Convergence) If the initial sup-optimality cosine $\rho^0 := \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|} \neq 0$, then with the number of total iterations at most $T = \mathcal{O}\left(\log \frac{1}{|\rho^0|} \vee \log \log \frac{1}{\varepsilon} \right)$, the error of EM update at the population level is bounded by $\frac{\|\theta^{T+1} - \mathrm{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} < \varepsilon$, and $\|\pi^{T+1} - \bar{\pi}^*\|_1 = \mathcal{O}(\sqrt{\varepsilon}) \cdot \|\frac{1}{2} - \pi^*\|_1$, where $\bar{\pi}^* := \frac{1}{2} - \mathrm{sgn}(\rho^0)(\frac{1}{2} - \pi^*)$.

Proof Sketch of Theorem 4.1. To estimate the errors of mixing weights and regression parameters, we need to study the trajectory of the regression parameters θ^t. Recall, Corollary 3.3 demonstrated that for the special case of the noiseless setting, the EM update for θ^t depends on the angle φ^t. Additionally, upon recalling $|\tanh(\nu)| = \|\frac{1}{2} - \pi^*\|_1$, we directly obtain the following Corollary 4.2 for the error of mixing weights. Regarding the regression

parameters, we derive the recurrence relation in Proposition 4.3 and further show the cycloid trajectory in Proposition 4.4. Hence, the error of regression parameters is only determined by the sup-optimality angle φ^t. Therefore, instead of directly analyzing the error of regression parameters, we could characterize its behavior by studying the evolution of the *sub-optimality angle* φ^t.

Corollary 4.2. (Error of Mixing Weights π^t) In the noiseless setting, the error of mixing weights for EM updates at the population level is

$$\|\pi^t - \bar{\pi}^*\|_1 = \left|1 - \frac{2}{\pi}\varphi^{t-1}\right| \cdot \left\|\frac{1}{2} - \pi^*\right\|_1 \tag{8}$$

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

Thus, at the population level, we investigate the evolution of the *sub-optimality angle* φ^t to characterize the convergence behavior of EM iterations. In Proposition 4.5 we show the linear convergence rate of $\tan\varphi^t$ when φ^t is relatively small and demonstrate the quadratic convergence rate when φ^t is large enough. Hence, we can divide the EM updates into two stages: In the first stage, it takes $T' = \mathcal{O}(\frac{1}{\tan\varphi^0}) = \mathcal{O}(\frac{1}{\sin\varphi^0}) = \mathcal{O}(\frac{1}{|\rho^0|})$ iterations to ensure $\tan\varphi^{T'}$ is large enough. Subsequently, in the second stage, it takes the other $T'' = \mathcal{O}(\log\log\frac{1}{\varepsilon})$ to guarantee $\phi^T = \mathcal{O}(\sqrt{\varepsilon})$. Hence, $\frac{\|\theta^{T+1} - \operatorname{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} = \frac{1}{\pi}\sqrt{(\phi^T - \sin\phi^T)^2 + (1-\cos\phi^T)^2} \le \frac{[\phi^T]^2}{2\pi} = \mathcal{O}(\varepsilon)$.

The next proposition derives a recurrence equation for the update of the sub-optimality angle φ.

Proposition 4.3. *(Recurrence Relation) Assume the initial sub-optimality cosine satisfies* $\rho^0 := \frac{\langle\theta^0,\theta^*\rangle}{\|\theta^0\|\cdot\|\theta^*\|} \neq \pm 1$, *or equivalently* $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in [0, \pi/2)$. *Then the recurrence relation for EM updates at population level characterized by the sub-optimality angle is*

$$\tan\varphi^t = \tan\varphi^{t-1} + \varphi^{t-1}(\tan^2\varphi^{t-1} + 1). \tag{9}$$

In the following Proposition, we show that the trajectory of iterations $\theta^t, \forall t \in \mathbb{N}_+$ is on the cycloid (Harris & Stöcker, 1998) on the plane $\operatorname{span}\{\theta^0, \theta^*\}$.

Proposition 4.4. *(Cycloid Trajectory) If* $\rho^0 := \frac{\langle\theta^0,\theta^*\rangle}{\|\theta^0\|\cdot\|\theta^*\|} \neq \pm 1$, *namely* $\phi^0 := 2\arccos|\rho^0| \in (0, \pi]$. *Then the coordinates* $\mathrm{x}^t, \mathrm{y}^t$ *of normalized vector* $\frac{\theta^t}{\|\theta^*\|} = \mathrm{x}^t\hat{e}_1 + \mathrm{y}^t\hat{e}_2^t = \mathrm{x}^t\hat{e}_1 + \mathrm{y}^t\hat{e}_2^0, \forall t \in \mathbb{N}_+$ *for EM updates at the Population level can be parameterized with the angle* $\phi^{t-1} := 2\arccos|\rho^{t-1}| \in (0, \pi]$ *as follows,*

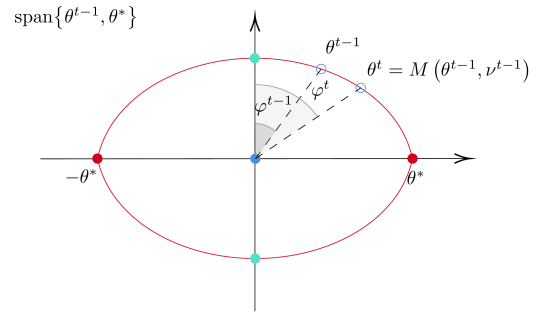


Figure 2: The cycloid trajectory for the EM update $M(\theta, \nu)$ of regression parameters θ. The figure further shows the two global solutions (red dots), the unstable solution (blue dot), and the two saddle points (green dots). As long as the initial suboptimality angle is sufficiently large, φ^t and in turn θ^t super-linearly converge to $\frac{\pi}{2}$ and θ^*.

where $\rho^{t-1} := \frac{\langle\theta^{t-1},\theta^*\rangle}{\|\theta^{t-1}\|\cdot\|\theta^*\|}$.

$$\begin{aligned} 1 - \operatorname{sgn}(\rho^0)\mathrm{x}^t &= \frac{1}{\pi}[\phi^{t-1} - \sin\phi^{t-1}] \\ \mathrm{y}^t &= \frac{1}{\pi}[1 - \cos\phi^{t-1}] \end{aligned} \tag{10}$$

Hence, the trajectory of iterations $\theta^t, \forall t \in \mathbb{N}_+$ *is on the cycloid with a parameter* $\frac{\|\theta^*\|}{\pi}$, *on the plane* $\operatorname{span}\{\theta^0, \theta^*\}$ *(see Fig. 2).*

Proposition 4.5. *(Quadratic Convergence Rate) If* $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in (0, \frac{\pi}{2})$, *then the EM updates at population level satisfies*

$$\tan\varphi^t \ge \frac{1+\sqrt{5}}{2} \cdot \tan\varphi^{t-1}. \tag{11}$$

Particularly, if $\varphi^{t-1} := \frac{\pi}{2} - \arccos|\rho^{t-1}| \in [\arctan 1.5, \frac{\pi}{2})$, *then the EM updates at the Population level satisfies*

$$\frac{\pi}{2}\left(\tan\varphi^t - \frac{\pi}{4}\right) \ge \left\{\frac{\pi}{2}\left(\tan\varphi^{t-1} - \frac{\pi}{4}\right)\right\}^2. \tag{12}$$

Proposition 4.5 states that as long as the initial sub-optimality angle is nonzero, the angle increases linearly which implies convergence to global optima. Note that if $\varphi^0 = 0$, then the update may converge to one of the saddle points (see Fig. 2 for a visualization). Let $\phi = 2(\frac{\pi}{2} - \varphi)$ be small enough, then $\tan(\varphi) = \cot(\frac{\phi}{2}) \approx \frac{2}{\phi}$, and by Proposition 4.5, $\frac{\pi}{2}\tan(\varphi^t) \approx [\frac{\pi}{2}\tan(\varphi^{t-1})]^2$, so we show that $\phi^t/\pi \approx [\phi^{t-1}/\pi]^2$, demonstrating quadratic convergence in the angle ϕ (see Appendix E, proof of Proposition 5.5). Further, note that at the population level, Kwon et al. (2021) (Page 5, Lemma 1) established the quadratic convergence when $C\sqrt{\log\|\bar{\theta}^*\|}/\|\bar{\theta}^*\| \le \frac{\|\theta - \theta^*\|}{\|\theta^*\|} \le 1/10$ for high SNR. Proposition 4.5 then extends the region of quadratic convergence from 1/10 to $\frac{1}{\pi}\sqrt{(\phi - \sin\phi)^2 + (1-\cos\phi)^2}_{\phi = 2(\frac{\pi}{2} - \arctan 1.5)} \approx 0.22$ in the noiseless setting.

5. Finite-sample Level Analysis

In this section, we give a finite-sample analysis by coupling the population EM with the corresponding finite-sample EM in the noiseless setting.

Theorem 5.1. (Convergence at Finite-sample Level) In the noiseless setting, suppose any initial mixing weights π^0 and any initial regression parameters $\theta^0 \in \mathbb{R}^d$ ensuring that $\varphi^0 \geq \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$. If we run finite-sample Easy EM for at most $T_1 = \mathcal{O}\left(\log\frac{n}{\log\frac{1}{\delta}}\right)$ iterations followed by the finite-sample standard EM for at most $T' = \mathcal{O}\left(\log\frac{n}{d} \wedge \log\frac{n}{\log\frac{1}{\delta}}\right)$ iterations with all the same $n = \Omega\left(d \vee \log\frac{1}{\delta}\right)$ samples, then we have

$$\frac{\|\theta^{T+1} - \mathrm{sgn}(\rho^{T+1})\theta^*\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

$$\|\pi^{T+1} - \bar{\pi}^*\|_1 = \left\|\frac{1}{2} - \pi^*\right\|_1 \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) + c(\pi^*) \cdot \mathcal{O}\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

with probability at least $1 - T\delta$, where $T := T_1 + T', \varphi^0 := \frac{\pi}{2} - \arccos\left|\frac{\langle\theta^0,\theta^*\rangle}{\|\theta^0\|\cdot\|\theta^*\|}\right|, \rho^{T+1} := \frac{\langle\theta^{T+1},\theta^*\rangle}{\|\theta^{T+1}\|\cdot\|\theta^*\|}, \bar{\pi}^* := \frac{1}{2} - \mathrm{sgn}(\rho^0)(\frac{1}{2} - \pi^*)$, and the coefficient $c(\pi^*) = \mathcal{O}(1)$, especially $c(\pi^*) = 0$ when $\pi^* = \{1, 0\}$ or $\{0, 1\}$.

We note that Ghosh & Kannan (2020) showed the quadratic convergence at the finite-sample level, only when the error of regression parameters $\lesssim \min\{\pi^*(1), \pi^*(2)\}\|\theta_1^* - \theta_2^*\|$ for a variant of EM in the noiseless setting. Theorem 5.1 on the other hand removes the restriction of mixing weights and still obtains the quadratic convergence rate. Our results for convergence rate and statistical error hold when $\frac{d}{n}$ is not greater than some constant, which is less than 1 (see Appendix E, proof of Proposition 5.5). However, if $\frac{d}{n} > 1$, recovering regression parameters $\theta \in \mathbb{R}^d$ from a reduced number n of measurements invalidates the EM update rules for regression parameters at the finite-sample level (see Eq. (5)), as $\mathrm{rank}(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top) \leq n < d$. Additional restrictions on θ, such as assuming some components of θ are zeros (Barik & Honorio, 2022), are necessary.

Sketch Proof of Theorem 5.1 The first step is to upper-bound the statistical error of the finite-sample EM. In the following Proposition 5.2, Proposition 5.3, we give the bounds for the projected statistical error $\mathcal{O}(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n})$ and the total statistical error $\mathcal{O}(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}})$.

Proposition 5.2. *(Projected Statistical Error) In the noiseless setting, the projection on 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 span(P_{θ,θ^*}) = span$\{\theta, \theta^*\}$.

Proposition 5.3. *(Statistical Error) In the noiseless setting, the statistical error of θ for EM updates at the Finite-sample level satisfies*

$$\frac{\|M_n(\theta,\nu) - M(\theta,\nu)\|_2}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

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.

Kwon et al. (2021) (Appendix E, page 17, Lemma 11) and Balakrishnan et al. (2017) upper-bounded $\mathcal{O}(\sqrt{\frac{d}{n}}\log\frac{n}{\delta})$ for the statistical error, while our finer analysis presents a tighter bound $\mathcal{O}(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}) \lesssim \mathcal{O}(\sqrt{\frac{d}{n}}\log\frac{n}{\delta})$. The difference arises from the techniques used for bounding the statistical error in the ℓ_2 norm. We obtain our additive log factor by leveraging the rotational invariance of Gaussians, rewriting the ℓ_2 norm of the error as the geometric mean of two Chi-square distributions (see Appendix E, proof of Proposition 5.3). The multiplicative log factor is achieved using the standard symmetrization technique and the Ledoux-Talagrand contraction argument (see Appendix E, proof of Lemma 11, page 17 of (Kwon et al., 2021)).

Proposition 5.4 shows in the first step when using the initialization for θ^0, π^0 the angle φ^t is larger than the projected statistical error, if we run at most $\mathcal{O}(\log\frac{1}{\delta})$ iterations of Easy EM with at most $\Theta(\frac{n}{\log\frac{1}{\delta}})$ fresh samples per iteration.

Proposition 5.4. *(Initialization with Easy EM) In the noiseless setting, suppose we run the sample-splitting finite-sample Easy EM with $n' := \Theta\left(\frac{n}{\log\frac{1}{\delta}} \wedge \left[\frac{n}{\log\frac{1}{\delta}}\right]^2\right)$ fresh samples for each iteration, then after at most $T_0 = \mathcal{O}\left(\log\frac{1}{\delta}\right)$ iterations, it satisfies $\varphi^{T_0} \geq \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$ with probability at least $1 - \delta$.*

This result allows us to divide the analysis of the convergence rate into three stages. In the first stage, we show that after at most $T_1 = \mathcal{O}(\log\frac{\text{statistical error}}{\text{projected statistical error}}) =$

$\mathcal{O}(\log \frac{n}{\log \frac{1}{\delta}})$ iterations of easy EM, we can ensure the angle φ^t is larger than the statistical error. Then, in the second stage, by using the linear convergence rate for φ^t established in Proposition 4.5, we show that the angle φ^t would be larger than $\arctan(1.5)$ after at most $T_2 = \mathcal{O}(\log \frac{1}{\text{statistical error}}) = \mathcal{O}(\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}})$ iterations of standard EM. In the third stage, with the quadratic convergence rate for $\varphi^t \geq \arctan(1.5)$ in Proposition 4.5, we further show the distance between θ^t and θ^* will decrease with the quadratic speed until the distance reaches the statistical error (using Proposition 5.5). Hence, with at most $T_3 = \mathcal{O}(\log\log \frac{1}{\text{statistical error}}) = \mathcal{O}(\log[\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}])$ iterations of standard EM, the error of regression parameters reaches the statistical error.

Proposition 5.5. *(Convergence of Angle) In the noiseless setting, suppose* $\varphi^0 \geq \Theta\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}} \vee \frac{\log \frac{1}{\delta}}{n}\right)$*, run finite-sample Easy EM for* $T_1 = \mathcal{O}\left(\log \frac{n}{\log \frac{1}{\delta}}\right)$ *iterations followed by the finite-sample standard EM for at most* $T' = \mathcal{O}\left(\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}}\right)$ *iterations with all the same* $n = \Omega\left(d \vee \log \frac{1}{\delta}\right)$ *samples, then it satisfies*

$$\varphi^T \geq \frac{\pi}{2} - \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log \frac{1}{\delta}}{n} \vee \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right), \tag{13}$$

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* $\varphi^T := \frac{\pi}{2} - \arccos\left|\frac{\langle \theta^T, \theta^* \rangle}{\|\theta^T\| \cdot \|\theta^*\|}\right|$.

Upon taking all these three stages into account, the total number of iterations until convergence with the initialization (i.e., $\varphi_0 \geq$ the projected statistical error) is $T = T_1 + T_2 + T_3 = \mathcal{O}(\log \frac{n}{\log \frac{1}{\delta}})$. In particular, with a good initialization (i.e., $\varphi_0 \geq$ the statistical error), the total number of iterations is $T' = T_2 + T_3 = \mathcal{O}(\log \frac{n}{d} \wedge \log \frac{n}{\log \frac{1}{\delta}})$.

For the error of the mixing weights, we first establish the upper bound for the error between the population EM update and the finite-sample EM update for mixing weights $\|N_n(\theta, \nu) - N(\theta, \nu)\|_1 = c(\pi^*) \cdot \mathcal{O}(\sqrt{\frac{\log \frac{1}{\delta}}{n}})$. We establish this result by estimating the Chernoff bound (the full proof of Theorem 5.1 is described in the supplementary materials, Appendix E). The final error is obtained by summing up the error stemming from the population EM update (Corollary 4.2) and the error between the population and finite-sample EM updates.

6. Experiments

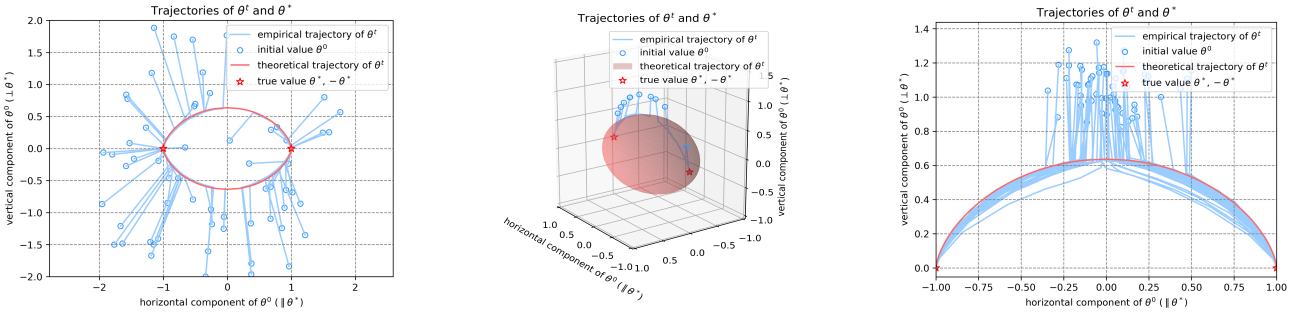
In this section of empirical experiments, we validate the theoretical findings established in the preceding sections. From a normal distribution $\mathcal{N}(0, I_d)$, we sample 5,000 independent and identically distributed (i.i.d.) d-dimensional covariates, denoted as $\{x_i\}_{i=1}^n$. The true parameters θ^* are randomly chosen from a d-dimensional unit sphere. We subsequently manually/randomly set the true mixture weights π^* for two components, utilizing them to generate latent variable samples $\{z_i\}_{i=1}^n$ from a categorical distribution $\mathcal{CAT}(\pi^*)$. Following this, we introduce Gaussian noise to the linear regression determined by these latent variables, yielding output samples $\{y_i\}_{i=1}^n$. In all experiments, we utilize the entire dataset for EM updates at every iteration. Each point on the plots of Fig. 4 is an average taken from 50 trials with different initial values for EM updates. The code for numerical experiments is available at `https://github.com/dassein/cycloid_em_mlr`.

Cycloid Trajectory of EM Iterations. At the population level, we show that the output of the t-th iteration lies on the cycloid of the spanning space span$\{\theta^{t-1}, \theta^*\}$ in the noiseless setting. In the corresponding experiments, we choose the signal-to-noise ratio SNR$= 10^8$ and consider different values of d (2, 3, and 50). In Fig. 3, all the iterations are near the theoretical cycloid. Thus, our experimental results validate our theoretical analysis in Proposition 4.4.

Quadratic Convergence for 2 Mixtures. We show the super-linear convergence of $\frac{\pi}{2}(\tan \varphi^t - \frac{\pi}{4})$ in Fig. 4a under high SNR regimes. We specified the dimension (d=50) and considered different high SNR values (SNR=$10^6, 10^7, 10^8$). We uniformly choose the initial values for the parameters and the mixing weights from a unit sphere and the interval $[0, 1]$, respectively. All the points of 4 EM iterations in Fig. 4a are the average of 50 trials with different initial values. The slope of the plot indicates the convergence rate exponent. Notably, the slopes of lines at different SNR values consistently hover around or slightly exceed 2. That aligns with our theoretical result of quadratic convergence rate in Proposition 4.5.

Error of Mixing Weights and Angle. In the noiseless setting, we prove that the error of mixing weights $\|\pi^t - \bar{\pi}^*\|_1$ at the Population level is proportional to $\frac{\pi}{2} - \varphi^{t-1} = \arccos\left|\frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \cdot \|\theta^*\|}\right|$ in Corollary 4.2. The angle $\frac{\pi}{2} - \varphi^{t-1}$ is determined by the output of EM updates at the $(t-1)$-th iteration and the true value of parameters θ^*. We demonstrate the linear correlation between the error of mixing weights and the angle in Fig. 4b. For the setting of experiments, we specify the dimension $d = 50$, and consider different high SNR values (SNR=$10^6, 10^7, 10^8$), respectively. We note that the error in the mixing weights during the t-th iteration is precisely quantified by the angle $\frac{\pi}{2} - \varphi^{t-1}$ at the preceding iteration. Hence, our experimental results validate Corollary 4.2.

Comparison with Different Mixing Weights. In the noiseless setting, we establish in Corollary 3.3 and Propo-

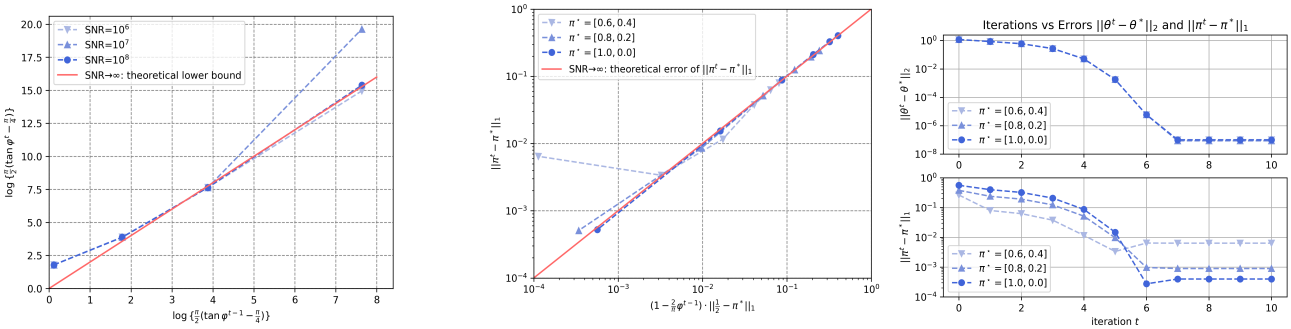


(a) $d = 2$, $\theta^* = [1, 0]$, $\pi^* = [0.7, 0.3]$, trajectories of θ^t for 60 trials with θ^0 and π^0 uniformly sampled from $[-2, 2]^2$ and $[0, 1]$, respectively.

(b) $d = 3$, trajectories of θ^t for 10 trials, with θ^*, θ^0 sampled from three-dimensional unit sphere, π^*, π^0 drawn uniformly from $[0, 1]$.

(c) $d = 50$, trajectories of θ^t are displayed across 60 trials, with θ^*, θ^0 sampled from $\mathcal{N}(0, I_d)$, π^*, π^0 drawn uniformly from $[0, 1]$.

Figure 3: Cycloid trajectory of EM iterations θ^t– we perform 100 iterations of Finite-sample EM with SNR=10^8, varying dimensions ($d = 2, 3, 50$).



(a) Quadratic convergence of $\frac{\pi}{2}(\tan\varphi^t - \frac{\pi}{4})$ with $\pi^*, \pi_0 \sim [0, 1]$.

(b) Correlation between $\|\pi^t - \bar{\pi}^*\|_1$ and $\frac{\pi}{2} - \varphi^{t-1} = \arccos\left|\frac{\langle\theta^{t-1}, \theta^*\rangle}{\|\theta^{t-1}\|\|\theta^*\|}\right|$.

(c) Estimation error of EM v.s. iteration for various mixing weights.

Figure 4: Left and Middle: Quadratic convergence and correlation are shown with θ^*, θ^0 from $d = 50$ unit sphere, s.t. $\varphi^0 = \arctan(1.5)$ in Panel (a), $\varphi^0 = 0.3$ in Panel (b). Right: The errors of regression parameters and mixing weights for ten EM iterations, with $d = 50$, $\varphi^0 = 0.3$, SNR=10^8 and different true mixing weights $\pi^* = \{0.6, 0.4\}, \{0.8, 0.2\}, \{1, 0\}$.

sition 4.3 that the EM update for regression parameters θ^t is independent of the true mixing weights π^*. The first subplot of Fig. 4c demonstrates that, at high SNR (10^8), the error in regression parameters (measured in ℓ_2 norm) remains nearly constant for varying true mixing weights $\pi^* = \{0.6, 0.4\}$, $\{0.8, 0.2\}$, and $\{1, 0\}$, thus affirming our theoretical analysis.

In Theorem 5.1, we prove that the final error (in ℓ_1 norm) in mixing weights depends on the error in regression parameters and true mixing weights. Specifically, when the error in regression parameters is relatively small, the closer the true mixing weights are to $\{1, 0\}$ or $\{0, 1\}$, the smaller the final mixing weight error. To validate our theoretical analysis in Theorem 5.1 concerning the statistical errors in regression parameters θ (measured in ℓ_2 norm) and mixing weights π (measured in ℓ_1 norm), we experiment with various true mixing weights $\pi^* = \{0.6, 0.4\}$, $\{0.8, 0.2\}$, and $\{1, 0\}$. The second subplot of Fig. 4c illustrates the relationship between the errors and the true mixing weights π^*, further supports our theoretical analysis. In our experimental setup, θ^*, θ^0 are sampled from 50-dimensional unit sphere, with $\varphi^0 = 0.3$.

7. Conclusion

We derived closed-form expressions for the EM updates in the 2MLR problem. Notably, in the noiseless setting we first showed and then analyzed the cycloid trajectory of EM updates. Additionally, we demonstrated the quadratic convergence rate for regression parameters, which is independent of mixing weights. We emphasized that errors in mixing weights primarily arise from the angle formed between true and estimated regression parameters. Finally, we conducted a detailed analysis of the statistical errors in the estimation of regression parameters and mixing weights. We investigate the special case of the noiseless setting, namely when SNR tends to infinity. Could we propose a more refined analysis using the recurrence relations outlined in Corollary 3.2 for weakly separated scenarios? These questions could guide our potential future endeavors.

Acknowledgements

We are grateful to the ICML area chair and anonymous reviewers for their insightful input that improved this paper.

Impact Statement

There are several potential applications of our theory in Mixed Linear Regression: **Analysis of Over-specified Model**: Corollary 3.2 enables a thorough analysis of no separation case as SNR$\rightarrow$ 0 to obtain a fundamental understanding of EM with over-specified mixture models (Dwivedi et al., 2020a;b). **Extension to Finite High/Low SNR Cases**: Leveraging the insights from Theorem 3.1, we can conduct asymptotic expansions of integrals (Wong, 2001; Bleistein & Handelsman, 1986), enabling the extension of results from limit cases (SNR $\rightarrow \infty, 0$) to practical, finite high and low SNR scenarios, exploring the transition from low SNR to high SNR regimes. **Generalization to Multiple Components**: We could expand our analysis from a mixture of two components to scenarios involving multiple components, albeit requiring strong separation of regression parameters compared to the noise level and good initialization (Kwon & Caramanis, 2020).

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1): 77–120, 2017. doi: 10.1214/16-AOS1435. 2, 3, 6

Barik, A. and Honorio, J. Sparse mixed linear regression with guarantees: Taming an intractable problem with invex relaxation. In *International Conference on Machine Learning*, pp. 1627–1646. PMLR, 2022. 6

Beale, E. M. and Little, R. J. Missing values in multivariate analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 37(1):129–145, 1975. 1

Bleistein, N. and Handelsman, R. A. *Asymptotic expansions of integrals*. Courier Corporation, 1986. 9

Chandrasekher, K. A., Pananjady, A., and Thrampoulidis, C. Sharp global convergence guarantees for iterative nonconvex optimization: A Gaussian process perspective. *arXiv preprint arXiv:2109.09859*, 2021. 2

Chen, Y. and Xi, X. Likelihood landscape and local minima structures of Gaussian mixture models. *arXiv preprint arXiv:2009.13040*, 2020. 2

Chen, Y., Yi, X., and Caramanis, C. Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory*, 64(3):1738–1766, 2018. doi: 10.1109/TIT.2017.2773474. 2

Dasgupta, S. and Schulman, L. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(7):203–226, 2007. 2

Daskalakis, C., Tzamos, C., and Zampetakis, M. Ten steps of EM suffice for mixtures of two Gaussians. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 704–710. PMLR, 07–10 Jul 2017. 2

De Veaux, R. D. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989. 1

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 1

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M., Jordan, M., and Yu, B. Sharp analysis of expectation-maximization for weakly identifiable models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1866–1876. PMLR, 2020a. 2, 9

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., and Yu, B. Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48 (6):3161 – 3182, 2020b. doi: 10.1214/19-AOS1924. 2, 9

Ghosh, A. and Kannan, R. Alternating minimization converges super-linearly for mixed linear regression. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1093–1103. PMLR, 26–28 Aug 2020. 2, 3, 6

Harris, J. W. and Stöcker, H. *Handbook of mathematics and computational science*. Springer Science & Business Media, 1998. 5, 32

Hong, J., Qian, W., Chen, Y., and Zhang, Y. A geometric approach to k-means. *arXiv preprint arXiv:2201.04822*, 2022. 2

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and Jordan, M. I. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. *Advances in neural information processing systems*, 29, 2016. 2

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6 (2):181–214, 1994. 1

Jordan, M. I. and Xu, L. Convergence results for the EM approach to mixtures of experts architectures. *Neural networks*, 8(9):1409–1431, 1995. 1

Katsevich, A. and Bandeira, A. S. Likelihood maximization and moment matching in low SNR Gaussian mixture models. *Communications on Pure and Applied Mathematics*, 76(4):788–842, 2023. 2

Klusowski, J. M. and Brinda, W. Statistical guarantees for estimating the centers of a two-component Gaussian mixture by EM. *arXiv preprint arXiv:1608.02280*, 2016. 2

Klusowski, J. M., Yang, D., and Brinda, W. D. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019. doi: 10.1109/TIT.2019.2891628. 2

Kong, W., Somani, R., Kakade, S., and Oh, S. Robust meta-learning for mixed linear regression with small batches. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4683–4696. Curran Associates, Inc., 2020a. 1

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. Meta-learning for mixed linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5394–5404. PMLR, 13–18 Jul 2020b. 1

Kwon, J. and Caramanis, C. EM converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1727–1736. PMLR, 2020. 2, 9

Kwon, J., Qian, W., Caramanis, C., Chen, Y., and Davis, D. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pp. 2055–2110. PMLR, 2019. 2

Kwon, J., Ho, N., and Caramanis, C. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1405–1413. PMLR, 2021. 2, 3, 5, 6

Kwon, J. et al. *Statistical learning with latent variables: mixture models and reinforcement learning*. PhD thesis, The University of Texas at Austin, 2022. 2

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000. 38

Li, Y. and Liang, Y. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pp. 1125–1144. PMLR, 2018. 1

Ndaoud, M. Sharp optimal recovery in the two-component Gaussian mixture model. *arXiv preprint arXiv:1812.08078*, 2018. 2

Olver, F. W., Lozier, D. W., Boisvert, R., and Clark, C. W. *NIST handbook of mathematical functions*. Cambridge university press, 2010. 2, 3, 14, 15

Qian, W. *Local Minima in Mixture Problems and Their Algorithmic Implications*. Cornell University, 2020. 2

Qian, W., Zhang, Y., and Chen, Y. Global convergence of least squares EM for demixing two log-concave densities. *Advances in Neural Information Processing Systems*, 32, 2019. 2

Qian, W., Zhang, Y., and Chen, Y. Structures of spurious local minima in k-means. *IEEE Transactions on Information Theory*, 68(1):395–422, 2022. doi: 10.1109/TIT.2021.3122465. 2

Ren, T., Cui, F., Sanghavi, S., and Ho, N. Beyond EM algorithm on over-specified two-component location-scale Gaussian mixtures. *arXiv preprint arXiv:2205.11078*, 2022. 2

Ross, N. Fundamentals of Stein's method. *Probability Surveys*, 8(none):210 – 293, 2011. doi: 10.1214/11-PS182. 42

Shen, Y. and Sanghavi, S. Iterative least trimmed squares for mixed linear regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

Thrampoulidis, C., Oymak, S., and Hassibi, B. The Gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014. 2

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015. 2

Tipping, M. E. and Bishop, C. M. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11 (2):443–482, 1999. 1

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019. 36, 39, 40

Wedel, M. and DeSarbo, W. S. A mixture likelihood approach for generalized linear models. *Journal of classification*, 12:21–55, 1995. 1

Wong, R. *Asymptotic approximations of integrals*. SIAM, 2001. 9

Wu, C. J. On the convergence properties of the EM algorithm. *The Annals of statistics*, pp. 95–103, 1983. 1

Wu, Y. and Zhou, H. H. Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4(3), 2021. 2

Xu, J., Hsu, D. J., and Maleki, A. Global analysis of expectation maximization for mixtures of two Gaussians. *Advances in Neural Information Processing Systems*, 29, 2016. 2

Xu, Y. and Zeevi, A. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020. 2

Yi, X., Caramanis, C., and Sanghavi, S. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pp. 613–621. PMLR, 2014. 2

Yi, X., Caramanis, C., and Sanghavi, S. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016. 2

Zhang, Y., Qu, Q., and Wright, J. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020. 2

Zhao, R., Li, Y., and Sun, Y. Statistical convergence of the EM algorithm on Gaussian mixture models. *Electronic Journal of Statistics*, 14:632–660, 2020. 2

Supplementary Materials: Unveiling the Cycloid Trajectory of EM Iterations in Mixed Linear Regression

Appendix

Table of Contents

A Lemmas: Integrals, Convolutions, Expectations **13**

B Derivations for EM Update Rules **20**

C Proof for Results of Population EM Updates **26**

D Proof for Results at the Population Level **31**

E Proof for Results at the Finite-sample Level **36**

We organize the Appendix as follows:

- In Section A, we prepare some lemmas for integrals, convolutions related to Bessel functions, expectations for Gaussiansused in proofs, etc.
- In Section B, we derive EM update rules at the population level and the finite-sample level.
- In Section C, we provide the proof for the explicit closed-form expressions with Bessel functions for Population EM Updates.
- In Section D, we give the full proof for the results at the population level in the noiseless setting.
- In Section E, we give the full proof for the results at the finite-sample level in the noiseless setting.

A. Lemmas: Integrals, Convolutions, Expectations

A.1. Relations between θ^*, θ and unit vecotors $\vec{e}_1, \vec{e}_2, \hat{e}_1, \hat{e}_2$

The following lemma shows the relations between θ^*, θ and unit vecotors $\vec{e}_1, \vec{e}_2, \hat{e}_1, \hat{e}_2$.

Lemma A.1. For $\hat{e}_1 := \frac{\theta^*}{\|\theta^*\|}$, $\hat{e}_2 := \frac{\theta - \hat{e}_1 \hat{e}_1^\top \theta}{\|\theta - \hat{e}_1 \hat{e}_1^\top \theta\|} = \frac{\frac{\theta}{\|\theta\|} - \rho \frac{\theta^*}{\|\theta^*\|}}{\sqrt{1-\rho^2}}$, and $\vec{e}_1 = \frac{\theta}{\|\theta\|}$, $\vec{e}_2 = \frac{\theta^* - \vec{e}_1 \vec{e}_1^\top \theta^*}{\|\theta^* - \vec{e}_1 \vec{e}_1^\top \theta^*\|}$, define $\rho := \frac{\langle \bar{\theta}^*, \bar{\theta} \rangle}{\|\bar{\theta}\| \cdot \|\bar{\theta}^*\|} = \frac{\langle \theta^*, \theta \rangle}{\|\theta\| \cdot \|\theta^*\|}$, then

$$\begin{aligned} \vec{e}_2 + \rho \hat{e}_2 &= \sqrt{1-\rho^2} \frac{\theta^*}{\|\theta^*\|} \\ \hat{e}_2 + \rho \vec{e}_2 &= \sqrt{1-\rho^2} \frac{\theta}{\|\theta\|} \end{aligned}$$

Proof. Let $\bar{\theta}^* := \frac{\theta^*}{\sigma}$ and $\bar{\theta} := \frac{\theta}{\sigma}$

Define $\vec{e}_1 = \frac{\theta}{\|\theta\|}$ and $\vec{e}_2 = \frac{\theta^* - \vec{e}_1 \vec{e}_1^\top \theta^*}{\|\theta^* - \vec{e}_1 \vec{e}_1^\top \theta^*\|}$ thus $\langle \vec{e}_1, \vec{e}_2 \rangle = 0, \|\vec{e}_1\| = \|\vec{e}_2\| = 1, \operatorname{span}\{\theta, \theta^*\} = \operatorname{span}\{\bar{\theta}^*, \bar{\theta}\} = \operatorname{span}\{\vec{e}_1, \vec{e}_2\}$

Let $\rho := \frac{\langle \bar{\theta}^*, \bar{\theta} \rangle}{\|\bar{\theta}\| \cdot \|\bar{\theta}^*\|} = \frac{\langle \theta^*, \theta \rangle}{\|\theta\| \cdot \|\theta^*\|}$, then $\left| \bar{\theta}^* - \frac{\langle \bar{\theta}^*, \bar{\theta} \rangle}{\|\bar{\theta}\|^2} \bar{\theta} \right| = \sqrt{1-\rho^2} \|\bar{\theta}^*\|$

If we define $\hat{e}_1 := \frac{\theta^*}{\|\theta^*\|}$, and $\hat{e}_2 := \frac{\theta - \hat{e}_1 \hat{e}_1^\top \theta}{\|\theta - \hat{e}_1 \hat{e}_1^\top \theta\|} = \frac{\frac{\theta}{\|\theta\|} - \rho \frac{\theta^*}{\|\theta^*\|}}{\sqrt{1-\rho^2}}$ and $\langle \hat{e}_1, \hat{e}_2 \rangle = 0, \|\hat{e}_1\| = \|\hat{e}_2\| = 1$

$$\begin{pmatrix} \theta \\ \theta^* \end{pmatrix} = \begin{pmatrix} \|\theta\| & \\ & \|\theta^*\| \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} \vec{e}_1 \\ \vec{e}_2 \end{pmatrix} = \begin{pmatrix} \|\theta\| & \\ & \|\theta^*\| \end{pmatrix} \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix}$$

Therefore

$$\begin{aligned} \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} &= \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} \vec{e}_1 \\ \vec{e}_2 \end{pmatrix} = \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ \frac{}{\sqrt{1-\rho^2}} & -\rho \end{pmatrix} \begin{pmatrix} \vec{e}_1 \\ \vec{e}_2 \end{pmatrix} \\ \begin{pmatrix} \vec{e}_1 \\ \vec{e}_2 \end{pmatrix} &= \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ \frac{}{\sqrt{1-\rho^2}} & -\rho \end{pmatrix} \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \vec{e}_1 - \frac{\rho\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2}{1+(1-\rho^2)\|\bar{\theta}^*\|^2} \vec{e}_2 &= \begin{pmatrix} 1 \\ -\frac{\rho\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2}{1+(1-\rho^2)\|\bar{\theta}^*\|^2} \end{pmatrix}^\top \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ \frac{}{\sqrt{1-\rho^2}} & -\rho \end{pmatrix} \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} \\ &= \frac{1}{\frac{1}{\|\bar{\theta}^*\|^2} + (1-\rho^2)} \begin{pmatrix} \frac{1}{\|\bar{\theta}^*\|^2} + (1-\rho^2) \\ -\rho\sqrt{1-\rho^2} \end{pmatrix}^\top \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ \frac{}{\sqrt{1-\rho^2}} & -\rho \end{pmatrix} \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} \\ &= \frac{\begin{pmatrix} \rho\frac{1}{\|\bar{\theta}^*\|^2} & \sqrt{1-\rho^2}\left(\frac{1}{\|\bar{\theta}^*\|^2} + 1\right) \end{pmatrix}}{\frac{1}{\|\bar{\theta}^*\|^2} + (1-\rho^2)} \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} \\ &= \frac{\frac{1}{\|\bar{\theta}^*\|^2}}{\frac{1}{\|\bar{\theta}^*\|^2} + (1-\rho^2)} \cdot \frac{\bar{\theta}}{\|\bar{\theta}\|} + \frac{\sqrt{1-\rho^2}}{\frac{1}{\|\bar{\theta}^*\|^2} + (1-\rho^2)} \cdot \hat{e}_2 \end{aligned}$$

$$\frac{1}{\sqrt{1-\rho^2}} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \frac{\theta}{\|\theta\|} \\ \frac{\theta^*}{\|\theta^*\|} \end{pmatrix} = \begin{pmatrix} \hat{e}_2 \\ \vec{e}_2 \end{pmatrix}$$

Therefore, we can show that

$$\vec{e}_2 + \rho \hat{e}_2 = \sqrt{1-\rho^2} \frac{\theta^*}{\|\theta^*\|}, \quad \hat{e}_2 + \rho \vec{e}_2 = \sqrt{1-\rho^2} \frac{\theta}{\|\theta\|}$$

□

A.2. Integrals and Expectation with Gaussian

A.2.1. INTEGRALS WITH GAUSSIAN

Lemma A.2. For $\forall a > 0$, then

$$(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left[-\frac{at^2+2bt}{2}\right] \mathrm{d}t = \frac{\exp\left[\frac{b^2}{2a}\right]}{\sqrt{a}}$$

$$(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left[-\frac{at^2+2bt}{2}\right] t\mathrm{d}t = -\frac{b}{a^{\frac{3}{2}}} \exp\left[\frac{b^2}{2a}\right]$$

A.2.2. EXPECTATIONS WITH GAUSSIAN

Lemma A.3. Let $\bar{\theta} := \frac{\theta}{\sigma}, \bar{\theta^*} := \frac{\theta^*}{\sigma}$, and $\nu \sim \mathcal{N}(-\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2)$ or $\nu \sim \mathcal{N}(\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2)$, then the expectations for the density functions are

$$\begin{aligned}
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathcal{N}(-\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2) \\
= \; & (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta^*}\|^2)^{-\frac{1}{2}} \exp\left[-\frac{\rho\|\bar{\theta^*}\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta^*}\|^2]}\right] \cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta^*}\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta^*}\|^2]}\right) \\
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathcal{N}(\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2) \\
= \; & (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta^*}\|^2)^{-\frac{1}{2}} \exp\left[+\frac{\rho\|\bar{\theta^*}\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta^*}\|^2]}\right] \cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta^*}\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta^*}\|^2]}\right)
\end{aligned}$$

Proof. We define $\rho := \frac{\langle\bar{\theta},\bar{\theta^*}\rangle}{\|\bar{\theta}\|\cdot\|\bar{\theta^*}\|}$, and $\lambda_1 := \langle\bar{\theta}, \vec{e}_1\rangle, \lambda_2 := \langle\bar{\theta}, \vec{e}_2\rangle$.

Then $\langle x, \bar{\theta}\rangle = \lambda_1\|\bar{\theta}\|, \langle x, \bar{\theta^*}\rangle = \left\langle \lambda_1\vec{e}_1 + \lambda_2\vec{e}_2, \rho\|\bar{\theta^*}\|\vec{e}_1 + \sqrt{1-\rho^2}\|\bar{\theta^*}\|\vec{e}_2\right\rangle = \lambda_1\rho\|\bar{\theta^*}\| + \lambda_2\sqrt{1-\rho^2}\|\bar{\theta^*}\|$.

For the evaluation of the first expectation, we let

$v \leftarrow 0, z \leftarrow \frac{\sqrt{1+\|\bar{\theta^*}\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta^*}\|^2]}$ in (Olver et al., 2010) Chapter 10 (Schläfli's Integral of $K_v(z)$).

$a \leftarrow 1+(1-\rho^2)\|\bar{\theta^*}\|^2$ and $b \leftarrow \frac{\nu\sqrt{1-\rho^2}\|\bar{\theta^*}\|}{\lambda_1\|\bar{\theta}\|} + \lambda_1\rho\sqrt{1-\rho^2}\|\bar{\theta^*}\|^2 = \sqrt{1-\rho^2}\|\bar{\theta^*}\|\left(\frac{\nu}{\lambda_1\|\bar{\theta}\|} + \lambda_1\rho\|\bar{\theta^*}\|\right)$ and $t \leftarrow \lambda_2$ in Lemma A.2.

For the evaluation of the second expectation, note that the following relation holds for these two density functions.

$$\mathcal{N}(\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2) = \left[\mathcal{N}(-\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2)\right]_{\bar{\theta}\to-\bar{\theta}}$$

Since $\bar{\theta} \to -\bar{\theta}$ implies $\rho \to -\rho$, we can obtain such a relation to derive the closed-form expression in this Lemma.

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathcal{N}(\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2) = \left[\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathcal{N}(-\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2)\right]_{\rho\to-\rho}$$

□

Lemma A.4. Let $\bar{\theta} := \frac{\theta}{\sigma}, \bar{\theta^*} := \frac{\theta^*}{\sigma}$, and $\lambda_1 := \langle x, \vec{e}_1\rangle, \lambda_2 := \langle x, \vec{e}_2\rangle$.
$\nu \sim \mathcal{N}(-\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2)$ or $\nu \sim \mathcal{N}(\langle x, \bar{\theta^*}\rangle\langle x, \bar{\theta}\rangle, \langle x, \bar{\theta}\rangle^2)$,

then the expectations for the the products of $\frac{\lambda_2}{\lambda_1}$ and density functions are

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_2}{\lambda_1}\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\\
=\;& -(\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{3}{2}}\sqrt{1-\rho^2}\|\bar{\theta}^*\|\exp\left[-\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right]\\
\cdot\;& \left[\mathrm{sgn}(\nu)[1+\|\bar{\theta}^*\|^2]^{\frac{1}{2}}K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)+\rho\|\bar{\theta}^*\|K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\right]\\
&\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_2}{\lambda_1}\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\\
=\;& (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{3}{2}}\sqrt{1-\rho^2}\|\bar{\theta}^*\|\exp\left[+\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right]\\
\cdot\;& \left[\mathrm{sgn}(\nu)[1+\|\bar{\theta}^*\|^2]^{\frac{1}{2}}K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)-\rho\|\bar{\theta}^*\|K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\right]
\end{aligned}$$

Proof. For the first expectation, let $v\leftarrow\{0,1\}, z\leftarrow\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}$ in (Olver et al., 2010) Chapter 10 (Schläfli's Integral of $K_v(z)$); $a\leftarrow 1+(1-\rho^2)\|\bar{\theta}^*\|^2$ and $b\leftarrow\frac{\nu\sqrt{1-\rho^2}\|\bar{\theta}^*\|}{\lambda_1\|\bar{\theta}\|}+\lambda_1\rho\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2=\sqrt{1-\rho^2}\|\bar{\theta}^*\|\left(\frac{\nu}{\lambda_1\|\bar{\theta}\|}+\lambda_1\rho\|\bar{\theta}^*\|\right)$ and $t\leftarrow\lambda_2$ in Lemma A.2.

For the second expectation, note that $\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)=\left[\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\right]_{\bar{\theta}\to-\bar{\theta}}$ holds for these two density functions. Since $\bar{\theta}\to-\bar{\theta}$ implies $\rho\to-\rho$, $\sqrt{1-\rho^2}\to-\sqrt{1-\rho^2}$, we can obtain such a relation.

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_2}{\lambda_1}\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)=\left[\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_2}{\lambda_1}\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\right]_{\rho\to-\rho,\sqrt{1-\rho^2}\to-\sqrt{1-\rho^2}}$$

□

Lemma A.5. Let $\bar{\theta}:=\frac{\theta}{\sigma},\bar{\theta^*}:=\frac{\theta^*}{\sigma}$, and $\lambda_1:=\langle x,\vec{e}_1\rangle,\lambda_2:=\langle x,\vec{e}_2\rangle$.
$\nu\sim\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)$ or $\nu\sim\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)$,

then the expectations for the the products of $\frac{x}{\langle x,\bar{\theta}\rangle}$ and density functions are

$$
\begin{aligned}
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x}{\langle x,\bar{\theta}\rangle}\right]\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) \\
= \; & \frac{(\pi\|\bar{\theta}\|^2)^{-1}}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{1}{2}}}\exp\left[-\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right]\cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\left[\vec{e}_1-\frac{\rho\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2}{1+(1-\rho^2)\|\bar{\theta}^*\|^2}\vec{e}_2\right] \\
- \; & \mathrm{sgn}(\nu)\frac{(\pi\|\bar{\theta}\|^2)^{-1}\sqrt{1-\rho^2}\|\bar{\theta}^*\|[1+\|\bar{\theta}^*\|^2]^{\frac{1}{2}}}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{3}{2}}}\exp\left[-\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right]K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\vec{e}_2 \\
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x}{\langle x,\bar{\theta}\rangle}\right]\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) \\
= \; & (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{3}{2}}\sqrt{1-\rho^2}\|\bar{\theta}^*\|\exp\left[+\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right] \\
= \; & \frac{(\pi\|\bar{\theta}\|^2)^{-1}}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{1}{2}}}\exp\left[+\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right]\cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\left[\vec{e}_1-\frac{\rho\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2}{1+(1-\rho^2)\|\bar{\theta}^*\|^2}\vec{e}_2\right] \\
+ \; & \mathrm{sgn}(\nu)\frac{(\pi\|\bar{\theta}\|^2)^{-1}\sqrt{1-\rho^2}\|\bar{\theta}^*\|[1+\|\bar{\theta}^*\|^2]^{\frac{1}{2}}}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{3}{2}}}\exp\left[+\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right]K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\vec{e}_2
\end{aligned}
$$

Proof. We express $x=(\lambda_1\vec{e}_1+\lambda_2\vec{e}_2)+\tilde{x}$ and r.v. $\tilde{x}\perp\!\!\!\perp\lambda_1,\lambda_2$ and $\mathbb{E}_{\tilde{x}}[\tilde{x}]=0$. Hence, for any function $\varphi(\lambda_1,\lambda_2)$, we have $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\tilde{x}\cdot\varphi(\lambda_1,\lambda_2)]=\mathbb{E}_{\lambda_1,\lambda_2\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\varphi(\lambda_1,\lambda_2)[\mathbb{E}_{\tilde{x}}[\tilde{x}\mid\lambda_1,\lambda_2]]=\mathbb{E}_{\lambda_1,\lambda_2\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\varphi(\lambda_1,\lambda_2)[\mathbb{E}_{\tilde{x}}[\tilde{x}]]=0$

Note that $\langle x,\bar{\theta}\rangle,\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2),\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)$ are functions of λ_1,λ_2.

Therefore, $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{\tilde{x}}{\langle x,\bar{\theta}\rangle}\right]\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)=0,\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{\tilde{x}}{\langle x,\bar{\theta}\rangle}\right]\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)=0.$

Subsequently, we can decompose thees expectations into two terms.

$$
\begin{aligned}
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x}{\langle x,\bar{\theta}\rangle}\right]\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)=\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_1\vec{e}_1+\lambda_2\vec{e}_2}{\lambda_1\|\bar{\theta}\|}\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) \\
= \; & \frac{\vec{e}_1}{\|\bar{\theta}\|}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)+\frac{\vec{e}_2}{\|\bar{\theta}\|}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_2}{\lambda_1}\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) \\
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x}{\langle x,\bar{\theta}\rangle}\right]\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) \\
= \; & \frac{\vec{e}_1}{\|\bar{\theta}\|}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)+\frac{\vec{e}_2}{\|\bar{\theta}\|}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\frac{\lambda_2}{\lambda_1}\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)
\end{aligned}
$$

Then, with the previous two Lemmas, we derive the closed-from expressions in this Lemma. □

A.2.3. EXPECTATIONS FOR 2MLR

Lemma A.6. For the 2MLR at the population level, $s:=(x,y)$

and for any function $\psi(y),\forall y\in\mathbb{R}$, the operator $\mathcal{F}_{-y}$ are defined by $\mathcal{F}_{-y}[\psi(y)]=\psi(-y)$;

then $\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}=\mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y|x\sim\pi^*(1)\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)+\pi^*(2)\mathcal{N}(-\langle x,\theta^*\rangle,\sigma^2)}$ satisfies

$$\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}=\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}[\pi^*(1)+\pi^*(2)\mathcal{F}_{-y}]$$

Proof. For any $\psi(y)$, it can be verified by letting $y \leftarrow -y$ in the second term, and note that $\mathcal{F}_{-y}[\psi(y)] = \psi(-y)$

$$\begin{aligned}
& \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\psi(y)\mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y|x\sim\pi^*(1)\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)+\pi^*(2)\mathcal{N}(-\langle x,\theta^*\rangle,\sigma^2)}\psi(y) \\
= \quad & \pi^*(1)\mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}\psi(y) + \pi^*(2)\mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y\sim\mathcal{N}(-\langle x,\theta^*\rangle,\sigma^2)}\psi(y) \\
& \pi^*(1)\mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}\psi(y) + \pi^*(2)\mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}\psi(-y) \\
= \quad & \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}[\pi^*(1) + \pi^*(2)\mathcal{F}_{-y}]\cdot\psi(y)
\end{aligned}$$

□

Lemma A.7. Let $\bar{\theta} := \frac{\theta}{\sigma}, \bar{\theta^*} := \frac{\theta^*}{\sigma}$ and $\rho := \frac{\langle\theta,\theta^*\rangle}{\|\theta\|\cdot\|\theta^*\|}, \nu^* := \frac{\log\pi^*(1)-\log\pi^*(2)}{2}$, then

$$\begin{aligned}
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\pi^*(1)\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) + \pi^*(2)\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\right] \\
= \quad & (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{1}{2}} \\
& K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\cosh^{-1}(\nu^*)\cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right)
\end{aligned}$$

Proof. With the previous Lemma and $\pi^*(1) = \frac{1+\tanh(\nu^*)}{2}, \pi^*(2) = \frac{1-\tanh(\nu^*)}{2}$, we obtain the follows.

$$\begin{aligned}
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\pi^*(1)\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) + \pi^*(2)\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\right] \\
= \quad & (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{1}{2}}K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right) \\
\cdot \quad & \left[\cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right) - \tanh(\nu^*)\sinh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\right]
\end{aligned}$$

Note that $\cosh(\alpha) - \tanh(\beta)\sinh(\alpha) = \frac{\cosh(\alpha-\beta)}{\cosh(\beta)}$, let $\alpha \leftarrow \frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}, \beta \leftarrow \nu^*$, we further prove this Lemma. □

Lemma A.8. Let $\bar{\theta} := \frac{\theta}{\sigma}, \bar{\theta^*} := \frac{\theta^*}{\sigma}$ and $\rho := \frac{\langle\theta,\theta^*\rangle}{\|\theta\|\cdot\|\theta^*\|}, \nu^* := \frac{\log\pi^*(1)-\log\pi^*(2)}{2}$, then

$$\begin{aligned}
& \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x}{\langle x,\bar{\theta}\rangle}\right]\left[\pi^*(1)\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2) + \pi^*(2)\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\right] \\
= \quad & \frac{1}{\pi}\cdot\frac{\|\bar{\theta}^*\|^2}{\|\bar{\theta}\|^2}\cdot\frac{\sqrt{1-\rho^2}\cosh^{-1}(\nu^*)}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{3}{2}}} \\
\Bigg\{ \quad & \cosh\left[\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right]\cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\left[\frac{1}{\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2}\cdot\frac{\bar{\theta}}{\|\bar{\theta}\|} + \hat{e}_2\right] \\
+ \quad & \mathrm{sgn}(\nu)\frac{\sqrt{1+\|\bar{\theta}^*\|^2}}{\|\bar{\theta}^*\|}\sinh\left[\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right]\cdot K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\vec{e}_2\Bigg\}
\end{aligned}$$

Proof. Note that $\frac{1+\tanh(\beta)}{2}\exp(-\alpha) + \frac{1-\tanh(\beta)}{2}\exp(\alpha) = \cosh(\alpha) - \tanh(\beta)\sinh(\alpha) = \frac{\cosh(\alpha-\beta)}{\cosh(\beta)}$,

and $-\frac{1+\tanh(\beta)}{2}\exp(-\alpha) + \frac{1-\tanh(\beta)}{2}\exp(\alpha) = \sinh(\alpha) - \tanh(\beta)\cosh(\alpha) = \frac{\sinh(\alpha-\beta)}{\cosh(\beta)}$,

let $\alpha \leftarrow \frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}, \beta \leftarrow \nu^*$, and use previous Lemma, we give the following. With Lemma for relations of unit vectors, we obtain $\hat{e}_2 + \rho\vec{e}_2 = \sqrt{1-\rho^2}\frac{\theta}{\|\theta\|}$, and by definition $\vec{e}_1 = \frac{\theta}{\|\theta\|}$ and $\vec{e}_2 = -\hat{e}_2 + \sqrt{1-\rho^2}\frac{\theta}{\|\theta\|}$.

$$\vec{e}_1 - \frac{\rho\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2}{1+(1-\rho^2)\|\bar{\theta}^*\|^2}\vec{e}_2 = \left(1+(1-\rho^2)\|\bar{\theta}^*\|^2\right)^{-1}\left[\frac{\theta}{\|\theta\|} + \sqrt{1-\rho^2}\|\bar{\theta}^*\|^2\hat{e}_2\right]$$

Hence, this Lemma is proved by rearranging the terms and using the above relation for vectors. □

Lemma A.9. For the 2MLR at the population level, let $\rho := \frac{\langle\theta,\theta^*\rangle}{\|\theta\|\cdot\|\theta^*\|}, \nu^* := \frac{\log\pi^*(1)-\log\pi^*(2)}{2}$,

$$\begin{aligned}
& \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\log\cosh\left(\frac{y\langle x,\theta\rangle}{\sigma^2}+\nu\right) \\
= \; & \log\cosh(\nu) * (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{1}{2}} \\
& K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\cosh^{-1}(\nu^*)\cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}-\nu^*\right)
\end{aligned}$$

Proof. Note that $\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}[\pi^*(1)+\pi^*(2)\mathcal{F}_{-y}]$ in a previous Lemma; and for convolution $(f*g)(\nu) = \int_{-\infty}^{+\infty} f(\nu-\nu')g(\nu')\mathrm{d}\nu'$, we can exchange expectation and convolution. Let $\nu' := -\frac{y\langle x,\theta\rangle}{\sigma^2}$, then $\nu' \sim \mathcal{N}\left(-\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2}, \frac{\langle x,\theta\rangle^2}{\sigma^2}\right)$; and $\nu'' := \frac{y\langle x,\theta\rangle}{\sigma^2}$, then $\nu'' \sim \mathcal{N}\left(\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2}, \frac{\langle x,\theta\rangle^2}{\sigma^2}\right)$.

$$\begin{aligned}
& \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\log\cosh\left(\frac{y\langle x,\theta\rangle}{\sigma^2}+\nu\right) = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}[\pi^*(1)+\pi^*(2)\mathcal{F}_{-y}]\log\cosh\left(\frac{y\langle x,\theta\rangle}{\sigma^2}+\nu\right) \\
= \; & \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}\left[\pi^*(1)\log\cosh\left(\frac{y\langle x,\theta\rangle}{\sigma^2}+\nu\right)+\pi^*(2)\log\cosh\left(-\frac{y\langle x,\theta\rangle}{\sigma^2}+\nu\right)\right] \\
= \; & \pi^*(1)\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{\nu'\sim\mathcal{N}\left(-\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2},\frac{\langle x,\theta\rangle^2}{\sigma^2}\right)}\log\cosh(\nu-\nu') \\
+ \; & \pi^*(2)\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{\nu''\sim\mathcal{N}\left(\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2},\frac{\langle x,\theta\rangle^2}{\sigma^2}\right)}\log\cosh(\nu-\nu'') \\
= \; & \log\cosh(\nu) * \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\pi^*(1)\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)+\pi^*(2)\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle,\langle x,\bar{\theta}\rangle^2)\right]
\end{aligned}$$

Then with closed-form expression in the previous Lemma, we further prove this Lemma. □

Lemma A.10. For the 2MLR at the population level, let $\rho := \frac{\langle\theta,\theta^*\rangle}{\|\theta\|\cdot\|\theta^*\|}, \nu^* := \frac{\log\pi^*(1)-\log\pi^*(2)}{2}$,

$$\begin{aligned}
& \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\tanh\left(\frac{y\langle x,\theta\rangle}{\sigma^2}+\nu\right) \\
= \; & \tanh(\nu) * (\pi\|\bar{\theta}\|)^{-1}(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{1}{2}} \\
& K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\cosh^{-1}(\nu^*)\cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}-\nu^*\right) \\
= \; & \frac{\left(1+\left(1-\rho^2\right)\|\bar{\theta}^*\|^2\right)^{-\frac{1}{2}}}{\pi\|\bar{\theta}\|\cosh\left(\nu^*\right)}\int_{\mathbb{R}}\tanh\left(\nu-\nu'\right)K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu'}{\|\bar{\theta}\|}\right|}{1+(1-\rho^2)\|\bar{\theta}^*\|^2}\right)\cosh\left(\frac{\rho\left\|\bar{\theta}^*\right\|\left(\frac{\nu'}{\|\bar{\theta}\|}\right)}{1+(1-\rho^2)\|\bar{\theta}^*\|^2}-\nu^*\right)\mathrm{d}\nu'
\end{aligned}$$

Proof. We follow the same steps in the proof of the previous Lemma, but substitute $\log\cosh$ with $\tanh$. □

Lemma A.11. For the 2MLR at the population level, let $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}, \nu^* := \frac{\log \pi^*(1) - \log \pi^*(2)}{2}$,

$$
\begin{aligned}
& \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \tanh\left(\frac{y\langle x,\theta\rangle}{\sigma^2} + \nu\right) yx \\
= \quad & -\sigma \frac{1}{\pi} \cdot \frac{\|\bar{\theta}^*\|^2}{\|\bar{\theta}\|^2} \cdot \frac{\sqrt{1-\rho^2}\cosh^{-1}(\nu^*)}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{\frac{3}{2}}} \tanh(\nu) * \nu \\
\Bigg\{ \quad & \cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right) \cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2} \cdot \left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right) \left[\frac{1}{\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2} \cdot \frac{\bar{\theta}}{\|\bar{\theta}\|} + \hat{e}_2\right] \\
& + \operatorname{sgn}(\nu) \frac{\sqrt{1+\|\bar{\theta}^*\|^2}}{\|\bar{\theta}^*\|} \sinh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right) \cdot K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2} \cdot \left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right) \vec{e}_2 \Bigg\}
\end{aligned}
$$

Proof. Note that $\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}[\pi^*(1) + \pi^*(2)\mathcal{F}_{-y}]$ in a previous Lemma; and for convolution $(f * g)(\nu) = \int_{-\infty}^{+\infty} f(\nu - \nu')g(\nu')\mathrm{d}\nu'$, we can exchange expectation and convolution. Let $\nu' := -\frac{y\langle x,\theta\rangle}{\sigma^2}$, then $\nu' \sim \mathcal{N}\left(-\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2}, \frac{\langle x,\theta\rangle^2}{\sigma^2}\right)$; and $\nu'' := \frac{y\langle x,\theta\rangle}{\sigma^2}$, then $\nu'' \sim \mathcal{N}\left(\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2}, \frac{\langle x,\theta\rangle^2}{\sigma^2}\right)$.

$$
\begin{aligned}
& \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \tanh\left(\frac{y\langle x,\theta\rangle}{\sigma^2} + \nu\right) xy \\
= \quad & \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} x \cdot \mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)}[\pi^*(1) + \pi^*(2)\mathcal{F}_{-y}] \tanh\left(\frac{y\langle x,\theta\rangle}{\sigma^2} + \nu\right) y \\
= \quad & \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} x \mathbb{E}_{y\sim\mathcal{N}(\langle x,\theta^*\rangle,\sigma^2)} \left[\pi^*(1)\tanh\left(\frac{y\langle x,\theta\rangle}{\sigma^2} + \nu\right) y - \pi^*(2)\tanh\left(-\frac{y\langle x,\theta\rangle}{\sigma^2} + \nu\right) y\right] \\
= \quad & -\pi^*(1)\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x\sigma^2}{\langle x,\theta\rangle}\right] \mathbb{E}_{\nu'\sim\mathcal{N}\left(-\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2}, \frac{\langle x,\theta\rangle^2}{\sigma^2}\right)} \tanh(\nu - \nu')\nu' \\
& -\pi^*(2)\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x\sigma^2}{\langle x,\theta\rangle}\right] \mathbb{E}_{\nu''\sim\mathcal{N}\left(\frac{\langle x,\theta^*\rangle\langle x,\theta\rangle}{\sigma^2}, \frac{\langle x,\theta\rangle^2}{\sigma^2}\right)} \tanh(\nu - \nu'')\nu'' \\
= \quad & -\sigma\tanh(\nu) * \nu \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\frac{x}{\langle x,\bar{\theta}\rangle}\right] (\pi^*(1)\mathcal{N}(-\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle, \langle x,\bar{\theta}\rangle^2) + \pi^*(2)\mathcal{N}(\langle x,\bar{\theta}^*\rangle\langle x,\bar{\theta}\rangle, \langle x,\bar{\theta}\rangle^2))
\end{aligned}
$$

Then, with the previous Lemma of expectaions for 2MLR, we further prove the closed-form expression. □

B. Derivations for EM Update Rules

Lemma B.1. The negative expected log-likelihood $f(\theta,\pi) := -\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}[\log p(s \mid \theta,\pi)]$ 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.

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

where $\operatorname{KL}_s, \mathcal{H}_s$ are KL divengence and Shannon's entropy wrt. $s = (x,y)$;

$\operatorname{KL}_z, \mathcal{H}_z, \operatorname{softmax}_z$ are KL divengence, Shannon's entropy and softmax wrt. $z \in \mathcal{Z} = [M]$;

$\{q_s(z) \mid s \in \mathcal{X}\times\mathcal{Y} = \mathbb{R}^d \times \mathbb{R}\}$ is a family of distributions wrt. $z \in \mathcal{Z} = [M]$, namely $\sum_{z\in\mathcal{Z}} q_s(z) = 1$.

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

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

□

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

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

that is $g^t(\theta,\pi) \geq f(\theta,\pi)$, and $g^t(\theta,\pi)\mid_{(\theta,\pi)=(\theta^{t-1},\pi^{t-1})} = f(\theta,\pi)\mid_{(\theta,\pi)=(\theta^{t-1},\pi^{t-1})}$,

$\nabla_\theta g^t(\theta,\pi)\mid_{(\theta,\pi)=(\theta^{t-1},\pi^{t-1})} = \nabla_\theta f(\theta,\pi)\mid_{(\theta,\pi)=(\theta^{t-1},\pi^{t-1})}$;

where $\operatorname{KL}_s, \mathcal{H}_s$ are KL divengence and Shannon's entropy wrt. $s = (x,y)$;

$\operatorname{KL}_z, \mathcal{H}_z, \operatorname{softmax}_z$ are KL divengence, Shannon's entropy and softmax wrt. $z \in \mathcal{Z} = [M]$

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

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

Hence, the gradients of r^t wrt. (θ,π) at $(t-1)$-th iteration (θ^{t-1},π^{t-1}) are all 0 by the chain rule. □

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

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

where $\mathrm{KL}_s, \mathcal{H}_s$ are KL divengence and Shannon's entropy wrt. $s=(x,y)$;

$\mathrm{KL}_z, \mathcal{H}_z, \mathrm{softmax}_z$ are KL divengence, Shannon's entropy and softmax wrt. $z \in \mathcal{Z} = [M]$;

$q_s(z) \leftarrow p\left(z \mid s;\theta^{t-1},\pi^{t-1}\right)$ and $\pi^t = \{\pi(z)\}_{z\in\mathcal{Z}}, \pi^t(z) := \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} q_s(z)$.

Proof.

- $(z;\pi) \perp\!\!\!\perp \theta$ 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 $p(z \mid s;\theta,\pi) = \frac{p(s,z|\theta,\pi)}{p(s|\theta,\pi)} = \frac{\pi(z)\cdot p(y|x,z;\theta)}{\sum_{z'\in\mathcal{Z}}\pi(z')\cdot p(y|x,z';\theta)} = \mathrm{softmax}_z(\log\pi(z) + \log p(y \mid x,z;\theta))$

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. □

Lemma B.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 $\varepsilon \sim \mathcal{N}(0,\sigma^2)$

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

where $c = -\frac{1}{2}\log\left(2\pi\sigma^2\right)$ and $q_s(z) \leftarrow p\left(z \mid s;\theta^{t-1},\pi^{t-1}\right) = \mathrm{softmax}_z\left(-\frac{\|y-\langle\theta_z^{t-1},x\rangle\|^2}{2\sigma^2} + \log\pi^{t-1}(z)\right)$, 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, $p(y, x, z; \theta) = p(\varepsilon, x, z; \theta) \left|\frac{\partial \varepsilon}{\partial y}\right| = p(\varepsilon) \cdot p(x, z; \theta)$

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

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

Subsequently, note that $g^t(\theta, \pi) = Q(\theta, \pi \mid \theta^{t-1}, \pi^{t-1}) - \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\mathcal{H}_z\left[q_s(z)\right]$, 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. □

Theorem B.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 $f(\theta, \pi) := -\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}[\log p(s \mid \theta, \pi)]$ and the surrogate function $g^t(\theta, \pi)$ can be expressed as follows.

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

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

Proof. Note that $\text{sigmoid}(2t) + \text{sigmoid}(-2t) = 1$ and $\text{sigmoid}(2t) - \text{sigmoid}(-2t) = \tanh(t)$ for $\forall t$,

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

$\nu^{t-1} := \frac{\log\pi^{t-1}(1) - \log\pi^{t-1}(2)}{2}$.

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

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 $\log 2 + \log\cosh\nu = -[\frac{\log\pi(1)+\log\pi(2)}{2}]$ and $\text{sigmoid}(2t) - \text{sigmoid}(-2t) = \tanh(t)$ for $\forall t$.

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

To sum up, we obtain the following.

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

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

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

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

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

□

Theorem B.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) $f_n(\theta, \pi) := -\frac{1}{n}\sum_{i=1}^n [\log p(s_i \mid \theta, \pi)]$ 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.

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

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

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

Theorem B.7. (Derivation for Eq. (3), (4) in Section 2) 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 $\theta, \tanh(\nu)$ at the population level, namely the minizer 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.

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

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.

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

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

Theorem B.8. (Derivation for Eq. (5) in Section 2) 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 $\theta, \tanh(\nu)$ at the finite-sample level, namely the minizer 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 susbstituting $\frac{1}{n}\sum_{i=1}^n, s_i := (x_i, y_i)$ for $\mathbb{E}_{s \sim p(s|\theta^*, \pi^*)}, s := (x, y)$ in the previous Theorem, but note that $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \not\equiv I_d$. □

C. Proof for Results of Population EM Updates

Theorem C.1. (Theorem 3.1 in Section 3) Let $\rho \equiv \rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \cdot \|\theta^*\|}, \bar{\theta} := \frac{\theta^{t-1}}{\sigma}, \bar{\theta}^* := \frac{\theta^*}{\sigma}$, then the EM update rules for $\theta^t, \tanh(\nu^t)$ at Population level are as follows.

$$
\begin{aligned}
& \theta^t := M(\theta^{t-1}, \nu^{t-1}) \\
= \; & \nabla_\theta \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \log\cosh\left(\frac{y\langle x, \theta\rangle}{\sigma^2} + \nu\right)|_{(\theta,\nu)=(\theta^{t-1},\nu^{t-1})} \\
= \; & \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \tanh\left(\frac{y\langle x, \theta^{t-1}\rangle}{\sigma^2} + \nu^{t-1}\right) yx \\
= \; & \left[-\frac{\sigma}{\pi} \cdot \frac{\|\bar{\theta}^*\|^2}{\|\bar{\theta}\|^2} \cdot \frac{\sqrt{1-\rho^2}}{\left(1+(1-\rho^2)\|\bar{\theta}^*\|^2\right)^{\frac{3}{2}}} \cosh^{-1}(\nu^*)\right] \\
\cdot \; & \left\{\tanh(\nu) * \nu \left[\alpha_{\rho,\|\bar{\theta}\|,\|\bar{\theta}^*\|,\nu^*}(\nu)\left(\frac{\frac{\bar{\theta}}{\|\bar{\theta}\|}}{\sqrt{1-\rho^2}\|\bar{\theta}^*\|^2} + \hat{e}_2\right) + \beta_{\rho,\|\bar{\theta}\|,\|\bar{\theta}^*\|,\nu^*}(\nu)\vec{e}_2\right]\right\}_{\nu\leftarrow\nu^{t-1}} \\
& \tanh(\nu^t) := N(\theta^{t-1}, \nu^{t-1}) \\
= \; & \nabla_\nu \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \log\cosh\left(\frac{y\langle x, \theta\rangle}{\sigma^2} + \nu\right)|_{(\theta,\nu)=(\theta^{t-1},\nu^{t-1})} \\
= \; & \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)} \tanh\left(\frac{y\langle x, \theta^{t-1}\rangle}{\sigma^2} + \nu^{t-1}\right) \\
= \; & \frac{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)^{-\frac{1}{2}}}{\pi\|\bar{\theta}\|\cosh(\nu^*)} \int_{\mathbb{R}} \tanh(\nu^{t-1}-\nu')K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu'}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)\cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu'}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right)\mathrm{d}\nu'
\end{aligned}
$$

where $\alpha_{\rho,\|\bar{\theta}\|,\|\bar{\theta}^*\|,\nu^*}(\nu) := \cosh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right) \cdot K_0\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)$

and $\beta_{\rho,\|\bar{\theta}\|,\|\bar{\theta}^*\|,\nu^*}(\nu) := \mathrm{sgn}(\nu)\frac{\sqrt{1+\|\bar{\theta}^*\|^2}}{\|\bar{\theta}^*\|}\sinh\left(\frac{\rho\|\bar{\theta}^*\|\left(\frac{\nu}{\|\bar{\theta}\|}\right)}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]} - \nu^*\right) \cdot K_1\left(\frac{\sqrt{1+\|\bar{\theta}^*\|^2}\cdot\left|\frac{\nu}{\|\bar{\theta}\|}\right|}{[1+(1-\rho^2)\|\bar{\theta}^*\|^2]}\right)$

Proof. With Lemma in Section B, we show that the EM update rules at population level are as follows.

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

Then, by using the Lemma for the evaluation of expectations in Section A, we complete the proof for these closed-from expression in this Theorem. □

Corollary C.2. (Corollary 3.2 in Section 3: EM Updates for No Separation Case) For the special case of no separation of parameters, namely $\mathrm{SNR} := \frac{\|\theta^*\|}{\sigma} \to 0$, the EM update rules for $\theta^t, \tanh(\nu^t)$ at Population level are

$$\begin{aligned}
\bar{\theta}^t &= \frac{\bar{\theta}^0}{\|\bar{\theta}^0\|} \cdot \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\|\bar{\theta}^{t-1}\| x - \nu^{t-1}) x K_0(|x|) \mathrm{d}x \\
\tanh(\nu^t) &= \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\nu^{t-1} - \|\bar{\theta}^{t-1}\| x) K_0(|x|) \mathrm{d}x
\end{aligned}$$

where $\bar{\theta}^t := \frac{\theta^t}{\sigma}$ and $\bar{\theta}^0 := \frac{\theta^0}{\sigma}$.

Remark C.3. *Note that we can rewrite the EM update rule as*

$$\tanh(\nu^t) = \tanh(\nu^{t-1}) \cdot \frac{2}{\pi} \int_{\mathbb{R}_{\geq 0}} \frac{\cosh(2\nu^{t-1}) + 1}{\cosh(2\nu^{t-1}) + \cosh(2\|\bar{\theta}^{t-1}\| x)} K_0(|x|) \mathrm{d}x.$$

Since $\frac{2}{\pi} \int_{\mathbb{R}_{\geq 0}} K_0(|x|) \mathrm{d}x = 1$ *and* $\cosh(2\|\bar{\theta}^{t-1}\| x) \geq 1$*, the EM update rule implies*

$$|\nu^t| \leq |\nu^{t-1}|, \quad \mathrm{sgn}(\nu^t) = \mathrm{sgn}(\nu^{t-1}).$$

If we take the ℓ_2 *norm on both sides of the EM update rule for regression parameters, it follows that*

$$\|\bar{\theta}^t\| = \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\|\bar{\theta}^{t-1}\| x - \nu^{t-1}) x K_0(|x|) \mathrm{d}x \leq \frac{1}{\pi} \int_{\mathbb{R}} |x| K_0(|x|) \mathrm{d}x = \frac{2}{\pi}$$

is bounded.

Proof. As $\mathrm{SNR} := \frac{\|\theta^*\|}{\sigma} \to 0$ in previous Theorem, then coefficients in the EM update rule for θ are as follows.

$$\begin{aligned}
\|\bar{\theta}^*\|^2 \cdot \alpha_{\rho, \|\bar{\theta}\|, \|\bar{\theta}^*\|, \nu^*}(\nu) &\to 0^2 \cdot \cosh(\nu^*) \cdot K_0\left(\frac{|\nu|}{\|\bar{\theta}\|}\right) = 0 \\
\|\bar{\theta}^*\|^2 \cdot \beta_{\rho, \|\bar{\theta}\|, \|\bar{\theta}^*\|, \nu^*}(\nu) &\to 0 \cdot \mathrm{sgn}(\nu) \sinh(-\nu^*) \cdot K_1\left(\frac{|\nu|}{\|\bar{\theta}\|}\right) = 0
\end{aligned}$$

Hence, θ^t has no $\hat{e}_2, \bar{e}_2$ components, and it only contains the $\frac{\bar{\theta}^{t-1}}{\|\bar{\theta}^{t-1}\|}$ component.

$$\begin{aligned}
\bar{\theta}^t := \frac{\theta^t}{\sigma} &= -\frac{\cosh^{-1}(\nu^*)}{\pi \|\bar{\theta}\|^2} \cdot \left\{ \tanh(\nu) * \nu \left[\cosh(\nu^*) \cdot K_0\left(\frac{|\nu|}{\|\bar{\theta}\|}\right) \frac{\bar{\theta}^{t-1}}{\|\bar{\theta}^{t-1}\|} \right] \right\}_{\nu \leftarrow \nu^{t-1}} \\
&= \frac{\bar{\theta}^{t-1}}{\|\bar{\theta}^{t-1}\|} \cdot \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\nu' - \nu^{t-1}) \frac{\nu'}{\|\bar{\theta}\|} \cdot K_0\left(\frac{|\nu'|}{\|\bar{\theta}\|}\right) \mathrm{d} \frac{\nu'}{\|\bar{\theta}\|} \\
&= \frac{\bar{\theta}^{t-1}}{\|\bar{\theta}^{t-1}\|} \cdot \frac{1}{\pi} \int_{\mathbb{R}} \tanh(\|\bar{\theta}^{t-1}\| x - \nu^{t-1}) x \cdot K_0(|x|) \mathrm{d}x
\end{aligned}$$

Note that $\tanh(\nu^{t-1} + \|\bar{\theta}^{t-1}\| x) - \tanh(\nu^{t-1} - \|\bar{\theta}^{t-1}\| x) > 0$ for $\|\bar{\theta}^{t-1}\| \neq 0, x > 0$.

$$\begin{aligned}
\int_{\mathbb{R}} \tanh(\|\bar{\theta}^{t-1}\| x - \nu^{t-1}) x \cdot K_0(|x|) \mathrm{d}x &= \left[\int_0^\infty + \int_{-\infty}^0 \right] \tanh(\|\bar{\theta}^{t-1}\| x - \nu^{t-1}) x \cdot K_0(|x|) \mathrm{d}x \\
&= \int_0^\infty \left(\tanh(\nu^{t-1} + \|\bar{\theta}^{t-1}\| x) - \tanh(\nu^{t-1} - \|\bar{\theta}^{t-1}\| x) \right) x \cdot K_0(|x|) \mathrm{d}x \\
&> 0
\end{aligned}$$

Hence, we conclude that $\frac{\bar{\theta}^t}{\|\bar{\theta}^t\|} = \frac{\bar{\theta}^{t-1}}{\|\bar{\theta}^{t-1}\|} = \cdots = \frac{\bar{\theta}^0}{\|\bar{\theta}^0\|}$.

Consequently, we prove the closed-form expression for the EM update rule for θ.

As $\mathrm{SNR} := \frac{\|\theta^*\|}{\sigma} \to 0$ in previous Theorem, the EM update rule for $\tanh(\nu)$ is as follows.

$$\tanh(\nu^t) = \frac{1}{\pi\|\bar{\theta}\|\cosh(\nu^*)}\int_{\mathbb{R}} \tanh(\nu^{t-1}-\nu')K_0\left(\frac{|\nu'|}{\|\bar{\theta}\|}\right)\cosh(\nu^*)\,\mathrm{d}\nu' = \frac{1}{\pi}\int_{\mathbb{R}} \tanh(\nu^{t-1}-\|\bar{\theta}^{t-1}\|x)K_0(|x|)\mathrm{d}x$$

□

Corollary C.4. (Corollary 3.3 in Section 3: EM Updates in Noiseless Setting) In the noiseless setting, namely $\mathrm{SNR} := \frac{\|\theta^*\|}{\sigma} \to \infty$, the EM update rules for $\theta^t, \tanh(\nu^t)$ at the Population level are

$$\frac{\theta^t}{\|\theta^*\|} = \frac{2}{\pi}\left[\mathrm{sgn}(\rho^{t-1})\varphi^{t-1}\frac{\theta^*}{\|\theta^*\|} + \cos\varphi^{t-1}\frac{\theta^{t-1}}{\|\theta^{t-1}\|}\right]$$
$$\tanh(\nu^t) = \mathrm{sgn}(\rho^{t-1})\left(\frac{2}{\pi}\varphi^{t-1}\right)\cdot\tanh(\nu^*),$$

where $\rho^{t-1} := \frac{\langle\theta^{t-1},\theta^*\rangle}{\|\theta^{t-1}\|\|\theta^*\|}, \varphi^{t-1} := \frac{\pi}{2} - \arccos|\rho^{t-1}|$.

Proof. For brevity, we let $\rho := \rho^{t-1}, \varphi^{t-1} = \frac{\pi}{2} - \arccos|\rho^{t-1}|$, thus $\sqrt{1-\rho^2} = \cos\varphi^{t-1}$

Denote $k := \frac{\|\bar{\theta}^*\|}{\|\bar{\theta}^{t-1}\|}$ and $\alpha^2 := \frac{1}{(1+(1-\rho^2)\|\bar{\theta}^*\|^2)} \to 0_+, x := k\alpha^2\cdot\nu$, when $\|\bar{\theta}^{t-1}\| \to \infty$.

Hence, $\sqrt{1+\|\bar{\theta}^*\|^2} \sim \|\bar{\theta}^*\| \sim \frac{\alpha^{-1}}{\sqrt{1-\rho^2}}$, $\|\bar{\theta}^{t-1}\|^{-1} = k\|\bar{\theta}^*\|^{-1} \sim k\sqrt{1-\rho^2}\alpha$.

Furthermore, $\tanh(\nu^{t-1}-\nu) = \tanh\left(\nu^{t-1} - \frac{x}{k\alpha^2}\right) \to -\mathrm{sgn}(x)$ as $\alpha \to 0_+$.

Evaluating the integral involving $K_0(|x|)$, we obtain the following expression as $\|\bar{\theta}^{t-1}\| \to \infty$, namely $\alpha \to 0_+$.

$$\begin{aligned}
&\tanh(\nu^t) = \mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\tanh\left(\frac{y\langle x,\theta^{t-1}\rangle}{\sigma^2} + \nu^{t-1}\right)\\
=\ &\lim_{\alpha\to 0_+}\frac{1}{\pi\cosh(\nu^*)}k\sqrt{1-\rho^2}\alpha^2\int_{\mathbb{R}}\tanh(\nu^{t-1}-\nu)K_0(k\alpha^2\cdot|\nu|)\cosh(\rho k\alpha^2\cdot\nu-\nu^*)\mathrm{d}\nu\\
=\ &\frac{\sqrt{1-\rho^2}}{\pi\cosh(\nu^*)}\int_{\mathbb{R}} -\mathrm{sgn}(x)K_0(|x|)\cosh(\rho x-\nu^*)\mathrm{d}x\\
=\ &\frac{\sqrt{1-\rho^2}}{\pi\cosh(\nu^*)}\int_0^{+\infty} 2\sinh(\nu^*)\sinh(\rho x)K_0(|x|)\mathrm{d}x\\
=\ &\frac{2\sqrt{1-\rho^2}}{\pi}\tanh(\nu^*)\,\mathrm{sgn}(\rho)\int_0^{+\infty}\sinh(|\rho|x)K_0(|x|)\mathrm{d}x\\
=\ &\mathrm{sgn}(\rho)\tanh(\nu^*)\left[1-\frac{2}{\pi}\arccos|\rho|\right]\\
=\ &\mathrm{sgn}(\rho^{t-1})\left(\frac{2}{\pi}\varphi^{t-1}\right)\cdot\tanh(\nu^*)
\end{aligned}$$

Evaluating the integrals involving $K_0(|x|), K_1(|x|)$ and the Lemma for the relations of unit vectors, we obtain the following expression as $\|\bar{\theta}^{t-1}\| \to \infty$, namely $\alpha \to 0_+$.

$$
\begin{aligned}
\frac{\theta^t}{\|\theta^*\|} &= \frac{\bar{\theta}^t}{\|\bar{\theta}^*\|} = \frac{1}{\|\bar{\theta}^*\|} \cdot \frac{\theta^t}{\sigma} \\
&= \lim_{\alpha \to 0_+} -\frac{\sqrt{1-\rho^2}\cosh^{-1}(\nu^*)}{\pi} k^2 \left(\sqrt{1-\rho^2}\alpha\right) \alpha^3 \tanh(\nu) * \nu \\
&\quad \left[\cosh(\rho k \alpha^2 \cdot \nu - \nu^*) K_0(k\alpha^2 \cdot |\nu|) \left(\frac{\alpha^2}{\sqrt{1-\rho^2}} \cdot \frac{\bar{\theta}}{\|\bar{\theta}\|} + \hat{e}_2\right) + \operatorname{sgn}(\nu) \sinh(\rho k \alpha^2 \cdot \nu - \nu^*) K_1(k\alpha^2 \cdot |\nu|) \vec{e}_2\right] \\
&= -\frac{(1-\rho^2)}{\pi} \cosh^{-1}(\nu^*) \int_{\mathbb{R}} -\operatorname{sgn}(x) x [\cosh(\rho x - \nu^*) K_0(|x|) \hat{e}_2 + \operatorname{sgn}(x) \sinh(\rho x - \nu^*) K_1(|x|) \vec{e}_2] \mathrm{d}x \\
&= \frac{(1-\rho^2)}{\pi} \cosh^{-1}(\nu^*) \left\{ \left[\int_{\mathbb{R}} |x| \cosh(\rho x - \nu^*) K_0(|x|) \mathrm{d}x\right] \hat{e}_2 + \left[\int_{\mathbb{R}} x \sinh(\rho x - \nu^*) K_1(|x|) \mathrm{d}x\right] \vec{e}_2 \right\} \\
&= \frac{(1-\rho^2)}{\pi} \left\{ \left(2\left[\frac{1}{1-\rho^2} + \frac{|\rho|\left(\frac{\pi}{2} - \arccos|\rho|\right)}{(1-\rho^2)^{\frac{3}{2}}}\right]\right) \hat{e}_2 + \left(2\operatorname{sgn}(\rho)\left[\frac{\left(\frac{\pi}{2} - \arccos|\rho|\right)}{(1-\rho^2)^{\frac{3}{2}}} + \frac{|\rho|}{1-\rho^2}\right]\right) \vec{e}_2 \right\} \\
&= \operatorname{sgn}(\rho) \frac{\left(\frac{\pi}{2} - \arccos|\rho|\right)}{\frac{\pi}{2}\sqrt{1-\rho^2}} (\vec{e}_2 + \rho \hat{e}_2) + \left(\frac{\pi}{2}\right)^{-1} (\hat{e}_2 + \rho \vec{e}_2) \\
&= \operatorname{sgn}(\rho) \frac{\left(\frac{\pi}{2} - \arccos|\rho|\right)}{\frac{\pi}{2}\sqrt{1-\rho^2}} \cdot \sqrt{1-\rho^2} \frac{\theta^*}{\|\theta^*\|} + \left(\frac{\pi}{2}\right)^{-1} \sqrt{1-\rho^2} \frac{\theta}{\|\theta\|} \\
&= \frac{2}{\pi}\left[\operatorname{sgn}(\rho^{t-1}) \varphi^{t-1} \frac{\theta^*}{\|\theta^*\|} + \cos\varphi^{t-1} \frac{\theta^{t-1}}{\|\theta^{t-1}\|}\right]
\end{aligned}
$$

□

Lemma C.5. In the noiseless setting, the EM update rules at the population level for 2MLR are

$$
\begin{aligned}
M(\theta^{t-1}, \nu^{t-1}) &= \mathbb{E}_{x \sim p(x)} |\langle x, \theta^* \rangle| \operatorname{sgn}\langle x, \theta^{t-1} \rangle x \\
N(\theta^{t-1}, \nu^{t-1}) &= \mathbb{E}_{x \sim p(x)} \mathbb{E}_{z \sim \mathcal{CAT}(\pi^*)} (-1)^{z+1} \operatorname{sgn}\langle x, \theta^* \rangle \operatorname{sgn}\langle x, \theta^{t-1} \rangle
\end{aligned}
$$

In the noiseless setting, the EM update rules at the finite-sample level for 2MLR are

$$
\begin{aligned}
M_n(\theta^{t-1}, \nu^{t-1}) &= \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n |\langle x_i, \theta^* \rangle| \operatorname{sgn}\langle x_i, \theta^{t-1} \rangle x_i \\
N_n(\theta^{t-1}, \nu^{t-1}) &= \frac{1}{n}\sum_{i=1}^n (-1)^{z_i+1} \operatorname{sgn}\langle x_i, \theta^* \rangle \operatorname{sgn}\langle x_i, \theta^{t-1} \rangle
\end{aligned}
$$

Proof. By letting $\sigma \to 0_+$, SNR $:= \frac{\|\theta^*\|}{\sigma} \to 0_+$, then $y \to (-1)^{z+1}\langle x, \theta^* \rangle, \tanh\left(\frac{y\langle x, \theta^{t-1} \rangle}{\sigma^2} + \nu^{t-1}\right) \to \operatorname{sgn}\langle x, \theta^* \rangle \operatorname{sgn}\langle x, \theta^{t-1} \rangle$. Hence, these expressions are proved by taking the limits. □

Lemma C.6. In the noiseless setting, the easy EM update rule for θ at the finite-sample level for 2MLR is

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

Proof. This Lemma is proved by taking the limit $\sigma \to 0_+$ for easy EM update. □

Lemma C.7. Let $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$ and $\varphi = \frac{\pi}{2} - \arccos|\rho|$, then the identity holds.

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}|\theta^{*\top} x x^\top \theta| &= \|\theta^*\|\|\theta\| \cdot \left\{ \left[1 - \frac{\arccos|\rho|}{\frac{\pi}{2}} \right] |\rho| + \left(\frac{\pi}{2}\right)^{-1} \sqrt{1-\rho^2} \right\} \\
&= \|\theta^*\|\|\theta\| \cdot \frac{2}{\pi}(\varphi \sin\varphi + \cos\varphi)
\end{aligned}$$

Proof. Decompose $x = \tilde{x} + \in \mathbb{R}^d$, where $\tilde{x} \in \operatorname{span}\{\theta, \theta^*\}$.

Let $\theta = \|\theta\| \left(\rho \hat{e}_1 + \sqrt{1-\rho^2}\hat{e}_2 \right), \theta^* = \|\theta^*\|\hat{e}_1$ and $\tilde{x} = \lambda_1 \hat{e}_1 + \lambda_2 \hat{e}_2$, since $\lambda_1, \lambda_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$

and $r = \sqrt{\lambda_1^2 + \lambda_2^2}, \lambda_1 + \mathtt{i}\lambda_2 = r\exp(\mathtt{i}\alpha)$ and $\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\alpha$

Note $(\pi - 2\arccos\rho)\rho$ is an even function

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}|\theta^{*\top} x x^\top \theta| &= \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}|\theta^{*\top} \tilde{x} \tilde{x}^\top \theta| \\
&= \|\theta^*\|\|\theta\| \mathbb{E}_{\lambda_1,\lambda_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)} \left| \lambda_1 \left(\lambda_1 \rho + \lambda_2 \sqrt{1-\rho^2} \right) \right| \\
&= \|\theta^*\|\|\theta\| \iint\limits_{\mathbb{R}^2} \left| \lambda_1 \left(\lambda_1 \rho + \lambda_2 \sqrt{1-\rho^2} \right) \right| \frac{1}{2\pi} \exp\left(-\frac{\lambda_1^2 + \lambda_2^2}{2} \right) \mathrm{d}x\mathrm{d}y \\
&= \|\theta^*\|\|\theta\| \left[\frac{1}{2\pi} \int\limits_{\mathbb{R}_{\geq 0}} r^3 \exp\left(-\frac{r^2}{2} \right) \mathrm{d}r \right] \cdot \left[\int\limits_{[0,2\pi]} |\cos\alpha \cdot \cos(\alpha - \arccos\rho)| \mathrm{d}\alpha \right] \\
&= \|\theta^*\|\|\theta\| \cdot \frac{1}{\pi} \cdot \left[(\pi - 2\arccos\rho)\rho + 2\sqrt{1-\rho^2} \right] \\
&= \|\theta^*\|\|\theta\| \cdot \left\{ \left[1 - \frac{\arccos\rho}{\frac{\pi}{2}} \right] \rho + \left(\frac{\pi}{2}\right)^{-1} \sqrt{1-\rho^2} \right\} \\
&= \|\theta^*\|\|\theta\| \cdot \left\{ \left[1 - \frac{\arccos|\rho|}{\frac{\pi}{2}} \right] |\rho| + \left(\frac{\pi}{2}\right)^{-1} \sqrt{1-\rho^2} \right\} \\
&= \|\theta^*\|\|\theta\| \cdot \frac{2}{\pi}(\varphi \sin\varphi + \cos\varphi)
\end{aligned}$$

□

Lemma C.8. (Grothendieck's Identity) Let $\rho := \frac{\langle \theta, \theta^* \rangle}{\|\theta\| \cdot \|\theta^*\|}$ and $\varphi = \frac{\pi}{2} - \arccos|\rho|$, then the identity holds.

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)} \operatorname{sgn}\langle x, \theta^* \rangle \operatorname{sgn}\langle x, \theta \rangle = \frac{2}{\pi} \operatorname{sgn}\langle \theta, \theta^* \rangle \varphi$$

Note that $\frac{1}{2} + \frac{1}{2}\operatorname{sgn}\langle x, \theta^* \rangle \operatorname{sgn}\langle x, \theta \rangle \in \{1, 0\}$ is a Binomial random variable, hence

$$\begin{aligned}
\mathbb{P}\left[\operatorname{sgn}\langle x, \theta^* \rangle \operatorname{sgn}\langle x, \theta \rangle = +1 \right] &= \frac{1}{2} + \frac{1}{\pi} \operatorname{sgn}\langle \theta, \theta^* \rangle \varphi \\
\mathbb{P}\left[\operatorname{sgn}\langle x, \theta^* \rangle \operatorname{sgn}\langle x, \theta \rangle = -1 \right] &= \frac{1}{2} - \frac{1}{\pi} \operatorname{sgn}\langle \theta, \theta^* \rangle \varphi
\end{aligned}$$

D. Proof for Results at the Population Level

Theorem D.1. (Proposition 4.3 in Section 4: Recurrence Relation) Assume the initial sub-optimality cosine satisfies $\rho^0 := \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|} \neq \pm 1$, or equivalently $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in [0, \frac{\pi}{2})$. Then the recurrence relation for EM updates at population level characterized by the sub-optimality angle is

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

Proof. As $\|\bar{\theta}^*\| \to \infty$, we can obtain the EM update rule for θ^t in previous Corollary 3.3.

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

Let $\varphi := \frac{\pi}{2} - \arccos|\rho| \in [0, \frac{\pi}{2})$, then since ρ^0, ρ^{t-1} have the same sign (validated by checking the sign of $\langle \theta^t, \theta^* \rangle$)

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

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

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

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

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 D.2. (Proposition 4.4 in Section 4: Cycloid Trajectory) If $\rho^0 := \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|} \neq \pm 1$, namely $\phi^0 := 2\arccos|\rho^0| \in (0, \pi]$. Then the coordinates $\mathrm{x}^t, \mathrm{y}^t$ of normalized vector $\frac{\theta^t}{\|\theta^*\|} = \mathrm{x}^t\hat{e}_1 + \mathrm{y}^t\hat{e}_2^t = \mathrm{x}^t\hat{e}_1 + \mathrm{y}^t\hat{e}_2^0, \forall t \in \mathbb{N}_+$ for EM updates at the Population level can be parameterized with the angle $\phi^{t-1} := 2\arccos|\rho^{t-1}| \in (0, \pi]$ as follows, where $\rho^{t-1} := \frac{\langle \theta^{t-1}, \theta^* \rangle}{\|\theta^{t-1}\| \cdot \|\theta^*\|}$.

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

Hence, the trajectory of iterations $\theta^t, \forall t \in \mathbb{N}_+$ is on the cycloid with a parameter $\frac{\|\theta^*\|}{\pi}$, on the plane $\operatorname{span}\{\theta^0, \theta^*\}$.

Proof. Let's prove this, by using the recurrence relation in Proposition 4.3.

Since $\tan\varphi^t = \tan\varphi^{t-1} + \varphi^{t-1}(\tan^2\varphi^{t-1}+1)$ in Proposition 4.3, it shows that $\tan\varphi^t \geq \tan\varphi^{t-1} \geq 0$, therefore $0 \leq \varphi^0 \leq \varphi^1 \leq \cdots \leq \varphi^{t-1} \leq \varphi^t < \frac{\pi}{2}$.

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

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

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 $\hat{e}_2^{t-1}, \hat{e}_2^t \in \mathrm{span}\{\theta^t, \theta^{t-1}, \theta^*\} \subset \mathrm{span}\{\theta^0, \theta^*\}$, $\|\hat{e}_2^{t-1}\| = \|\hat{e}_2^t\| = 1$, we validate $\hat{e}_2^0 = \cdots = \hat{e}_2^{t-1} = \hat{e}_2^t$.

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

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

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

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

With the recurrence relation $\sin\varphi^t \cdot \frac{\|\theta^t\|}{\|\theta^*\|} = \left(\frac{\pi}{2}\right)^{-1}[\varphi^{t-1} + \cos\varphi^{t-1}\sin\varphi^{t-1}]$, $\cos\varphi^t \cdot \frac{\|\theta^t\|}{\|\theta^*\|} = \left(\frac{\pi}{2}\right)^{-1}\cos^2\varphi^{t-1}$, which we showed in the proof of Proposition 4.3, we derive the implicit equation of $\mathrm{x}^t, \mathrm{y}^t (t \geq 1)$

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

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

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

Let $\phi := 2\left(\frac{\pi}{2} - \varphi\right) \in (0, \pi]$, then we rewrite the implicit equations of $\mathrm{x}^t, \mathrm{y}^t (t \geq 1)$ (Harris & Stöcker, 1998).

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

□

Theorem D.3. (Proposition 4.5 in Section 4: Quadratic Convergence Rate) If $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in \left(0, \frac{\pi}{2}\right)$, then the EM updates at population level satisfies

$$\tan\varphi^t \geq \frac{1+\sqrt{5}}{2} \cdot \tan\varphi^{t-1}.$$

Particularly, if $\varphi^{t-1} := \frac{\pi}{2} - \arccos|\rho^{t-1}| \in [\arctan 1.5, \frac{\pi}{2})$, then the EM updates at the Population level satisfies

$$\frac{\pi}{2}\left(\tan\varphi^t - \frac{\pi}{4}\right) \geq \left\{\frac{\pi}{2}\left(\tan\varphi^{t-1} - \frac{\pi}{4}\right)\right\}^2.$$

Proof. Let's prove the inequalites in the Propostion 4.5 as follows.

Let's define $a^t := \tan\varphi^t$, then start from the recurrence relation in Proposition 4.3, note that $a^0 > 0$.

$$a^t - a^{t-1} = \arctan a^{t-1}([a^{t-1}]^2 + 1) > 0$$

Firstly, for the case of $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in \left(0, \frac{\pi}{2}\right)$, since $\frac{\mathrm{d}^2 \arctan(a)}{\mathrm{d}a^2} = -\frac{2a}{(a^2+1)^2} < 0$, it is a concave function.

$$\frac{\arctan a^{t+1} - \arctan a^t}{a^{t+1} - a^t} > \frac{\mathrm{d}\arctan(a)}{\mathrm{d}a}\Big|_{a=a^{t+1}} = \frac{1}{1+[a^{t+1}]^2} = \frac{\arctan a^{t+1}}{a^{t+2} - a^{t+1}}$$

Then for $\forall t \geq 0$, we obtain the following inequality.

$$\frac{a^{t+2} - a^{t+1}}{a^{t+1} - a^t} > \frac{1}{1 - \frac{\arctan a^t}{\arctan a^{t+1}}} > 1 + \frac{\arctan a^t}{\arctan a^{t+1}} = 1 + \left[\frac{a^{t+2} - a^{t+1}}{a^{t+1} - a^t}\right]^{-1} \frac{1+[a^{t+2}]^2}{1+[a^{t+1}]^2} > 1 + \left[\frac{a^{t+2} - a^{t+1}}{a^{t+1} - a^t}\right]^{-1}$$

Hence, by solving $x > 1 + x^{-1} (x > 0)$, we show that

$$\frac{a^{t+2} - a^{t+1}}{a^{t+1} - a^t} = \left(\frac{a^{t+2}}{a^{t+1}} - 1\right)\left\{1 + \frac{1}{\frac{a^{t+1}}{a^t} - 1}\right\} > \frac{\sqrt{5}+1}{2} \quad \forall t \geq 0$$

With $a^1 - a^0 = \arctan a^0([a^0]^2 + 1) > a^0$, we conclude that $\frac{a^1}{a^0} - 1 > 1 > \frac{\sqrt{5}-1}{2}$.

If we assume that $\frac{a^{t+1}}{a^t} - 1 > \frac{\sqrt{5}-1}{2}$, then the inequality below shows that $\frac{a^{t+2}}{a^{t+1}} - 1 > \frac{\sqrt{5}-1}{2}$.

$$\left(\frac{a^{t+2}}{a^{t+1}} - 1\right) \cdot \frac{\frac{\sqrt{5}+1}{2}}{\frac{\sqrt{5}-1}{2}} = \left(\frac{a^{t+2}}{a^{t+1}} - 1\right)\left\{1 + \frac{1}{\frac{\sqrt{5}-1}{2}}\right\} > \left(\frac{a^{t+2}}{a^{t+1}} - 1\right)\left\{1 + \frac{1}{\frac{a^{t+1}}{a^t} - 1}\right\} > \frac{\sqrt{5}+1}{2}$$

By mathematical induction, $\frac{a^{t+1}}{a^t} > \frac{\sqrt{5}+1}{2}$ for $\forall t \geq 0$. Therefore, we have proved the following inequality.

$$\tan\varphi^t \geq \frac{1+\sqrt{5}}{2} \cdot \tan\varphi^{t-1} \quad \forall t \in \mathbb{N}_+$$

Secondly, for the case of $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in \left(\arctan 1.5, \frac{\pi}{2}\right)$, then we have $a^0 = \tan\varphi^0 \geq 1.5$.

Applying the elementary inequality $\arctan a > \frac{\pi a}{2a+\pi}$, $\forall a > 0$, and noting the fact that $a^t \geq a^0 \geq 1.5$, and $1.5 - \frac{4\pi^2+\pi^4}{8(2\cdot 1.5+\pi)} + \frac{1}{8}\pi(4+\pi^2) \approx 4.16 > \frac{\pi}{4} + \frac{\pi^3}{32} \approx 1.75$

$$\begin{aligned}
a^{t+1} &= a^t + \arctan a^t(1+[a^t]^2) \\
&> a^t + \frac{\pi a^t}{2a^t+\pi} \cdot (1+[a^t]^2) \\
&= \frac{\pi}{2}[a^t]^2 - \frac{\pi^2}{4}a^t + a^t - \frac{4\pi^2+\pi^4}{8(2a^t+\pi)} + \frac{1}{8}\pi(4+\pi^2) \\
&\geq \frac{\pi}{2}[a^t]^2 - \frac{\pi^2}{4}a^t + 1.5 - \frac{4\pi^2+\pi^4}{8(2\cdot 1.5+\pi)} + \frac{1}{8}\pi(4+\pi^2) \\
&> \frac{\pi}{2}[a^t]^2 - 2\cdot\frac{\pi}{2}\cdot\frac{\pi}{4}a^t + \frac{\pi}{4} + \frac{\pi^3}{32} \\
&= \frac{\pi}{4} + \frac{\pi}{2}\left[a^t - \frac{\pi}{4}\right]^2
\end{aligned}$$

Therefore, we have proved the following inequality.

$$\frac{\pi}{2}\left(\tan\varphi^t - \frac{\pi}{4}\right) \geq \left\{\frac{\pi}{2}\left(\tan\varphi^{t-1} - \frac{\pi}{4}\right)\right\}^2 \quad \forall t \in \mathbb{N}_+$$

□

Corollary D.4. (Corollary 4.2 in Section 4: Error of Mixing Weights π^t) In the noiseless setting, the error of mixing weights for EM updates at the population level is

$$\|\pi^t - \bar{\pi}^*\|_1 = \left|1 - \frac{2}{\pi}\varphi^{t-1}\right| \cdot \left\|\frac{1}{2} - \pi^*\right\|_1$$

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

Proof. Using Corollary 3.3, and note that $\operatorname{sgn}(\rho^{t-1}) = \operatorname{sgn}(\rho^0)$, we obtain that equation.

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

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

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

In the above equation, we use such an identity $\tanh(\nu^*) = \left\|\frac{1}{2} - \pi^*\right\|_1$. □

Theorem D.5. (Theorem 4.1 in Section 4: Population Level Convergence) If the initial sup-optimality cosine $\rho^0 := \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \cdot \|\theta^*\|} \neq 0$, then with the number of total iterations at most $T = \mathcal{O}\left(\log \frac{1}{|\rho^0|} \vee \log\log \frac{1}{\varepsilon}\right)$, the error of EM update at the population level is bounded by $\frac{\|\theta^{T+1} - \operatorname{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} < \varepsilon$, and $\|\pi^{T+1} - \bar{\pi}^*\|_1 = \mathcal{O}(\sqrt{\varepsilon}) \cdot \|\frac{1}{2} - \pi^*\|_1$, where $\bar{\pi}^* := \frac{1}{2} - \operatorname{sgn}(\rho^0)(\frac{1}{2} - \pi^*)$.

Proof. Let's prove Theorem 4.1, and consider the convergence rate of EM updates at population level.

Let $a^t := \tan\varphi^t$, and $\varphi^0 := \frac{\pi}{2} - \arccos|\rho^0| \in \left(0, \frac{\pi}{2}\right)$.

Step 1. Determine the minmum iteration number T' required to ensure $a^{T'} \geq 1.5$

If $a^0 \geq 1.5$, then $T' = 0$.

Otherwise $a^0 < 1.5$, with the inequality in Proposition 4.5, we obtain the following result for $\forall t \in \mathbb{N}_+$.

$$a^t = \left[\prod_{t'=1}^{t} \frac{a^{t'}}{a^{t'-1}}\right] \cdot a^0 \geq \left(\frac{\sqrt{5}+1}{2}\right)^t a^0$$

Hence, if $T' \geq \left\lceil \frac{\log \frac{1.5}{a^0}}{\log \frac{\sqrt{5}+1}{2}} \right\rceil$, we shows that $a^{T'} \geq \left(\frac{\sqrt{5}+1}{2}\right)^{T'} a^0 \geq 1.5$

Step 2. Determine the minimum iteration number T'' required to ensure $a^{T'+T''} > N_\varepsilon$, where N_ε is a big number that replies on ε.

With the inequality in Proposition 4.5, we obtain the following result for $\forall t \in \mathbb{N}_+$ and $t \geq T'$

$$\frac{\pi}{2}\left[a^{t+1} - \frac{\pi}{4}\right] > \left(\frac{\pi}{2}\left[a^t - \frac{\pi}{4}\right]\right)^2$$

Hence, if $T'' \geq \left\lceil \frac{\log\left[\log\left(N_\varepsilon - \frac{\pi}{4}\right) + \log\frac{\pi}{2}\right] - \log\left(\log\left(\frac{\pi}{4}\left(1.5 - \frac{\pi}{4}\right)\right)\right)}{\log 2} \right\rceil = \mathcal{O}(\log[\log N_\varepsilon])$

$$\frac{\pi}{2}\left[a^{T'+T''} - \frac{\pi}{4}\right] > \left(\frac{\pi}{2}\left[a^{T'} - \frac{\pi}{4}\right]\right)^{2^{T''}} \geq \left(\frac{\pi}{2}\left[1.5 - \frac{\pi}{4}\right]\right)^{2^{T''}} \geq \frac{\pi}{2}\left[N_\varepsilon - \frac{\pi}{4}\right]$$

Hence $T'' \leq \mathcal{O}(\log[\log N_\varepsilon])$, then $T := T' + T'' = \mathcal{O}\left(\log \frac{1}{a^0} \vee \log[\log N_\varepsilon]\right)$ ensures $a^T > N_\varepsilon$.

Step 3. Determine the relationship between N_ε and the desired relative error ε.

With Proposition 4.4, we can write the expression for relative error of θ as follows.

$$\begin{aligned}\frac{\|\theta^{T+1}-\operatorname{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|} &= \pi^{-1}\sqrt{(\phi^T-\sin\phi^T)^2+(1-\cos\phi^T)^2} \\ &= \pi^{-1}\left[\frac{\phi^2}{2}-\frac{\phi^4}{72}+\frac{\phi^6}{6480}-o(\phi^6)\right]_{\phi=\phi^T}\end{aligned}$$

By letting $\phi^T=\mathcal{O}\left(\sqrt{\varepsilon}\right)$, we ensure that $\frac{\|\theta^{T+1}-\operatorname{sgn}(\rho^0)\theta^*\|}{\|\theta^*\|}<\varepsilon$.

$\phi^T=\mathcal{O}\left(\sqrt{\varepsilon}\right)$, that is $\frac{\pi}{2}-\varphi^T=\frac{\phi^T}{2}=\mathcal{O}\left(\sqrt{\varepsilon}\right)$.

With Corollary 4.2, we show that $\|\pi^{T+1}-\bar{\pi}^*\|_1=\frac{2}{\pi}\left|\frac{\pi}{2}-\varphi^T\right|\cdot\left\|\frac{1}{2}-\pi^*\right\|_1=\mathcal{O}\left(\sqrt{\varepsilon}\right)\left\|\frac{1}{2}-\pi^*\right\|_1$.

Note that $\phi^T:=2\left(\frac{\pi}{2}-\varphi^T\right)$ and $a^T:=\tan\varphi^T=\tan\left(\frac{\pi}{2}-\frac{\phi^T}{2}\right)$, expand $\tan\left(\frac{\pi}{2}-x\right)=\frac{1}{x}-\frac{x}{3}-\frac{x^3}{45}+o(x^3)$.

$$a^T=\frac{1}{\mathcal{O}\left(\sqrt{\varepsilon}\right)}-\mathcal{O}\left(\sqrt{\varepsilon}\right)-\mathcal{O}\left(\sqrt{\varepsilon}\right)^3+o\left(\mathcal{O}\left(\sqrt{\varepsilon}\right)^3\right)=\Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$$

Let $N_\varepsilon:=\Theta\left(\frac{1}{\sqrt{\varepsilon}}\right)$, with $T=\mathcal{O}\left(\log\frac{1}{\arctan a^0}\vee\log[\log N_\varepsilon]\right)$, we ensure $a^T>N_\varepsilon=\Theta\left(\frac{1}{\sqrt{\varepsilon}}\right)$, then we obtain

$$T=\mathcal{O}\left(\log\frac{1}{\arctan a^0}\vee\log\left[\log\left(\frac{1}{\sqrt{\varepsilon}}\right)\right]\right)=\mathcal{O}\left(\log\frac{1}{\arctan a^0}\vee\log\log\frac{1}{\varepsilon}\right)$$

Note that $\varphi^0:=\frac{\pi}{2}-\arccos|\rho^0|$, and $a^0:=\tan\varphi^0$, then $|\rho^0|=\sin(\varphi^0)=\Theta(\varphi^0)=\Theta(a^0)$ when $a^0<1.5$, thus

$$T=\mathcal{O}\left(\log\frac{1}{a^0}\vee\log\log\frac{1}{\varepsilon}\right)=\mathcal{O}\left(\log\frac{1}{|\rho^0|}\vee\log\log\frac{1}{\varepsilon}\right)$$

□

E. Proof for Results at the Finite-sample Level

E.1. Upper-bound for Statistical Error

Theorem E.1. (Proposition 5.2 in Section 5: Projected Statistical Error) 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. In the noiseless setting, the statistical error is

$$\begin{aligned} M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t) &= \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\right\} \tanh\left(\frac{y_i\langle x_i,\theta^t\rangle}{\sigma^2} + \nu^t\right) y_i x_i \\ &\to \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\} |\langle x_i,\theta^*\rangle| \,\text{sgn}\langle x_i,\theta^t\rangle x_i \end{aligned}$$

Let $x_i = \tilde{x}_i + x_i^\perp$ and $\tilde{x}_i \in \text{span}\{\theta^*,\theta^t\}, x_i^\perp \perp \text{span}\{\theta^*,\theta^t\}$, we may assume $\dim\text{span}\{\theta^*,\theta^t\} = 2$ without loss of generality, and decompose the space into $\mathbb{R}^d = \text{span}\{\theta^*,\theta^t\} \oplus \text{span}\{\theta^*,\theta^t\}^\perp$.

It implies that 2nd term of statistical error of (Easy EM) doesn't depends on dimension d of the space $\mathbb{R}^d$.

$$\begin{aligned} M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t) &= \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\} |\langle x_i,\theta^*\rangle| \,\text{sgn}\langle x_i,\theta^t\rangle x_i \\ &= \frac{1}{n}\sum_{i\in[n]} |\langle \tilde{x}_i,\theta^*\rangle| \,\text{sgn}\langle \tilde{x}_i,\theta^t\rangle x_i^\perp + \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\} |\langle \tilde{x}_i,\theta^*\rangle| \,\text{sgn}\langle \tilde{x}_i,\theta^t\rangle \tilde{x}_i \end{aligned}$$

Let the orthogonal projection matrix P_{θ,θ^*} satisfy $\text{span}(P_{\theta,\theta^*}) = \text{span}\{\theta,\theta^*\}$, the 2nd term is the projected statistical error.

$$\|P_{\theta,\theta^*}[M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t)]\| = \left\|\left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\} |\langle \tilde{x}_i,\theta^*\rangle| \,\text{sgn}\langle \tilde{x}_i,\theta^t\rangle \tilde{x}_i\right\|_2$$

To estimate the projected statistical error, we begin with decompose it into two parts in terms of $\hat{e}_1, \hat{e}_2$.

Let $\tilde{x}_i = \lambda_{1i}\hat{e}_1 + \lambda_{2i}\hat{e}_2, \lambda_{1i}, \lambda_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, so $\tilde{x}_i = \left\langle \begin{pmatrix} \rho \\ \sqrt{1-\rho^2} \end{pmatrix}, \begin{pmatrix} \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right\rangle \vec{e}_1 + \left\langle \begin{pmatrix} \sqrt{1-\rho^2} \\ -\rho \end{pmatrix}, \begin{pmatrix} \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right\rangle \vec{e}_2$.

Let $Z_i := |\lambda_{1i}| \cdot |\rho\lambda_{1i} + \sqrt{1-\rho^2}\lambda_{2i}|$, $Z_i' := |\lambda_{1i}| \,\text{sgn}(\rho\lambda_{1i} + \sqrt{1-\rho^2}\lambda_{2i}) \cdot (-\sqrt{1-\rho^2}\lambda_{1i} + \rho\lambda_{2i})$.

$$\frac{\|P_{\theta,\theta^*}[M_n^{\text{easy}}(\theta,\nu) - M(\theta,\nu)]\|}{\|\theta^*\|} = \frac{1}{\|\theta^*\|}\sqrt{\left|\frac{1}{n}\sum_{i\in[n]}(Z_i - \mathbb{E}[Z_i])\right|^2 + \left|\frac{1}{n}\sum_{i\in[n]}(Z_i' - \mathbb{E}[Z_i'])\right|^2}$$

Note that both $Z_i, Z_i', i \in [n]$ are sub-exponential with parameters $\left(2\cdot 2.91\text{e}(2\pi)^{-\frac{1}{4}}, 2\cdot 2.91\text{e}\right)$ (we explain the reason later). By using the concentration inequality in (Wainwright, 2019) page 29, equation (2.18) for iid sub-exponential r.v.

with parameters, with probability at least $1-2\delta$

$$\left|\frac{1}{n}\sum_{i\in[n]}(Z_i-\mathbb{E}[Z_i])\right|,\left|\frac{1}{n}\sum_{i\in[n]}(Z_i'-\mathbb{E}[Z_i'])\right|\le\max\left(2^{\frac{1}{2}}\left(2\cdot 2.91\mathrm{e}(2\pi)^{-\frac{1}{4}}\right)\sqrt{\frac{\log\frac{1}{\delta}}{n}},2(2\cdot 2.91\mathrm{e})\frac{\log\frac{1}{\delta}}{n}\right)$$

Hence, the proof is complete.

$$\begin{aligned}\frac{\|P_{\theta,\theta^*}[M_n^{\text{easy}}(\theta,\nu)-M(\theta,\nu)]\|}{\|\theta^*\|} &\le \sqrt{2}\max\left(2^{\frac{1}{2}}\left(2\cdot 2.91\mathrm{e}(2\pi)^{-\frac{1}{4}}\right)\sqrt{\frac{\log\frac{1}{\delta}}{n}},2(2\cdot 2.91\mathrm{e})\frac{\log\frac{1}{\delta}}{n}\right)\\ &< \max\left(20\sqrt{\frac{\log\frac{1}{\delta}}{n}},45\frac{\log\frac{1}{\delta}}{n}\right)=\mathcal{O}\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\right)\end{aligned}$$

Now, let's explain why both $Z_i, Z_i', i\in[n]$ are sub-exponential. For brevity, we write Z, Z' for Z_i, Z_i' instead.

$$Z\le\frac{1+|\rho|}{2}(\lambda_1^2+\lambda_2^2)\le\lambda_1^2+\lambda_2^2,\quad Z'\le\frac{1+\sqrt{1-\rho^2}}{2}(\lambda_1^2+\lambda_2^2)\le\lambda_1^2+\lambda_2^2,\quad |\mathbb{E}[Z]|\le 1,\quad |\mathbb{E}[Z']|\le 1$$

Let $W:=\lambda_1^2+\lambda_2^2\sim\chi^2(2)$, then the q-th moment of $|Z|$ for $q\ge 2$, $\mathbb{E}[|Z|^q]\le\mathbb{E}_{W\sim\chi^2(2)}[W^q]=2^q q!$. By Minkowski's Inequality and Stirling's approximation, $-x\log x\le\mathrm{e}^{-1}$ and $\exp\left(\frac{\log q}{q}\right)\le\exp\left(\frac{\log 3}{3}\right)=3^{\frac{1}{3}}$ for $q\ge 2$ and $\left\{\frac{2\sqrt{2\pi}\mathrm{e}^{\frac{1}{12}}}{\mathrm{e}}3^{\frac{1}{6}}+\frac{1}{2}\right\}\approx 2.907<2.91$

$$\begin{aligned}\mathbb{E}[|Z-\mathbb{E}[Z]|^q]^{\frac{1}{q}} &\le \mathbb{E}[|Z|^q]^{\frac{1}{q}}+\mathbb{E}[|\mathbb{E}[Z]|^q]^{\frac{1}{q}}=\mathbb{E}[|Z|^q]^{\frac{1}{q}}+|\mathbb{E}[Z]|\le[2^q q!]^{\frac{1}{q}}+1\\ &\le \frac{2\left[\sqrt{2\pi}\mathrm{e}^{\frac{1}{12}}\right]^{\frac{1}{q}}}{\mathrm{e}}\sqrt{\exp\left(\frac{\log q}{q}\right)}q+1\le\frac{2\sqrt{2\pi}\mathrm{e}^{\frac{1}{12}}}{\mathrm{e}}3^{\frac{1}{6}}q+1\le\left\{\frac{2\sqrt{2\pi}\mathrm{e}^{\frac{1}{12}}}{\mathrm{e}}3^{\frac{1}{6}}+\frac{1}{2}\right\}q<2.91q\end{aligned}$$

Then for $0\le|\lambda|\le\frac{1}{2\cdot 2.91\mathrm{e}}$

$$\begin{aligned}&\mathbb{E}[\exp(\lambda\{Z-\mathbb{E}[Z]\})]\le\mathbb{E}[\exp(|\lambda||Z-\mathbb{E}[Z]|)]=1+\sum_{q=2}^{\infty}\frac{|\lambda|^q\mathbb{E}[|Z-\mathbb{E}[Z]|^q]}{q!}\\ \le\ & 1+\frac{1}{\sqrt{2\pi}}\sum_{q=2}^{\infty}\frac{(2.91\mathrm{e}|\lambda|)^q}{q^{\frac{1}{2}}}\le 1+\frac{1}{\sqrt{2\pi}}\frac{(2.91\mathrm{e}|\lambda|)^2}{1-2.91\mathrm{e}|\lambda|}\le 1+\frac{2}{\sqrt{2\pi}}(2.91\mathrm{e}|\lambda|)^2\le\exp\left(\frac{\left(2\cdot 2.91\mathrm{e}(2\pi)^{-\frac{1}{4}}\right)^2\lambda^2}{2}\right)\end{aligned}$$

Therefore, Z and Z' (the same reason) are sub-exponential with parameters $\left(2\cdot 2.91\mathrm{e}(2\pi)^{-\frac{1}{4}},2\cdot 2.91\mathrm{e}\right)$.

□

Theorem E.2. (Proposition 5.3 in Section 5: Statistical Error) In the noiseless setting, the statistical error of θ for EM updates at the Finite-sample level satisfies

$$\frac{\|M_n(\theta,\nu)-M(\theta,\nu)\|_2}{\|\theta^*\|}=\mathcal{O}\left(\sqrt{\frac{d}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\vee\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

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.

$$
\begin{aligned}
M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t) &= \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\right\}\tanh\left(\frac{y_i\langle x_i,\theta^t\rangle}{\sigma^2}+\nu^t\right)y_i x_i \\
&\to \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\}|\langle x_i,\theta^*\rangle|\operatorname{sgn}\langle x_i,\theta^t\rangle x_i
\end{aligned}
$$

Let $x_i = \tilde{x}_i + x_i^\perp$ and $\tilde{x}_i \in \operatorname{span}\{\theta^*,\theta^t\}, x_i^\perp \perp \operatorname{span}\{\theta^*,\theta^t\}$, we may assume $\dim\operatorname{span}\{\theta^*,\theta^t\} = 2$ without loss of generality, and decompose the space into $\mathbb{R}^d = \operatorname{span}\{\theta^*,\theta^t\}\oplus\operatorname{span}\{\theta^*,\theta^t\}^\perp$.

It implies that 2nd term of statistical error of (Easy EM) doesn't depends on dimension d of the space $\mathbb{R}^d$

$$
\begin{aligned}
M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t) &= \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\}|\langle x_i,\theta^*\rangle|\operatorname{sgn}\langle x_i,\theta^t\rangle x_i \\
= \frac{1}{n}\sum_{i\in[n]}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle x_i^\perp &+ \left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle\tilde{x}_i
\end{aligned}
$$

In the previous Proposition 5.2, we bound the ℓ_2 norm of the second term with

$$
\begin{aligned}
\frac{\|P_{\theta,\theta^*}[M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t)]\|}{\|\theta^*\|} &= \frac{1}{\|\theta^*\|}\left\|\left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle\tilde{x}_i\right\|_2 \\
&= \mathcal{O}\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\right),
\end{aligned}
$$

Let's focus on the first term, we start by rewriting the ℓ_2 norm of the first term in a different notation.

Let the projection matrix $P := \hat{e}_1\hat{e}_1^\top + \hat{e}_2\hat{e}_2^\top$, select an orthonormal basis $\{\hat{e}_3,\cdots,\hat{e}_d\}$ to form $\operatorname{span}\{\theta^*,\theta^t\}^\perp$, and let $Q := (\hat{e}_1,\hat{e}_2,\hat{e}_3,\cdots,\hat{e}_d)$.

Let $\left(\tilde{x}_i', x_i'^\perp\right) := Q^\top x_i$, where $\tilde{x}_i' \in \mathbb{R}^2, x_i'^\perp \in \mathbb{R}^{d-2}$ are independent, and $\tilde{x}_i' \overset{\text{iid}}{\sim} \mathcal{N}(0,I_2), x_i'^\perp \overset{\text{iid}}{\sim} \mathcal{N}(0,I_{d-2})$.

Note that $x_i''^\perp := \operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle Q^\top x_i^\perp = \operatorname{sgn}\langle\tilde{x}_i',(1,0)^\top\rangle\operatorname{sgn}\left\langle\tilde{x}_i',\left(\rho,\sqrt{1-\rho^2}\right)^\top\right\rangle x_i'^\perp$ are indepent from $\tilde{x}_i'$ and $\tilde{x}_i'' := \langle\tilde{x}_i',(1,0)^\top\rangle \overset{\text{iid}}{\sim} \mathcal{N}(0,1), x_i''^\perp \overset{\text{iid}}{\sim} \mathcal{N}(0,I_{d-2})$. We define $\tilde{x}'' := \{\tilde{x}_i''\}_{i\in[n]} \in \mathbb{R}^n, x_j''^\perp := \left\{x_{ij}''^\perp\right\}_{i\in[n]} \in \mathbb{R}^n$ for $j\in[d-2]$ (Remark: $x_{ij}''^\perp$ is the j-th component of $x_i''^\perp$).

The components of projected vector $\alpha_j = \left\langle\frac{\tilde{x}''}{\|\tilde{x}''\|}, x_j''^\perp\right\rangle \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ are independent from each other, and let $Z_1 := \left[\sum_{i\in[n]}(\tilde{x}_i'')^2\right] \sim \chi^2(n), Z_2 := \left[\sum_{j\in[d-2]}\alpha_j^2\right] \sim \chi^2(d-2)$ are independent from each other.

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{i\in[n]}|\langle\tilde{x},\theta^*\rangle|\operatorname{sgn}\langle\tilde{x},\theta^t\rangle x^\perp\right\|_2 &= \frac{\|\theta^*\|}{n}\left\|\sum_{i\in[n]}\tilde{x}_i''x_i''^\perp\right\|_2 = \frac{\|\theta^*\|}{n}\sqrt{\sum_{j\in[d-2]}\left(\sum_{i\in[n]}\tilde{x}_i''x_{ij}''^\perp\right)^2} \\
&= \frac{\|\theta^*\|}{n}\sqrt{\sum_{j\in[d-2]}\|\tilde{x}''\|^2\alpha_j^2} = \frac{\|\theta^*\|}{n}\sqrt{\left[\sum_{i\in[n]}(\tilde{x}_i'')^2\right]\cdot\left[\sum_{j\in[d-2]}\alpha_j^2\right]} = \frac{\|\theta^*\|}{n}\cdot\sqrt{Z_1Z_2}
\end{aligned}
$$

By using the concentration inequality for Chi-square distribution (see Lemma 1, page 1325 in (Laurent & Massart, 2000)),

then with at least probability at least $\left(1-\frac{\delta}{2}\right)^2 \geq 1-\delta$

$$Z_1 \leq \left(\sqrt{n}+\sqrt{\log\frac{2}{\delta}}\right)^2 + \log\frac{2}{\delta}, \quad Z_2 \leq \left(\sqrt{d-2}+\sqrt{\log\frac{2}{\delta}}\right)^2 + \log\frac{2}{\delta}$$

Therefore

$$\sqrt{Z_1 Z_2} \leq 2\left(\sqrt{n}+\sqrt{\log\frac{2}{\delta}}\right)\left(\sqrt{d-2}+\sqrt{\log\frac{2}{\delta}}\right) = 2\sqrt{n(d-2)} + 2\log\frac{2}{\delta} + 2(\sqrt{n}+\sqrt{d-2})\sqrt{\log\frac{2}{\delta}}$$

Note that $d \leq n$, hence we upper-bound the ℓ_2 norm for the first term.

$$\frac{1}{\|\theta^*\|}\left\|\frac{1}{n}\sum_{i\in[n]}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle x_i^{\perp}\right\|_2 \leq 2\sqrt{\frac{(d-2)}{n}} + 2\frac{\log\frac{2}{\delta}}{n} + 4\sqrt{\frac{\log\frac{2}{\delta}}{n}}$$

Bound for Easy EM For easy EM in the noiseless setting, we show the following upper bound for the statistical error (in ℓ_2 norm) by combining the upper-bounds for the 1st and 2nd terms.

$$\begin{aligned}
&\frac{\|M_n^{\text{easy}}(\theta^t,\nu^t) - M(\theta^t,\nu^t)\|_2}{\|\theta^*\|} \\
\leq\ & \frac{1}{\|\theta^*\|}\left\|\frac{1}{n}\sum_{i\in[n]}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle x_i^{\perp}\right\|_2 + \frac{1}{\|\theta^*\|}\left\|\left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)}\right\}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle\tilde{x}_i\right\|_2 \\
\leq\ & \left(2\sqrt{\frac{(d-2)}{n}} + 2\frac{\log\frac{2}{\delta}}{n} + 4\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right) + \max\left(20\sqrt{\frac{\log\frac{8}{\delta}}{n}}, 45\frac{\log\frac{8}{\delta}}{n}\right) = \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)
\end{aligned}$$

Bound for EM For the standard EM, We update parameters θ with

$$\theta^{t+1} \leftarrow M_n(\theta^t,\nu^t) \quad = \quad \left[\tfrac{1}{n}\textstyle\sum_{i\in[n]} x_i x_i^\top\right]^{-1} \tfrac{1}{n}\textstyle\sum_{i\in[n]} \tanh\left(\frac{y_i\langle x_i,\theta^t\rangle}{\sigma^2} + \nu^t\right) y_i x_i$$

In noiseless setting ($\sigma \to 0$, SNR$\to \infty$), consider the difference between EM updates at the finite-sample/population level.

$$\begin{aligned}
M_n(\theta^t,\nu^t) - M(\theta^t,\nu^t) \quad \to \quad & \left[\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right]^{-1}\left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}\right\}|\langle x_i,\theta^*\rangle|\operatorname{sgn}\langle x_i,\theta^t\rangle x_i \\
& + \left[\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right]^{-1}\left\{I_d - \left[\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right]\right\}\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}|\langle x_i,\theta^*\rangle|\operatorname{sgn}\langle x_i,\theta^t\rangle x_i
\end{aligned}$$

By using (Wainwright, 2019) page 162, equation (6.9), where $\gamma_{\min}$ is the minimum eigen value.

$$\mathbb{P}\left[\sqrt{\gamma_{\min}\left[\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right]} \leq (1-\delta) - \sqrt{\frac{d}{n}}\right] \leq \mathrm{e}^{-n\delta^2/2}$$

Let $\mathrm{e}^{-n\delta^2/2} \leftarrow \delta, \delta \leftarrow \sqrt{2}\sqrt{\frac{\log\frac{1}{\delta}}{n}}$, with probability at least $1-\delta$

$$\sqrt{\gamma_{\min}\left[\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right]} \geq 1 - \sqrt{2}\sqrt{\frac{\log\frac{1}{\delta}}{n}} - \sqrt{\frac{d}{n}} = \mathcal{O}(1)$$

Using $\frac{1}{\|\theta^*\|}\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}|\langle x,\theta^*\rangle|\operatorname{sgn}\langle x,\theta^t\rangle x = \operatorname{sgn}(\rho)\left[1-\frac{\arccos|\rho|}{\frac{\pi}{2}}\right]\frac{\theta^*}{\|\theta^*\|}+\left(\frac{\pi}{2}\right)^{-1}\sqrt{1-\rho^2}\frac{\theta^t}{\|\theta^t\|}$

Hence $\frac{\|\mathbb{E}_{s\sim p(s|\theta^*,\pi^*)}|\langle x,\theta^*\rangle|\operatorname{sgn}\langle x,\theta^t\rangle x\|_2}{\|\theta^*\|} = \sqrt{\left[1-\frac{\arccos|\rho|}{\frac{\pi}{2}}\right]^2+\left(\frac{\pi}{2}\right)^{-2}(1-\rho^2)+\frac{4}{\pi}|\rho|\sqrt{1-\rho^2}\left[1-\frac{\arccos|\rho|}{\frac{\pi}{2}}\right]} \in [0,1]$

By using (Wainwright, 2019) page 162, equation (6.12)

$$\mathbb{P}\left[\left\|\left[\frac{1}{n}\sum_{i\in[n]}x_ix_i^\top\right]-I_d\right\|_2 \geq 2\sqrt{\frac{d}{n}}+2\delta+\left(\sqrt{\frac{d}{n}}+\delta\right)^2\right] \leq 2\mathrm{e}^{-n\delta^2/2}$$

let $2\mathrm{e}^{-n\delta^2/2}\leftarrow\delta, \delta\leftarrow\sqrt{2}\sqrt{\frac{\log\frac{2}{\delta}}{n}}$, with probability at least $1-\delta$

$$\left\|\left[\frac{1}{n}\sum_{i\in[n]}x_ix_i^\top\right]-I_d\right\|_2 \leq 2\left(\sqrt{\frac{d}{n}}+2\sqrt{2}\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right)+\left(\sqrt{\frac{d}{n}}+\sqrt{2}\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right)^2$$

$$\begin{aligned}
&\frac{\|M_n(\theta^t,\nu^t)-M(\theta^t,\nu^t)\|_2}{\|\theta^*\|}\\
\leq\ &\left\{\gamma_{\min}\left[\frac{\sum_{i\in[n]}}{n}x_ix_i^\top\right]\right\}^{-1}\cdot\left\{\frac{1}{\|\theta^*\|}\left\|\left\{\frac{\sum_{i\in[n]}}{n}-\mathbb{E}\right\}|\langle x_i,\theta^*\rangle|\operatorname{sgn}\langle x_i,\theta^t\rangle x_i\right\|_2+\left\|\left[\frac{\sum_{i\in[n]}}{n}x_ix_i^\top\right]-I_d\right\|_2\right\}\\
=\ &\frac{\mathcal{O}\left(\sqrt{\frac{d}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\vee\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)+\left\{2\left(\sqrt{\frac{d}{n}}+2\sqrt{2}\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right)+\left(\sqrt{\frac{d}{n}}+\sqrt{2}\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right)^2\right\}}{\mathcal{O}(1)}\\
=\ &\mathcal{O}\left(\sqrt{\frac{d}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\vee\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)
\end{aligned}$$

□

Lemma E.3. (Convergence of θ^t in Single Iteration) 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 $\operatorname{span}\{\hat{e}_1,\hat{e}_2\}\subset\mathbb{R}^d$, and $\varphi\in\left(0,\frac{\pi}{2}\right)$; with a pertubation vector $\varrho\in\mathbb{R}^d$ with lengh $\|\varrho\|=r\in(0,\sin\varphi)$; 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 pertubation 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 $\hat{e}\in\operatorname{span}\{\hat{e}_1,\hat{e}_2\}^\perp, \|\hat{e}\|=1$

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

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

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

For $\psi(p) := \frac{[1-rp]\sin\varphi - r\sqrt{1-p^2}\cos\varphi}{\sqrt{[1+r^2]-2rp}}$ for $p \in [0,1]$, note that $\cos\left(\varphi + \left[\frac{\pi}{2} - \arcsin p\right]\right) \leq \cos\varphi$

Thus $\cos\varphi - r\cos\left(\varphi + \left[\frac{\pi}{2} - \arcsin p\right]\right) > 0$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}p}\log\psi &= \frac{-r\sin\varphi + r\frac{p}{\sqrt{1-p^2}}\cos\varphi}{[1-rp]\sin\varphi - r\sqrt{1-p^2}\cos\varphi} - \frac{1}{2}\cdot\frac{-2r}{[1+r^2]-2rp} \\
&= \frac{r\cdot\left\{\cos\varphi - r\left[p\cos\varphi - \sqrt{1-p^2}\sin\varphi\right]\right\}(p-r)}{\sqrt{1-p^2}\left\{[1-rp]\sin\varphi - r\sqrt{1-p^2}\cos\varphi\right\}\cdot\{[1+r^2]-2rp\}} \\
&= \frac{r\cdot\left\{\cos\varphi - r\cos\left(\varphi + \left[\frac{\pi}{2} - \arcsin p\right]\right)\right\}(p-r)}{\sqrt{1-p^2}\left\{[1-rp]\sin\varphi - r\sqrt{1-p^2}\cos\varphi\right\}\cdot\{[1+r^2]-2rp\}}
\end{aligned}$$

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

$$\begin{aligned}
\sin\varphi' &\geq \psi(p) \\
&\geq \psi(p)\mid_{p=r} \\
&= \sqrt{1-r^2}\sin\varphi - r\cos\varphi \\
&= \sin(\varphi - \arcsin r)
\end{aligned}$$

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

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

□

E.2. Initialization and Convergence of θ

Theorem E.4. (Proposition 5.4 in Section 5: Initialization with Easy EM) In the noiseless setting, suppose we run the sample-splitting finite-sample Easy EM with $n' := \Theta\left(\frac{n}{\log\frac{1}{\delta}} \wedge \left[\frac{n}{\log\frac{1}{\delta}}\right]^2\right)$ fresh samples for each iteration, then after at most $T_0 = \mathcal{O}\left(\log\frac{1}{\delta}\right)$ iterations, it satisfies $\varphi^{T_0} \geq \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$ with probability at least $1-\delta$.

Proof. Suppose we run finite-sample easy EM with refresh samples n' for each iterations, then after some iterations $\varphi \geq \sqrt{\frac{1}{n'}}$. We will prove this in the followings. Let's denote $\hat{\theta}, \hat{\varphi}$ the EM update at population level. With the EM update $\hat{\theta}^{t+1} := M(\theta^t)$ and let $\tilde{x}_i = \lambda_{1i}\vec{e}_1 + \lambda_{2i}\vec{e}_2$ then $\theta^* = \rho\vec{e}_1 + \sqrt{1-\rho^2}\vec{e}_2$. We begin with evaluating or estimating some expectations of Gaussian.

$$\mathbb{E}_{\lambda_{1i},\lambda_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\left|\rho\lambda_{1i} + \sqrt{1-\rho^2}\lambda_{2i}\right|\operatorname{sgn}(\lambda_{1i})\lambda_{2i} = \left(\frac{\pi}{2}\right)^{-1}\operatorname{sgn}(\rho)\sqrt{1-\rho^2}\left[\frac{\pi}{2} - \arccos|\rho|\right] \in [-0.357205, 0.357205]$$

The 2nd moment is $\mathbb{E}_{\lambda_{1i},\lambda_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right|^2\operatorname{sgn}(\lambda_{1i})^2\lambda_{2i}^2 = 3-2\rho^2$. The 3rd moment is bounded by

$$\begin{aligned}
&\mathbb{E}_{\lambda_{1i},\lambda_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\left\{\left|\left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right|\operatorname{sgn}(\lambda_{1i})\lambda_{2i}\right|^3\right\} \leq \left[\mathbb{E}_{\lambda_{1i},\lambda_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right|^4|\lambda_{2i}|^4\right]^{\frac{3}{4}} \\
= \;& [105-120\rho^2+24\rho^4]^{\frac{3}{4}} \leq 105^{\frac{3}{4}}
\end{aligned}$$

Let $X := \left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right|\operatorname{sgn}(\lambda_{1i})\lambda_{2i}$, thus $3 \geqslant \operatorname{Var}[X] \geq 3-2\rho^2-0.357205^2 \geq 2.8724-2\rho^2$

By Minkovski inequality, $\mathbb{E}[|X-\mathbb{E}[X]|^3] \leq (\mathbb{E}[|X|^3]^{\frac{1}{3}}+|\mathbb{E}[X]|)^3 \leq \left(\left[105^{\frac{3}{4}}\right]^{\frac{1}{3}}+0.357205\right)^3 \leq 45.054$.

Then, let's decompose the statistical error into three terms.

$$\begin{aligned}
&\theta^{t+1}-\hat{\theta}^{t+1} = M_n^{\text{easy}}(\theta^t)-M(\theta^t) \\
= \;& \frac{1}{n'}\sum_{i\in[n']}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^t\rangle x_i^{\perp} + \|\theta^*\|\underbrace{\left[\left\{\frac{1}{n'}\sum_{i\in[n']}-\mathbb{E}_{\lambda_{1i},\lambda_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\right\}\left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right|\operatorname{sgn}(\lambda_{1i})\lambda_{2i}\right]}_{T_2}\vec{e}_2 \\
&+\|\theta^*\|\underbrace{\left[\left\{\frac{1}{n'}\sum_{i\in[n']}-\mathbb{E}_{\lambda_{1i},\lambda_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)}\right\}\left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right||\lambda_{1i}|\right]}_{T_1}\vec{e}_1
\end{aligned}$$

Consider such events $\mathcal{E}_1 := \left\{|T_1| \leq \frac{2}{\pi}\right\}, \mathcal{E}_2 := \left\{|T_2| \geq c\sqrt{\frac{1}{n'}}\right\}$. Note that the variance of T_1

$$\operatorname{Var}[T_1] = \frac{1}{n'}\operatorname{Var}\left[\left|\rho\lambda_{1i}+\sqrt{1-\rho^2}\lambda_{2i}\right||\lambda_{1i}|\right] \leq \frac{1}{n'}[3-1^2] = \frac{2}{n'}.$$

Hence, $\mathbb{P}[\mathcal{E}_1] = 1-\mathbb{P}\left\{|T_1| \geq \frac{2}{\pi}\right\} \geq 1-2\frac{\operatorname{Var}[T_1]}{\left(\frac{2}{\pi}\right)^2} \geq 1-2\frac{\frac{2}{n'}}{\left(\frac{2}{\pi}\right)^2} = 1-\frac{\pi^2}{n'}$.

Consider $\mathcal{E}_2$, with **Berry-Esseen** bound for the central limit theorem, (Ross, 2011) theorem 1.1, where $T_2 = \frac{1}{n'}\sum_{i\in[n']}.[X_i-\mathbb{E}[X_i]]$

$$\left|\mathbb{P}\left[\frac{\sqrt{n'}T_2}{\sqrt{\operatorname{Var}[X]}} \leq \frac{c}{\sqrt{\operatorname{Var}[X]}}\right]-\Phi\left(\frac{c}{\sqrt{\operatorname{Var}[X]}}\right)\right| \leq \frac{0.4785}{(\operatorname{Var}[X])^{\frac{3}{2}}}\mathbb{E}[|X-\mathbb{E}[X]|^3]$$

Thus, with $3 \geqslant \operatorname{Var}[X] \geq 3-2\rho^2-0.357205^2 \geq 2.8724-2\rho^2 = 2.8724-\sin^2\varphi^t \geq 2.8724-\frac{1}{n'}$,
and $\mathbb{E}[|X-\mathbb{E}[X]|^3] \leq 45.054$, $2.8724-\frac{1}{n'} \geq 2.8724-\frac{1}{3} \geq 2.539$.

$$\begin{aligned}
\mathbb{P}\left[T_2 \geq c\sqrt{\frac{1}{n'}}\right] &= 1-\mathbb{P}\left[\frac{\sqrt{n'}T_2}{\sqrt{\operatorname{Var}[X]}} \leq \frac{c}{\sqrt{\operatorname{Var}[X]}}\right] \\
&\geq 1-\Phi\left(\frac{c}{\sqrt{\operatorname{Var}[X]}}\right)-\frac{0.4785}{(\operatorname{Var}[X])^{\frac{3}{2}}}\frac{\mathbb{E}[|X-\mathbb{E}[X]|^3]}{\sqrt{n'}} \\
&\geq 1-\Phi\left(\frac{c}{\sqrt{3}}\right)-\frac{0.4785}{\left(2.8724-\frac{1}{n'}\right)^{\frac{3}{2}}}\cdot\frac{45.054}{\sqrt{n'}} \\
&\geqslant 1-\Phi\left(\frac{c}{\sqrt{3}}\right)-\left(\frac{0.4785\cdot 45.054}{2.539^{\frac{3}{2}}}\right)\frac{1}{\sqrt{n'}} \\
&\geq 1-\Phi\left(\frac{c}{\sqrt{3}}\right)-5.3287\cdot\frac{1}{\sqrt{n'}}
\end{aligned}$$

$$\mathbb{P}[\mathcal{E}_2] = \mathbb{P}\left[T_2 \geq c\sqrt{\frac{1}{n'}}\right] + \mathbb{P}\left[-T_2 \geq c\sqrt{\frac{1}{n'}}\right] \geq 2\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right] - 2\cdot 5.3287\cdot\frac{1}{\sqrt{n'}}$$

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}_1\wedge\mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1]+\mathbb{P}[\mathcal{E}_2]-\mathbb{P}[\mathcal{E}_1\vee\mathcal{E}_2] \geq \mathbb{P}[\mathcal{E}_1]+\mathbb{P}[\mathcal{E}_2]-1\\
\geq\ & \left(1-\frac{\pi^2}{n'}\right)+\left(2\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right]-2\cdot 5.3287\cdot\frac{1}{\sqrt{n'}}\right)-1 \geq 2\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right]-\left[\frac{\pi^2}{n'}+\frac{11}{\sqrt{n'}}\right]
\end{aligned}$$

After runing Easy-EM for $\mathcal{T}$ times with independent batches (batch size is n')

To ensure that the probability of $\overline{\mathcal{E}_1\wedge\mathcal{E}_2}$ happens for $\mathcal{T}$ times is less than δ, and let $\mathcal{T} = \Theta\left(\log\frac{1}{\delta}\right)$ and $\left\{2\Phi\left(\frac{c}{\sqrt{3}}\right)-1+\left[\frac{\pi^2}{n'}+\frac{11}{\sqrt{n'}}\right]\right\} < 1$ for large enough $n' > \left\{\frac{11+\sqrt{11^2+8\pi^2\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right]}}{4\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right]}\right\}^2$

$$(1-\mathbb{P}[\mathcal{E}_1\wedge\mathcal{E}_2])^{\mathcal{T}} \leq \left\{1-2\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right]+\left[\frac{\pi^2}{n'}+\frac{11}{\sqrt{n'}}\right]\right\}^{\mathcal{T}} = \left\{2\Phi\left(\frac{c}{\sqrt{3}}\right)-1+\left[\frac{\pi^2}{n'}+\frac{11}{\sqrt{n'}}\right]\right\}^{\mathcal{T}} \leq \delta$$

Otherwise, if $\mathcal{E}_1\wedge\mathcal{E}_2 = \left\{|T_1|\leq\frac{2}{\pi}\right\}\wedge\left\{|T_2|\geq c\sqrt{\frac{1}{n'}}\right\}$ happens in the $\mathcal{T}$-th iteration

$$\begin{aligned}
\theta^{\mathcal{T}+1} &= \hat{\theta}^{\mathcal{T}+1}+\frac{1}{n'}\sum_{i\in[n']}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^{\mathcal{T}}\rangle x_i^{\perp}+\|\theta^*\|[T_1\vec{e}_1+T_2\vec{e}_2]\\
&= M(\theta^{\mathcal{T}})+\frac{1}{n'}\sum_{i\in[n']}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^{\mathcal{T}}\rangle x_i^{\perp}+\|\theta^*\|[T_1\vec{e}_1+T_2\vec{e}_2]
\end{aligned}$$

Then, since $\left\|\frac{M(\theta^{\mathcal{T}})}{\|\theta^*\|}\right\|\leq 1$, with probability at least $1-\delta$,

$$\frac{1}{\|\theta^*\|}\left\|\frac{1}{n'}\sum_{i\in[n']}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^{\mathcal{T}}\rangle x_i^{\perp}\right\| \leq 2\sqrt{\frac{d-2}{n'}}\left[1+\frac{1}{(d-2)-1}\left(2\log\frac{1}{\delta}+1.62\right)\right] = \mathcal{O}\left(\sqrt{\frac{d}{n'}}\vee\frac{\log\frac{1}{\delta}}{n}\right)$$

If we assume $2\sqrt{\frac{d-2}{n'}}\left[1+\frac{1}{(d-2)-1}\left(2\log\frac{1}{\delta}+1.62\right)\right]\leq 0.1$ for large n'

Thus, we conclude that $n'\geq\frac{8\cdot 2.62^2}{0.1^2}$

$$\begin{aligned}
\frac{\|\theta^{\mathcal{T}+1}\|}{\|\theta^*\|} &= \sqrt{\left\|\frac{M(\theta^{\mathcal{T}})}{\|\theta^*\|}+T_1\vec{e}_1+T_2\vec{e}_2\right\|^2+\frac{1}{\|\theta^*\|}\left\|\frac{1}{n'}\sum_{i\in[n']}|\langle\tilde{x}_i,\theta^*\rangle|\operatorname{sgn}\langle\tilde{x}_i,\theta^{\mathcal{T}}\rangle x_i^{\perp}\right\|^2}\\
&\leq \left\|\frac{M(\theta^{\mathcal{T}})}{\|\theta^*\|}+T_1\vec{e}_1+T_2\vec{e}_2\right\|+0.1 \leq \left(1+\frac{2}{\pi}+|T_2|\right)+0.1 = \left(1.1+\frac{2}{\pi}\right)+|T_2|
\end{aligned}$$

Using the results, $\left\langle\frac{M(\theta^{\mathcal{T}})-\theta^{\mathcal{T}}}{\|\theta^*\|},\vec{e}_2\right\rangle = \frac{2}{\pi}\varphi^{\mathcal{T}}\cos(\varphi^{\mathcal{T}})$, $\langle\theta^{\mathcal{T}},\vec{e}_2\rangle = 0$ and $\frac{\pi}{2}\leq\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}},\vec{e}_1\rangle = \frac{\theta^{\mathcal{T}}}{\|\theta^*\|}\leq 1$ and $|T_1|\leq\frac{2}{\pi}$

$$\begin{aligned}
\left|\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\vec{e}_1\rangle\right| &= \left|\left\langle\frac{M(\theta^{\mathcal{T}})-\theta^{\mathcal{T}}}{\|\theta^*\|},\vec{e}_1\right\rangle+\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}},\vec{e}_1\rangle+T_1\right| \leq \left|T_1+\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}},\vec{e}_1\rangle\right|+\left|\left\langle\frac{M(\theta^{\mathcal{T}})-\theta^{\mathcal{T}}}{\|\theta^*\|},\vec{e}_1\right\rangle\right|\\
&\leq \left(\frac{2}{\pi}+1\right)+\frac{\|M(\theta^{\mathcal{T}})-\theta^{\mathcal{T}}\|}{\|\theta^*\|} \leq \left(\frac{2}{\pi}+1\right)+\left(1-\frac{2}{\pi}\right) = 2
\end{aligned}$$

$$\begin{aligned}
\left|\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\vec{e}_2\rangle\right| &= \left|\left\langle\frac{M(\theta^{\mathcal{T}})-\theta^{\mathcal{T}}}{\|\theta^*\|},\vec{e}_2\right\rangle+\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}},\vec{e}_2\rangle+T_2\right| \geq |T_2|-\left\langle\frac{M(\theta^{\mathcal{T}})-\theta^{\mathcal{T}}}{\|\theta^*\|},\vec{e}_2\right\rangle\\
&\geq |T_2|-\frac{2}{\pi}\varphi^{\mathcal{T}}\cos(\varphi^{\mathcal{T}}) \geq |T_2|-\frac{2}{\pi}\sqrt{\frac{1}{n'}}
\end{aligned}$$

Then, with $n' \geq \frac{8\cdot 2.62^2}{0.1^2}$, we conclude that $\sqrt{\frac{1}{n'}} \leq \frac{0.1}{2\sqrt{2}\cdot 2.62} \leq 0.0135$, $\sqrt{1-\frac{1}{n'}} \geq \sqrt{1-\frac{0.1^2}{8\cdot 2.62^2}} \geq 1-0.1^4$

$$\begin{aligned}
&\left|\frac{1}{\|\theta^*\|^2}\langle\theta^{\mathcal{T}+1},\theta^*\rangle\right| = \left|\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\hat{e}_1\rangle\right| = \left|\sin\varphi^t\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\vec{e}_1\rangle + \cos\varphi^t\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\vec{e}_2\rangle\right| \\
\geq\ & \cos\varphi^t\left|\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\vec{e}_2\rangle\right| - \sin\varphi^t\left|\frac{1}{\|\theta^*\|}\langle\theta^{\mathcal{T}+1},\vec{e}_1\rangle\right| \geq \sqrt{1-\frac{1}{n'}}\cdot\left[|T_2| - \frac{2}{\pi}\sqrt{\frac{1}{n'}}\right] - \sqrt{\frac{1}{n'}}\cdot 2 \\
\geq\ & \left[(1-0.1^4)\left(\frac{|T_2|}{\sqrt{\frac{1}{n'}}} - \frac{2}{\pi}\right) - 2\right]\sqrt{\frac{1}{n'}}
\end{aligned}$$

and $|\rho| = \sin\varphi^{\mathcal{T}} \leq \varphi^{\mathcal{T}} < \sqrt{\frac{1}{n'}}$, with $\mathcal{E}_1 \wedge \mathcal{E}_2 = \left\{|T_1| \leq \frac{2}{\pi}\right\} \wedge \left\{|T_2| \geq c\sqrt{\frac{1}{n'}}\right\}$.

By solving $\frac{\left[(1-0.1^4)\left(c-\frac{2}{\pi}\right)-2\right]}{\left(1.1+\frac{2}{\pi}\right)+0.0135c} \geq 1$, we obtain $c \geq 4.43347$, then

$$\begin{aligned}
\varphi^{\mathcal{T}+1} \geq \sin\varphi^{\mathcal{T}+1} &= \frac{|\langle\theta^{\mathcal{T}+1},\theta^*\rangle|}{\|\theta^{\mathcal{T}+1}\|\cdot\|\theta^*\|} = \frac{\left|\frac{1}{\|\theta^*\|^2}\langle\theta^{\mathcal{T}+1},\theta^*\rangle\right|}{\frac{\|\theta^{\mathcal{T}+1}\|}{\|\theta^*\|}} \geq \frac{\left[(1-0.1^4)\left(\frac{|T_2|}{\sqrt{\frac{1}{n'}}} - \frac{2}{\pi}\right) - 2\right]\sqrt{\frac{1}{n'}}}{\left(1.1+\frac{2}{\pi}\right)+|T_2|} \\
\geq\ & \frac{\left[(1-0.1^4)\left(c-\frac{2}{\pi}\right)-2\right]}{\left(1.1+\frac{2}{\pi}\right)+c\sqrt{\frac{1}{n'}}}\sqrt{\frac{1}{n'}} \geq \frac{\left[(1-0.1^4)\left(c-\frac{2}{\pi}\right)-2\right]}{\left(1.1+\frac{2}{\pi}\right)+0.0135c}\sqrt{\frac{1}{n'}} \geq \sqrt{\frac{1}{n'}}
\end{aligned}$$

By choosing $c = 4.43347$, for large $n' \geq 1.103582\times 10^6$, we have $(1-0.01048) + \left[\frac{\pi^2}{n'} + \frac{11}{\sqrt{n'}}\right] < 1$

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_1\wedge\mathcal{E}_2] &\geq 2\left[1-\Phi\left(\frac{c}{\sqrt{3}}\right)\right] - \left[\frac{\pi^2}{n'} + \frac{11}{\sqrt{n'}}\right] = 0.01048 - \left[\frac{\pi^2}{n'} + \frac{11}{\sqrt{n'}}\right] \\
1-\mathbb{P}[\mathcal{E}_1\wedge\mathcal{E}_2] &\leq (1-0.01048) + \left[\frac{\pi^2}{n'} + \frac{11}{\sqrt{n'}}\right] < 1
\end{aligned}$$

With at most $\mathcal{T} = \frac{\log\frac{1}{\delta}}{-\log(1-\mathbb{P}[\mathcal{E}_1\wedge\mathcal{E}_2])} = \mathcal{O}\left(\log\frac{1}{\delta}\right)$ iterations, we can ensure $\varphi^{T_0} > \sqrt{\frac{1}{n'}}$ for some $T_0 \in [\mathcal{T}+1]$. Hence, $T_0 = \mathcal{O}(\mathcal{T}) = \mathcal{O}(\log\frac{1}{\delta})$ and by choosing $n' = \Theta(\frac{n}{\log\frac{1}{\delta}} \wedge (\frac{n}{\log\frac{1}{\delta}})^2)$

$$\varphi^{T_0} > \sqrt{\frac{1}{n'}} = \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$$

The proof is complete. □

Theorem E.5. (Proposition 5.5 in Section 5: Convergence of Angle) In the noiseless setting, suppose $\varphi^0 \geq \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$, run Easy finite-sample EM for $T_1 = \mathcal{O}\left(\log\frac{n}{\log\frac{1}{\delta}}\right)$ iterations followed by the standard finite-sample EM for at most $T' = \mathcal{O}\left(\log\frac{n}{d} \wedge \log\frac{n}{\log\frac{1}{\delta}}\right)$ iterations with all the same $n = \Omega\left(d \vee \log\frac{1}{\delta}\right)$ samples, then it satisfies

$$\varphi^T \geq \frac{\pi}{2} - \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right), \tag{14}$$

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. In this proof, we assume that $\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \leq \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) < 0.1$, and denote by $\Theta := \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$ the threshold for φ.

Besides, we denote $\bar{\theta}, \bar{\varphi}$ for the EM update at population level.

We divide the whole procedure inito three stages.

In Stage 1, $\varphi \geq \text{const}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \Rightarrow \varphi \geq \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$ after at most T_1 iterations of Easy EM.

In Stage 2, $\varphi \geq 4\Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \Rightarrow \varphi > \arctan 1.5$ after at most T_2 iterations of standard EM.

In Stage 3, $\varphi > \arctan 1.5 \Rightarrow \varphi^t > \frac{\pi}{2} - 1.775\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$ after at most T_2 iterations of standard EM.

Stage 1: $\varphi \geq \text{const}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \Rightarrow \varphi \geq \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$

In the first Stage, we run Easy EM $\theta^{t+1} \leftarrow M_n^{\text{easy}}(\theta^t, \nu^t)$, and note that the length of the projected vector is less than or equal to the length of the original one

$$\frac{1}{\|\theta^*\|}\langle\theta^{t+1} - M(\theta^t), \hat{e}_1\rangle \leq \frac{1}{\|\theta^*\|}\left\|\left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\sim\mathcal{N}(0,I_d)}\right\}|\langle\tilde{x},\theta^*\rangle|\text{sgn}\langle\tilde{x},\theta^t\rangle\tilde{x}\right\|_2.$$

With probability at least $1 - 4\delta$

$$\frac{1}{|\theta^*|}\left\|\left\{\frac{1}{n}\sum_{i\in[n]} -\mathbb{E}_{\{x_i\}_{i\in[n]}\sim\mathcal{N}(0,I_d)}\right\}|\langle\tilde{x},\theta^*\rangle|\text{sgn}\langle\tilde{x},\theta^t\rangle\tilde{x}\right\|_2 < \max\left(20\sqrt{\frac{\log\frac{1}{\delta}}{n}}, 45\frac{\log\frac{1}{\delta}}{n}\right) := \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$$

By using the assumption $\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \leq \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) < 0.1$

$$\frac{1}{\|\theta^*\|}\langle\theta^{t+1} - M(\theta^t), \hat{e}_1\rangle < 0.1$$

Then $\frac{\|\theta^{t+1}\|}{\|\theta^*\|} \leq 1 + \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) < 1.1$

Use $\left|\frac{1}{\|\theta^*\|}\langle\theta^{t+1}, \vec{e}_1\rangle\right| \leq 2$, and with assumption $\varphi^t < \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \leq 0.1$

We denote $\Theta := \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$ the threshold for φ.

$$\begin{aligned}
\sin\varphi^{t+1} &= \left(\frac{\|\theta^{t+1}\|}{\|\theta^*\|}\right)^{-1}\left|\frac{1}{\|\theta^*\|^2}\langle\theta^{t+1},\theta^*\rangle\right| = \left(\frac{\|\theta^{t+1}\|}{\|\theta^*\|}\right)^{-1}\left|\frac{1}{\|\theta^*\|}\langle\theta^{t+1},\hat{e}_1\rangle\right| \\
&\geq 1.1^{-1}\left|\frac{1}{\|\theta^*\|}\langle\theta^{t+1},\hat{e}_1\rangle\right| \\
&= 1.1^{-1}\left|\frac{1}{\|\theta^*\|}\langle\theta^{t+1}-M(\theta^t),\hat{e}_1\rangle + \frac{1}{\|\theta^*\|}\langle M(\theta^t),\hat{e}_1\rangle\right| \\
&\geq 1.1^{-1}\left|\frac{1}{\|\theta^*\|}\langle M(\theta^t),\hat{e}_1\rangle\right| - 1.1^{-1}\left|\frac{1}{\|\theta^*\|}\langle\theta^{t+1}-M(\theta^t),\hat{e}_1\rangle\right| \\
&\geq 1.1^{-1}\left|\frac{1}{\|\theta^*\|}\langle M(\theta^t),\hat{e}_1\rangle\right| - 1.1^{-1}\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \\
&\geq 1.1^{-1}\cdot 1.239\cdot\left|\frac{1}{\|\theta^*\|}\langle\theta^t,\hat{e}_1\rangle\right| - 1.1^{-1}\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \\
&\geq 1.1263\cdot\sin\varphi^t - 1.1^{-1}\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)
\end{aligned}$$

That is

$$1.1263\left(\sin\varphi^{t+1} - \frac{1}{1.1\cdot 0.1263}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)\right) \geq \sin\varphi^t - \frac{1}{1.1\cdot 0.1263}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$$

For $\varphi^t < \Theta < 0.1$, we have $\varphi^t \geq \sin\varphi^t \geq (1-0.002)\varphi^t$

When $\varphi^0 \geq \frac{\left(1+\frac{1}{1.1\cdot 0.1263}\right)}{(1-0.002)}\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \approx 8.2143\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$

$$\begin{aligned}
\sin\varphi^0 - \frac{1}{1.1\cdot 0.1263}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) &\geq (1-0.002)\varphi^0 - \frac{1}{1.1\cdot 0.1263}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \\
&\geq \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)
\end{aligned}$$

We could **assume** that $\varphi^0 \geq 8.3\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$, with at most T_1 iterations

$$\begin{aligned}
\varphi^{T_1} \geq \sin\varphi^{T_1} &\geq 1.1263^{T_1}\left[\sin\varphi^0 - \frac{1}{1.1\cdot 0.1263}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)\right] + \frac{1}{1.1\cdot 0.1263}\cdot\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \\
&\geq 1.1263^{T_1}\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right) \\
&\geq \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \equiv \Theta
\end{aligned}$$

Let $T_1 = \mathcal{O}\left(\log\frac{1}{\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)} - \log\frac{1}{\Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)}\right) = \mathcal{O}\left(\log\frac{1}{\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)}\right) = \mathcal{O}\left(\log\frac{n}{\log\frac{1}{\delta}}\right)$.

We use this Lemma in the previous analysis in Stage 1.

Lemma

with $\frac{1}{\|\theta^*\|}|\langle\theta^t,\hat{e}_1\rangle| = \sin\varphi^t \le \varphi^t < \Theta < 0.1$, $\frac{|\langle M(\theta^t),\hat{e}_1\rangle|}{\|M(\theta^t)\|} = \sqrt{[1-\pi^{-1}(\phi^t-\sin\phi^t)]^2+[\pi^{-1}(1-\cos\phi^t)]^2} \le 0.643$ for $\phi^t := 2\left(\frac{\pi}{2}-\varphi^t\right) \in [\pi-0.2,\pi]$

$$\frac{\frac{|\langle M(\theta^t),\hat{e}_1\rangle|}{\|M(\theta^t)\|}}{\sqrt{1-\left(\frac{|\langle M(\theta^t),\hat{e}_1\rangle|}{\|M(\theta^t)\|}\right)^2}} \ge \frac{1+\sqrt{5}}{2}\cdot\frac{\frac{|\langle\theta^t,\hat{e}_1\rangle|}{\|\theta^t\|}}{\sqrt{1-\left(\frac{|\langle\theta^t,\hat{e}_1\rangle|}{\|\theta^t\|}\right)^2}}$$

we conclude that

$$\frac{|\langle M(\theta^t),\hat{e}_1\rangle|}{\|M(\theta^t)\|}\cdot\frac{1}{\sqrt{1-0.643^2}} \ge \frac{1+\sqrt{5}}{2}\cdot\frac{|\langle\theta^t,\hat{e}_1\rangle|}{\|\theta^t\|}$$

with $\frac{\|M(\theta^t)\|}{\|\theta^t\|} \ge 1$

$$\begin{aligned}\left|\frac{1}{\|\theta^*\|}\langle M(\theta^t),\hat{e}_1\rangle\right| &\ge \left[\frac{1+\sqrt{5}}{2}\cdot\sqrt{1-0.643^2}\cdot\frac{\|M(\theta^t)\|}{\|\theta^t\|}\right]\cdot\frac{1}{\|\theta^*\|}|\langle\theta^t,\hat{e}_1\rangle| \\ &\ge \left[\frac{1+\sqrt{5}}{2}\cdot\sqrt{1-0.643^2}\right]\cdot\frac{1}{\|\theta^*\|}|\langle\theta^t,\hat{e}_1\rangle| \\ &\ge 1.239\cdot\left|\frac{1}{\|\theta^*\|}\langle\theta^t,\hat{e}_1\rangle\right|\end{aligned}$$

Stage 2: $\varphi \ge 4\Theta\left(\sqrt{\frac{d}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\vee\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \Rightarrow \varphi > \arctan 1.5$

We denote $r^{t+1} := \frac{\|M_n(\theta^t)-M(\theta^t)\|}{\|\theta^*\|} < \Theta\left(\sqrt{\frac{d}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\vee\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$, and $\varphi^t := \arcsin\frac{|\langle\theta^t,\hat{e}_1\rangle|}{\|\theta^t\|}$, $\bar{\varphi}^{t+1} := \arcsin\frac{\langle M(\theta^t),\hat{e}_1\rangle}{\|\theta^t\|}$ and the update rule $\theta^{t+1} = M_n(\theta^t)$.

$$\varphi^{t+1} \ge \bar{\varphi}^{t+1} - \arcsin r^{t+1}$$

Note that $\arcsin r^{t+1} \le \frac{\pi}{2}r^{t+1}$ and the Lemma E.3 that we just proved

$$\begin{aligned}\tan\bar{\varphi}^{t+1} &> \frac{1+\sqrt{5}}{2}\tan\varphi^t \quad \forall\varphi^t>0 \\ \varphi^{t+1} &\ge \bar{\varphi}^{t+1} - \arcsin r^{t+1}\end{aligned}$$

With the assumption $r^{t+1} \le \Theta := \Theta\left(\sqrt{\frac{d}{n}}\vee\frac{\log\frac{1}{\delta}}{n}\vee\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \le 0.1$, and $\tan(\arcsin\Theta) = \frac{\Theta}{\sqrt{1-\Theta^2}}$

Note that $1-\frac{4}{5}\left[\frac{\Theta}{\sqrt{1-\Theta^2}}\cdot\frac{1+\sqrt{5}}{2}\tan\varphi^t\right] > \frac{1}{1+\frac{\Theta}{\sqrt{1-\Theta^2}}\cdot\frac{1+\sqrt{5}}{2}\tan\varphi^t}$ for $\tan\varphi^t \le 1.5$, since $\frac{\Theta}{\sqrt{1-\Theta^2}}\cdot\frac{1+\sqrt{5}}{2}\tan\varphi^t \le \frac{0.1}{\sqrt{1-0.1^2}}\cdot\frac{1+\sqrt{5}}{2}\cdot 1.5 < \frac{1}{4} = \frac{1-\frac{4}{5}}{\frac{4}{5}}$

If $\tan\varphi^t \ge \frac{1}{5}\Theta$, then $\frac{4(1+\sqrt{5})}{5}\tan\varphi^t \ge \frac{4(1+\sqrt{5})}{5}\cdot\frac{1}{5}\Theta\cdot\frac{\sqrt{1-0.1^2}}{\sqrt{1-\Theta^2}} \ge \frac{8(1+\sqrt{5})\sqrt{1-0.1^2}}{25}\cdot\frac{1}{2}\left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right] > \frac{1}{2}\left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]$

Then, with $\frac{x}{\sqrt{1+x^2}} > x - \frac{x^3}{2}$

$$\begin{aligned}
\tan\varphi^{t+1} &\geq \tan(\bar{\varphi}^{t+1} - \arcsin\Theta) = \frac{\tan\bar{\varphi}^{t+1} - \frac{\Theta}{\sqrt{1-\Theta^2}}}{1 + \frac{\Theta}{\sqrt{1-\Theta^2}} \cdot \tan\bar{\varphi}^{t+1}} \\
&= \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^{-1} - \frac{\frac{\Theta}{\sqrt{1-\Theta^2}} + \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^{-1}}{1 + \frac{\Theta}{\sqrt{1-\Theta^2}} \cdot \tan\bar{\varphi}^{t+1}} \\
&> \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^{-1} - \frac{\frac{\Theta}{\sqrt{1-\Theta^2}} + \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^{-1}}{1 + \frac{\Theta}{\sqrt{1-\Theta^2}} \cdot \frac{1+\sqrt{5}}{2}\tan\varphi^t} \\
&> \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^{-1} - \left\{\frac{\Theta}{\sqrt{1-\Theta^2}} + \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^{-1}\right\}\left\{1 - \frac{4}{5}\left[\frac{\Theta}{\sqrt{1-\Theta^2}} \cdot \frac{1+\sqrt{5}}{2}\tan\varphi^t\right]\right\} \\
&= -\frac{\Theta}{\sqrt{1-\Theta^2}} + \frac{4}{5} \cdot \frac{1+\sqrt{5}}{2}\left[1 + \left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^2\right]\tan\varphi^t \\
&\geq \frac{2\left(1+\sqrt{5}\right)}{5}\tan\varphi^t - \left[\frac{\Theta}{\sqrt{1-\Theta^2}} - \frac{1}{2}\left[\frac{\Theta}{\sqrt{1-\Theta^2}}\right]^3\right] \\
&> \frac{2\left(1+\sqrt{5}\right)}{5}\tan\varphi^t - \Theta \\
&> \frac{22}{17}\tan\varphi^t - \Theta
\end{aligned}$$

Thus, we conclude that if $\frac{1}{5}\Theta \leq \tan\varphi^t \leq 1.5$, then

$$[\tan\varphi^{t+1} - 3.4\Theta] \quad > \quad \frac{22}{17}[\tan\varphi^t - 3.4\Theta]$$

We could assume $\varphi^0 \geq 4\Theta$ then $\tan\varphi^t - 3.4\Theta > \tan\varphi^0 - 3.4\Theta \geq 0.6\Theta$, thus

$$[\tan\varphi^t - 3.4\Theta] \quad > \quad \left(\frac{22}{17}\right)^t [\tan\varphi^0 - 3.4\Theta] \geq \left(\frac{22}{17}\right)^t \cdot 0.6\Theta$$

Therefore, after runing at most $T_2 = \lceil \frac{\log\frac{1.5}{0.6\Theta}}{\log\frac{22}{17}} \rceil = \lceil \frac{\log\frac{1}{\Theta} + \log 2.5}{\log\frac{22}{17}} \rceil = \mathcal{O}\left(\log\frac{1}{\Theta}\right)$

$$\tan\varphi^{T_2} > 3.4\Theta + \left(\frac{22}{17}\right)^{T_2} \cdot 0.6\Theta > 3.4\Theta + 1.5 > 1.5$$

Stage 3: $\varphi > \arctan 1.5 \Rightarrow \varphi^t > \frac{\pi}{2} - 1.775\Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$

Let's start from the following relation

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

We denote $\phi := 2\left(\frac{\pi}{2} - \varphi\right) \in (0, \pi)$, note that $\frac{2}{\phi} - \frac{\phi}{4.93} < \tan\left(\frac{\pi}{2} - \frac{\phi}{2}\right) < \frac{2}{\phi} - \frac{\phi}{6}$

$$\tan\left(\frac{\pi}{2} - \frac{\bar{\phi}^{t+1}}{2}\right) \quad = \quad \tan\left(\frac{\pi}{2} - \frac{\phi^t}{2}\right) + \left(\frac{\pi}{2} - \frac{\phi^t}{2}\right)\left[\tan^2\left(\frac{\pi}{2} - \frac{\phi^t}{2}\right) + 1\right]$$

$$\begin{aligned}
\frac{2}{\bar{\phi}^{t+1}} &> \frac{2}{\bar{\phi}^{t+1}} - \frac{\bar{\phi}^{t+1}}{6} \\
&> \tan\left(\frac{\pi}{2} - \frac{\bar{\phi}^{t+1}}{2}\right) \\
&= \tan\left(\frac{\pi}{2} - \frac{\phi^t}{2}\right) + \left(\frac{\pi}{2} - \frac{\phi^t}{2}\right)\left[\tan^2\left(\frac{\pi}{2} - \frac{\phi^t}{2}\right) + 1\right] \\
&> \frac{2}{\phi^t} - \frac{\phi^t}{4.93} + \left(\frac{\pi}{2} - \frac{\phi^t}{2}\right)\left[\left(\frac{2}{\phi^t} - \frac{\phi^t}{4.93}\right)^2 + 1\right]
\end{aligned}$$

For $0 < \phi^t \leq 1.18805$, i.e. $\varphi^t = \frac{\pi}{2} - \frac{\phi^t}{2} \geq 0.976772, \tan(\varphi^t) \geq 1.48061$, then $45849\pi - 144449\phi^t + 10000\pi[\phi^t]^2 - 10000[\phi^t]^3 \geq 0$

$$\begin{aligned}
\bar{\phi}^{t+1} &< \frac{2}{\frac{2}{\phi^t} - \frac{\phi^t}{4.93} + \left(\frac{\pi}{2} - \frac{\phi^t}{2}\right)\left[\left(\frac{2}{\phi^t} - \frac{\phi^t}{4.93}\right)^2 + 1\right]} \\
&= \frac{[\phi^t]^2}{\pi + \frac{[\phi^t]^2}{972196}\{45849\pi - 144449\phi^t + 10000\pi[\phi^t]^2 - 10000[\phi^t]^3\}} \\
&\leq \frac{2}{\frac{2\pi}{[\phi^t]^2}} = \frac{[\phi^t]^2}{\pi}
\end{aligned}$$

By using the previous Lemma, $\varphi^{t+1} \geq \bar{\varphi}^{t+1} - \arcsin r^{t+1}$

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

Hence, with $\arcsin r^{t+1} \leq \frac{\pi}{2} r^{t+1}$, for $0 < \frac{\phi^t}{\pi} \leq \frac{1.18805}{\pi} \approx 0.378167$

$$\begin{aligned}
\phi^{t+1} - 2\arcsin r^{t+1} \leq \bar{\phi}^{t+1} &< \frac{[\phi^t]^2}{\pi} \\
\frac{\phi^{t+1}}{\pi} < \left[\frac{\phi^t}{\pi}\right]^2 + \frac{2}{\pi}\arcsin r^{t+1} &\leq \left[\frac{\phi^t}{\pi}\right]^2 + r^{t+1}
\end{aligned}$$

Suppose $r^{t+1} \leq \Theta := \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \leq 0.1$, then

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

Suppose $0 < \frac{\phi^t}{\pi} \leq \frac{1.18805}{\pi} \approx 0.378167$, with $\Theta < 0.1$ then in one iteration

$$\frac{\phi^{t+1}}{\pi} < 0.3782^2 + 0.1 < \frac{1}{4}$$

Thus, we could assume $0 < \frac{\phi^t}{\pi} \leq \frac{\phi^0}{\pi} < \frac{1}{4}$, then

$$\begin{aligned}
\left[\frac{\phi^{t+1}}{\pi} - 2\Theta\right] &< \left[\frac{\phi^t}{\pi}\right]^2 - \Theta \\
&= \left[\frac{\phi^t}{\pi} - 2\Theta\right]^2 - \Theta\left(\left[1 - 4\frac{\phi^t}{\pi}\right] + 4\Theta\right) \\
&< \left[\frac{\phi^t}{\pi} - 2\Theta\right]^2
\end{aligned}$$

If $\frac{\phi^0}{\pi} < 2\Theta$, then $\frac{\phi^0}{\pi} < 3\Theta$; otherwise, with at most $T_3 = \lceil \frac{\log\log\frac{1}{\Theta} - \log\log 4}{\log 2} \rceil = \mathcal{O}\left(\log\log\frac{1}{\Theta}\right)$ iterations, $\frac{\phi^{T_3}}{\pi} < 3\Theta$

$$\begin{aligned}
\left[\frac{\phi^{T_3}}{\pi} - 2\Theta\right] &\leq \left[\frac{\phi^0}{\pi} - 2\Theta\right]^{2^{T_3}} < \left[\frac{1}{4} - 2\Theta\right]^{2^{T_3}} < \exp(-\log 4 \cdot 2^{T_3}) \\
&= \frac{1}{\exp\left(\log 4 \cdot 2^{\lceil \frac{\log\log\frac{1}{\Theta} - \log\log 4}{\log 2} \rceil}\right)} \leq \frac{1}{\exp\left(\log 4 \cdot 2^{\frac{\log\log\frac{1}{\Theta} - \log\log 4}{\log 2}}\right)} \\
&= \Theta
\end{aligned}$$

After at most T_3' interations, $0 < \frac{\phi^t}{\pi} < 3\Theta$ is satisfied, then run three more iteration, with $\Theta < 0.1$

$$\begin{aligned}
\frac{\phi^{t+1}}{\pi} &< \left[\frac{\phi^t}{\pi}\right]^2 + \Theta < (9\Theta + 1)\Theta < 1.9\Theta \\
\frac{\phi^{t+2}}{\pi} &< \left[\frac{\phi^{t+1}}{\pi}\right]^2 + \Theta < (1.9^2\Theta + 1)\Theta < 1.361\Theta \\
\frac{\phi^{t+3}}{\pi} &< \left[\frac{\phi^{t+2}}{\pi}\right]^2 + \Theta < (1.361^2\Theta + 1)\Theta < 1.13\Theta
\end{aligned}$$

To sum up, after running for at most $T_3 = 1 + T_3' + 3 = T_3 + 4 = \mathcal{O}\left(\log\log\frac{1}{\Theta}\right)$ iterations, then $0 < \phi^t < (1.13\pi)\Theta$ with $\phi^t = 2\left(\frac{\pi}{2} - \varphi^t\right)$, we conclude that $\varphi^t \in \left(0, \frac{\pi}{2}\right)$ satisfies

$$\varphi^t > \frac{\pi}{2} - \left(1.13\frac{\pi}{2}\right)\Theta > \frac{\pi}{2} - 1.775\Theta$$

Number of Iterations, Statistical Error

We denote $\Theta := \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) < 0.1$.

Note that

$$\begin{aligned}
\mathcal{O}\left(\log\frac{1}{\Theta}\right) &= \mathcal{O}\left(\log\frac{1}{\Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)}\right) = \mathcal{O}\left(\log\frac{1}{\sqrt{\frac{d}{n}}} \wedge \log\frac{1}{\frac{\log\frac{1}{\delta}}{n}} \wedge \log\frac{1}{\sqrt{\frac{\log\frac{1}{\delta}}{n}}}\right) \\
&= \mathcal{O}\left(\log\frac{n}{d} \wedge \log\frac{n}{\log\frac{1}{\delta}}\right)
\end{aligned}$$

In Stage 1, it taks at most $T_1 = \mathcal{O}\left(\log\frac{n}{\log\frac{1}{\delta}}\right)$.

In Stage 2, it taks at most $T_2 = \mathcal{O}\left(\log\frac{1}{\Theta}\right) = \mathcal{O}\left(\log\frac{n}{d} \wedge \log\frac{n}{\log\frac{1}{\delta}}\right)$.

In Stage 3, it taks at most $T_3 = \mathcal{O}\left(\log\log\frac{1}{\Theta}\right) = \mathcal{O}\left(\log\left[\log\frac{n}{d} \wedge \log\frac{n}{\log\frac{1}{\delta}}\right]\right)$.

Hence, the iteration numbers at most to ensure the convergence $\varphi^T \geq \frac{\pi}{2} - \Theta$ with probability at least $1 - T\delta$ is

$$T = T_1 + T_2 + T_3 = \mathcal{O}\left(\log\frac{n}{\log\frac{1}{\delta}}\right)$$

For a good initialization $\varphi^0 \geq 4\Theta$

$$T' = T_2 + T_3 = \mathcal{O}\left(\log\frac{n}{d} \wedge \log\frac{n}{\log\frac{1}{\delta}}\right)$$

□

E.3. Error of Mixing Weights π and Convergence at the Finite-Sample Level

Lemma E.6. Let $q := \max(p, 1-p)$ for $V_i \overset{\text{iid}}{\sim} \text{Bern}(p), \forall i \in [n]$, then

for $t \in \mathbb{R}_{\geq 0}$

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

for $t \in [\mathrm{e}(1-q), q) \neq \varnothing$

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

for $t \in [q, \infty)$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]}(V_i - \mathbb{E}[V_i]) \geq t\right) = 0$$

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

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

Hence, the first probability inequality is proved.

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

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

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

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

If $t \geq q$, then $\tau := \frac{t}{(1-q)} \in [\frac{1}{\gamma}, \infty)$.

$$n\gamma\inf_{\mu'>0}\left\{\frac{1}{\gamma}\log(1+\gamma\mu') - \tau\log(\mu')\right\} = -\infty$$

If $t \in [\mathrm{e}(1-q), q)$, then $\tau := \frac{t}{(1-q)} \in [\mathrm{e}, \frac{1}{\gamma})$.

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

Therefore, the third probability inequality is proved, and we show the probability is 0 when $t \geq q$. □

Theorem E.7. (Theorem 5.1 in Section 5: Convergence at the Finite-Sample Level) In the noiseless setting, suppose any initial mixing weights π^0 and any initial regression parameters $\theta^0 \in \mathbb{R}^d$ ensuring that $\varphi^0 \geq \Theta\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}} \vee \frac{\log\frac{1}{\delta}}{n}\right)$. If we run finite-sample Easy EM for at most $T_1 = \mathcal{O}\left(\log \frac{n}{\log\frac{1}{\delta}}\right)$ iterations followed by the finite-sample standard EM for at most $T' = \mathcal{O}\left(\log \frac{n}{d} \wedge \log \frac{n}{\log\frac{1}{\delta}}\right)$ iterations with all the same $n = \Omega\left(d \vee \log\frac{1}{\delta}\right)$ samples, then we have

$$
\frac{\|\theta^{T+1} - \mathrm{sgn}(\rho^{T+1})\theta^*\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),
$$

$$
\begin{aligned}
\|\pi^{T+1} - \bar{\pi}^*\|_1 = & \left\|\frac{1}{2} - \pi^*\right\|_1 \cdot \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \\
& + c(\pi^*) \cdot \mathcal{O}\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),
\end{aligned}
$$

with probability at least $1 - T\delta$, where $T := T_1 + T', \varphi^0 := \frac{\pi}{2} - \arccos\left|\frac{\langle\theta^0,\theta^*\rangle}{\|\theta^0\|\cdot\|\theta^*\|}\right|, \rho^{T+1} := \frac{\langle\theta^{T+1},\theta^*\rangle}{\|\theta^{T+1}\|\cdot\|\theta^*\|}, \bar{\pi}^* := \frac{1}{2} - \mathrm{sgn}(\rho^0)(\frac{1}{2} - \pi^*)$, and the coefficient $c(\pi^*) = \mathcal{O}(1)$, especially $c(\pi^*) = 0$ when $\pi^* = \{1, 0\}$ or $\{0, 1\}$.

Proof. In the proof of Proposition 5.5 for the converge of angle φ, we show that EM upate rules ensure $\varphi^T > \frac{\pi}{2} - 1.775\Theta$ after enough T iterations, where $\Theta := \Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$ is the threshold for φ.

Using the relation $\phi^T := 2\left(\frac{\pi}{2} - \varphi^T\right) \in (0, \pi)$, namely

$$
\phi^T < 3.55\Theta
$$

In the following proof, we use $\bar{\nu}^{T+1}$ for the EM update at the population level.

Final Statistical Error in Regression Parameters θ

Note that $\sqrt{(\phi - \sin\phi)^2 + (1-\cos\phi)^2} \le \frac{\phi^2}{2}$ for $\forall \phi \in \left[0, \frac{\pi}{2}\right]$, and use tha assumption $\Theta < 0.1$ in Proposition 5.5, the upper bound for the relative error is

$$\begin{aligned}
\frac{\left\|M_n(\theta^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\theta^*\right\|}{\|\theta^*\|} &\le \frac{\left\|M(\theta^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\theta^*\right\|}{\|\theta^*\|} + \frac{\|M_n(\theta^T) - M(\theta^T)\|}{\|\theta^*\|} \\
&\le \frac{1}{\pi}\sqrt{(\phi^T - \sin\phi^T)^2 + (1-\cos\phi^T)^2} + \Theta \\
&\le \frac{[\phi^T]^2}{2\pi} + \Theta \\
&< \frac{3.55^2}{2\pi}\Theta^2 + \Theta \\
&< \left(\frac{3.55^2}{2\pi}\cdot 0.1 + 1\right)\Theta \\
&< 1.21\Theta
\end{aligned}$$

Hence, $\frac{\|\theta^{T+1} - \mathrm{sgn}(\rho^{T+1})\theta^*\|}{\|\theta^*\|} = \frac{\left\|M_n(\theta^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\theta^*\right\|}{\|\theta^*\|} = \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$

Final Statistical Error in Mixing Weights π

By using the Corollary 3.3 in Section 3, and note that $\tanh(\bar{\nu}^{T+1}) := N(\theta^T, \nu^T)$.

$$\tanh(\bar{\nu}^{T+1}) = N(\theta^T, \nu^T) = \tanh(\nu^*)\cdot \mathrm{sgn}\langle\theta^T,\theta^*\rangle\left[\frac{2}{\pi}\varphi^T\right]$$

Note that $\mathbb{E}\left[N_n(\theta^T,\nu^T)\right] = N(\theta^T,\nu^T) = \tanh(\bar{\nu}^{T+1})$, in the noiseless setting, Lemma in Appendix C gives that.

$$\tanh(\nu^{t+1}) = N_n(\theta^T,\nu^T) = \frac{1}{n}\sum_{i\in[n]} \mathrm{sgn}\langle x_i,\theta^*\rangle\mathrm{sgn}\langle x_i,\theta^T\rangle\cdot(-1)^{z_i+1}$$

Note that x_i, z_i are independent, let $W_i := \mathrm{sgn}\langle x_i,\theta^*\rangle\mathrm{sgn}\langle x_i,\theta^T\rangle\cdot(-1)^{z_i+1}$.

with $\mathbb{P}(z_i = 1) = \frac{1}{2} + \frac{1}{2}\tanh(\nu^*), \mathbb{P}(z_i = 2) = \frac{1}{2} - \frac{1}{2}\tanh(\nu^*)$

Using Lemma C.8 (Grothendieck's Identity), we show that

$$\begin{aligned}
\mathbb{P}[\mathrm{sgn}\langle x_i,\theta^*\rangle\mathrm{sgn}\langle x_i,\theta^T\rangle = +1] &= \left[\frac{1}{2} + \mathrm{sgn}\langle\theta^T,\theta^*\rangle\frac{\varphi^T}{\pi}\right] \\
\mathbb{P}[\mathrm{sgn}\langle x_i,\theta^*\rangle\mathrm{sgn}\langle x_i,\theta^T\rangle = -1] &= \left[\frac{1}{2} - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\frac{\varphi^T}{\pi}\right]
\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}
\mathbb{P}[W_i = +1] &= \mathbb{P}[\mathrm{sgn}\langle x_i,\theta^*\rangle\mathrm{sgn}\langle x_i,\theta^T\rangle = +1]\cdot\mathbb{P}(z_i = 1) + \mathbb{P}[\mathrm{sgn}\langle x_i,\theta^*\rangle\mathrm{sgn}\langle x_i,\theta^T\rangle = -1]\cdot\mathbb{P}(z_i = 2) \\
&= \frac{1}{2} + \mathrm{sgn}\langle\theta^T,\theta^*\rangle\frac{\varphi^T}{\pi}\cdot\tanh(\nu^*) \\
\mathbb{P}[W_i = -1] &= \frac{1}{2} - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\frac{\varphi^T}{\pi}\cdot\tanh(\nu^*)
\end{aligned}$$

Let $V_i := \frac{1}{2}(W_i + 1) \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$ be Bernoulli distribution with the parameter $p := \frac{1}{2} + \mathrm{sgn}\langle\theta^T,\theta^*\rangle\frac{\varphi^T}{\pi}\cdot\tanh(\nu^*)$.

Therefore, with $\frac{1}{2}N_n(\theta^T,\nu^T) + \frac{1}{2} = \frac{1}{n}\sum_{i\in[n]} V_i$ and $\frac{1}{2}N(\theta^T,\nu^T) + \frac{1}{2} = \mathbb{E}[V_i] = \frac{1}{n}\sum_{i\in[n]}\mathbb{E}[V_i]$.

By using the probability inequalities in Lemma E.6

$$\begin{aligned}
\mathbb{P}[|N_n(\theta^T,\nu^T) - N(\theta^T,\nu^T)| \ge 2t] &= \mathbb{P}\left[\left|\frac{1}{n}\sum_{i\in[n]}(V_i - \mathbb{E}[V_i])\right| \ge t\right] \\
&\le 2\exp(-2nt^2)
\end{aligned}$$

Consequently, by letting $\exp(-2nt^2) \leftarrow \delta$, with probability at least $1-2\delta$

$$|N_n(\theta^T,\nu^T) - N(\theta^T,\nu^T)| < \sqrt{\frac{2\log\frac{1}{\delta}}{n}}$$

Furthermore, when $\pi^* = [1,0]$ or $[0,1]$, namely $\nu^* \to \pm\infty$, then $q := \max(p, 1-p) \to \frac{1}{2} + \frac{\varphi^T}{\pi}$. Since after enough iterations, EM updates ensure that $\varphi^T > \frac{\pi}{2} - 1.775$. Hence, $q := \frac{1}{2} + \frac{\varphi^T}{\pi} > 1 - 0.565\Theta, 1-q < 0.565\Theta$.

Using Lemma E.6, we show that for $t \in [\mathrm{e}(1-q), q) \neq \varnothing$

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

By choosing $t = 0.565\mathrm{e}\Theta \in [\mathrm{e}(1-q), \infty)$, note that $\Theta^2 \geq \frac{\log\frac{1}{\delta}}{n}$, then with probability at least

$$\begin{aligned}
1 - \exp\left\{-n\left\{\frac{t}{q}\left[\log\frac{t}{(1-q)} - 1\right] + \frac{t^2}{2q^2}\right\}\right\} &\geq 1 - \exp\left\{-n\frac{0.565^2\mathrm{e}^2}{2}\Theta^2\right\} \\
&\geq 1 - \exp\left\{-1.179n\cdot\frac{\log\frac{1}{\delta}}{n}\right\} \\
&\geq 1 - \delta
\end{aligned}$$

Hence $|N_n(\theta^T,\nu^T) - N(\theta^T,\nu^T)| = 2\frac{1}{n}\sum_{i\in[n]}|V_i - \mathbb{E}[V_i]| \geq 2\cdot 0.565\mathrm{e}\Theta = 1.13\mathrm{e}\Theta$ with probability at least $1-2\delta$, when $\pi^* = [1,0]$ or $[0,1]$.

$$\begin{aligned}
|N_n(\theta^T,\nu^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\tanh(\nu^*)| &\leq |N(\theta^T,\nu^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\tanh(\nu^*)| + |N_n(\theta^T,\nu^T) - N(\theta^T,\nu^T)| \\
&\leq \left|1 - \frac{2}{\pi}\varphi^T\right|\cdot|\tanh(\nu^*)| + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&< 1.775\frac{2}{\pi}\Theta\cdot|\tanh(\nu^*)| + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&= 1.13\Theta\cdot|\tanh(\nu^*)| + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&= 1.13|\tanh(\nu^*)|\Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}
\end{aligned}$$

In the proof, we use $|N(\theta^T,\nu^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\tanh(\nu^*)| = \left|1 - \frac{2}{\pi}\varphi^T\right|\cdot|\tanh(\nu^*)|$, which is provided in Corollary 4.2.

Particularly, when $\pi^* = \{1,0\}$ or $\{0,1\}$, then $|\tanh(\nu^*)| = 1$

$$\begin{aligned}
|N_n(\theta^T,\nu^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\tanh(\nu^*)| &\leq |N(\theta^T,\nu^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\tanh(\nu^*)| + |N_n(\theta^T,\nu^T) - N(\theta^T,\nu^T)| \\
&< 1.13\Theta\cdot|\tanh(\nu^*)| + 1.13\mathrm{e}\Theta \\
&= 1.13(1+\mathrm{e})|\tanh(\nu^*)|\Theta\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)
\end{aligned}$$

Therefore $\|\pi^{T+1} - \bar{\pi}^*\|_1 = |N_n(\theta^T,\nu^T) - \mathrm{sgn}\langle\theta^T,\theta^*\rangle\tanh(\nu^*)| = \left\|\frac{1}{2} - \pi^*\right\|_1 \cdot \mathcal{O}\left(\sqrt{\frac{d}{n}} \vee \frac{\log\frac{1}{\delta}}{n} \vee \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) + c(\pi^*)\cdot \mathcal{O}\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$, where $c(\pi^*) = 0$ when $\pi^* = \{1,0\}$ or $\{0,1\}$, and $c(\pi^*) = \mathcal{O}(1)$.

□