Proceedings of Machine Learning Research vol 99:1–56, 2019 32nd Annual Conference on Learning Theory

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

Jeongyeol Kwon* KWONCHUNGLI@UTEXAS.EDU
The University of Texas at Austin
Wei Qian* WQ34@CORNELL.EDU
Cornell University
Constantine Caramanis CONSTANTINE@UTEXAS.EDU
The University of Texas at Austin
Yudong Chen YUDONG.CHEN@CORNELL.EDU
Cornell University
Damek Davis DSD95@CORNELL.EDU
Cornell University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

The Expectation-Maximization algorithm is perhaps the most broadly used algorithm for inference of latent variable problems. A theoretical understanding of its performance, however, largely remains lacking. Recent results established that EM enjoys global convergence for Gaussian Mixture Models. For Mixed Linear Regression, however, only local convergence results have been established, and those only for the high SNR regime. We show here that EM converges for mixed linear regression with two components (it is known that it may fail to converge for three or more), and moreover that this convergence holds for random initialization. Our analysis reveals that EM exhibits very different behavior in Mixed Linear Regression from its counterpart in Gaussian Mixture Models, and hence our proofs require the development of several new ideas.[1]

1. Introduction

The expectation-maximization (EM) algorithm is a general-purpose technique for computing the maximum likelihood solution for problems with missing data, often modeled as latent variables (Dempster et al., 1977; Wu, 1983). In general, maximizing the likelihood in the presence of missing data is an intractable problem due to the non-convexity of the log-likelihood function. EM is an iterative procedure that computes successively tighter lower bounds of the log-likelihood function. Despite its simplicity and its widespread use in practice, relatively little is understood about the theoretical properties of EM. Recent results have demonstrated that in the high SNR regime (and under additional regularity assumptions), EM converges locally (e.g., Yi and Caramanis 2015; Balakrishnan et al. 2017; Klusowski et al. 2019; Yi et al. 2014, 2016). For the special case of Gaussian Mixture Models (GMM) with two components, very recent work (Daskalakis et al., 2017) has shown that a two-phase version of EM converges from random initialization. As far as we know,

* These two authors equally contributed to this work.
1. This paper results from a merger of work from two groups who work on the problem at the same time.

no comparable global convergence result is known for Mixed Linear Regression (MLR), despite the empirical success of EM in this problem (Jordan and Jacobs, 1994; De Veaux, 1989).

The lack of global convergence guarantees for EM under MLR is not simply an oversight. Rather, as we show later, MLR exhibits very different behavior from GMM, even on the population (infinite sample) level. Existing techniques used to analyze EM under GMM—often based on ℓ_2 distance contraction—are fundamentally insufficient for establishing global convergence of EM for MLR.

In this work, we show for the first time that EM for MLR with two components converges globally without the need for any special initialization. Moreover, our proof reveals (a bound on) the rate of convergence of EM as a function of how far it is from the true parameter. Locally, we improve upon past results, as these not only required an initialization step and a high SNR assumption, but also failed to provide a tight final error bound that correctly captures the improvement achieved by EM over the initial solution. We explain connections to prior art in more details in Section 1.2.

1.1. Basic Setup and the EM Algorithm

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

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

where $z_i \in \{\pm 1\}$ are the hidden/latent variables, which play the role of labels denoting whether a data point $(\boldsymbol{x}_i, y_i)$ is generated by $+\boldsymbol{\beta}^*$ or $-\boldsymbol{\beta}^*$. Finding the true parameter $\boldsymbol{\beta}^*$ is known to be NP-hard in general (Yi et al., 2014) even without noise. Accordingly, a common assumption in the literature stipulates that the covariates and noise terms, $\boldsymbol{x}_i$ and e_i, are sampled independently from Gaussian distributions, that is, $\boldsymbol{x}_i \sim \mathcal{N}(0, I_d)$ and $e_i \sim \mathcal{N}(0, \sigma^2)$, where σ is known. We assume, moreover, that the hidden variables z_i take values ± 1 with equal probability and are independent of everything else.

At each iteration, the EM algorithm performs two steps: the E-step that computes the expectation of the log likelihood function conditioned on the current estimate of $\boldsymbol{\beta}^*$, and the M-step that maximizes this expectation. For MLR, when we plug in the likelihood of the assumed Gaussian distribution and replace the expectation with an empirical average over observed data $\{\boldsymbol{x}_i, y_i\}$, the M-step becomes the familiar (weighted) least squared loss minimization problem. In this case, the sample-based EM update with the current estimator $\boldsymbol{\beta}$ has the following closed form expression (for a derivation see Balakrishnan et al. (2017); Klusowski et al. (2019)):

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

In the setting where the covariates $\{\boldsymbol{x}_i\}$ have identity covariance (or have been normalized to so), it is also interesting to consider the following simplified version of EM (we call it "Easy-EM") that replaces the matrix $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ with its expectation:

$$\text{(Easy-EM)} \qquad \tilde{\boldsymbol{\beta}}'' = \frac{1}{n}\sum_{i=1}^{n} \tanh\left(\frac{\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle}{\sigma^2} y_i\right) y_i \boldsymbol{x}_i. \tag{3}$$

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

1.2. Related Work and Main Contributions

As mentioned above, our knowledge of when EM converges to a true solution is still limited. In general, it is known that the EM algorithm may settle in a bad local optimum unless it starts from a well initialized point (Wu, 1983). Recent progress on the theoretical understanding of EM has been made in Balakrishnan et al. (2017), which proposed a novel framework to analyze the EM algorithm. Motivated by this work, there have been some positive results for two related problems: GMMs and MLRs (Daskalakis et al., 2017; Xu et al., 2016; Yi and Caramanis, 2015; Balakrishnan et al., 2017; Klusowski et al., 2019; Yi et al., 2014, 2016).

For GMM with two components, Daskalakis et al. (2017); Xu et al. (2016) provide a global analysis of EM for the mixture of two Gaussians and deliver results that guarantee convergence of EM for this specific problem from a random initialization. For GMM with more components, however, it is known that EM does not converge globally (Jin et al., 2016).

For MLR with two components, only the local convergence of EM has been recently established: it is known that the EM algorithm does converge to the global optimum if we start from a point sufficiently close to the true parameter (Yi and Caramanis, 2015; Balakrishnan et al., 2017; Yi et al., 2014, 2016). A better local contraction region was suggested in Klusowski et al. (2019), where the convergence is guaranteed inside a region where the angle formed by the initialization with the true parameter is small. Still, all known results remain inherently local for MLR, and in particular, are not satisfied by a random initialization, even when a norm bound on the true parameter is known.

MLR is an interesting problem by itself, for which many algorithms have been proposed. The work in Chen et al. (2014) developed a lifted convex formulation approach that achieves tight minimax error rates. A good initialization strategy for EM based on Stein's second-order lemma was proposed in Yi et al. (2014), though this seems to rely on the noiseless setting which they study. The above two papers have focused on the mixture of two components case. Recent work has extended the focus to more components. Work in Li and Liang (2018); Zhong et al. (2016) develops gradient descent based algorithms. In parallel, the work in Yi et al. (2016); Chaganty and Liang (2013); Sedghi et al. (2016) utilizes tensor decomposition of third order moments.

The question of whether EM converges from a random initialization for MLR with two components still remains open. Our main contribution is to resolve this question affirmatively.

Main Contributions. We prove the global convergence of the EM algorithm, i.e., it converges with probability one from a random initialization. We first establish this result for the infinite sample limit, i.e., the population EM, by analyzing its trajector along the landscape of the likelihood function. We then couple the finite-sample EM with the population EM, thereby providing a finite sample analysis. This coupling idea is inspired by Balakrishnan et al. (2017), but our strategy and most of the technical details differ. In particular, we control not only the ℓ_2 distance between the population and finite-sample EM, but also, crucially, the *angle* between them. As we comment on in greater details below, this coupling is strong enough to guarantee convergence even when the current EM iterate is far from the desired solution; moreover, it yields near-optimal sample complexity bounds that improve upon the results in Balakrishnan et al. (2017), particularly in terms of the dependence on the signal-to-noise ratio.

1.3. A Roadmap and Proof Outline

We provide a brief outline of the main steps of the paper.

Analysis of the population EM:

- **Landscape**. As mentioned, previous work on analyzing the EM algorithm for MLR relies on demonstrating that the ℓ_2 distance between the current iterate and the true solution $\boldsymbol{\beta}^*$, contracts at every iteration provided that the initial distance is already small. Such a contraction, however, cannot hold globally, as the EM update initialized randomly may in fact result in a *larger distance* from $\boldsymbol{\beta}^*$. This phenomenon was pointed out in Klusowski et al. (2019). We provide a geometric explanation in this paper by showing the existence of saddle points of the log-likelihood function in the direction orthogonal to $\boldsymbol{\beta}^*$. These saddle points prevent a global convergence in ℓ_2 distance of EM (which is equivalent to gradient ascent). On the other hand, we show that $\pm\boldsymbol{\beta}^*$ are the only local maxima, hence suggesting that global convergence can be proved by other means.

- **Decreasing Angle**. Instead of proving a global convergence via the ℓ_2 distance, we show that the angle between the iterate and $\boldsymbol{\beta}^*$ is always decreasing (unless we start from an exactly orthogonal vector—a measure zero event). Consequently, EM quickly enters a local region where the current iterate is well aligned with the direction of $\boldsymbol{\beta}^*$. In this local region, we show that a contraction in distance indeed holds.

- **Escaping Nearly Orthogonal Region**. Random initialization in a d-dimensional space typically yields a vector whose correlation with $\boldsymbol{\beta}^*$ is $O(1/\sqrt{d})$. In this region, the contracting behavior of the angle can be very subtle. We provide a fine-grained analysis in this region, showing that first the cosine of the angle increases geometrically, and then the sine of the angle decreases geometrically afterwards. Consequently, in a logarithmic number of steps EM escapes this nearly orthogonal region and attains a constant correlation with $\boldsymbol{\beta}^*$.

- **Low SNR**. Besides being local in nature, previous results are dependent on the high SNR assumption (Yi et al., 2014; Balakrishnan et al., 2017; Yi and Caramanis, 2015; Wu et al., 2016); that is, the standard deviation, σ, of the additive noise, is sufficiently smaller than the norm of the true parameter. Our analysis is applicable in both low and high SNR regimes, and reveals an explicit convergence rate as a function of the noise level.

Analysis of the Finite Sample EM:

- **Coupling in Angle**. We analyze the finite-sample EM update by coupling it with the population EM. Balakrishnan et al. (2017) provided a bound between these two updates in l_2 distance. Since our argument is based on contraction of angle, we need to establish additional concentration inequalities in order to bound the cosine and sine of the angle. We then conclude that, starting from a random initial guess in d-dimensional space, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$ fresh samples in each iteration, the finite-sample EM yields an estimate with an l_2 error bounded by $O(\epsilon)$ after $T = O(\max(1, \eta^{-2}) \max(\log d, \log(1/\epsilon)))$ iterations. η here is the notation for signal-to-noise ratio (SNR).

- **Statistical Error**. In the high SNR regime, we further refine the finite sample analysis and show that the EM algorithm in fact achieves an error of $\tilde{O}\left(\sigma\sqrt{d/n}\right)$. Note that the error rate is independent of the signal strength $\|\boldsymbol{\beta}^*\|$. This is in a stark contrast to all the previous analysis of EM which proved an error of $\tilde{O}\left(\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\sqrt{d/n}\right)$ for MLR (Balakrishnan

et al., 2017; Klusowski et al., 2019) —such an error bound is no better than the bound achieved by a simple spectral initialization and in particular cannot guarantee exact recovery in the noiseless setting.

- **Analysis of Easy-EM**. For the early iterations of EM where the cosine between the estimate (or a random initialized point) and $\boldsymbol{\beta}^*$ can be as small as $1/\sqrt{d}$, we can instead run Easy-EM, which does not need the computation of the inverse of the sample covariance. Easy-EM also provides a guarantee for reaching an angle larger than $O(1)$, while in our analysis, standard EM requires an additional condition that the statistical fluctuation, due to the size finite samples, should be less than $O(1/\sqrt{d})$. Therefore our results indicate that one can run Easy-EM until the cosine of the angle between the current estimate and the true parameter is large, and subsequently run EM.

Paper Organization In Section 2, we derive a closed form equation of the population EM and prove some of its structural properties. Section 3 is devoted to summarize our results on the global convergence of the population EM. The analysis on the finite-sample EM is provided in Section 4. All technical proofs that are not given in the main paper are deferred to the Appendix.

2. The Population EM Update

In this section, we consider the infinite-sample limit of the EM update (i.e., the population EM) and discuss its basic properties. This discussion highlights the main challenges in the MLR problem and the reasons why they can be resolved. It also serves as a starting point of our subsequent proof for global convergence.

2.1. Basic Notation

We use $\angle(\boldsymbol{u}, \boldsymbol{v})$ to denote the angle between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. The norm operator $\|\cdot\|$ without subscript is taken as the l_2 norm for a vector or the operator norm for a matrix. $\langle \cdot, \cdot \rangle$ denotes the usual inner product: $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$.

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

σ is the standard deviation of the noise e and assumed to be known. We define the signal-to-noise ratio (SNR) of the problem as $\eta := \frac{\|\boldsymbol{\beta}^*\|}{\sigma}$.

2.2. An Explicit Expression for the Population EM Update

As in Balakrishnan et al. (2017), we consider the following population EM update

$$\boldsymbol{\beta}_{t+1} = \mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[\left(\mathbb{E}_{Y|X\sim\mathcal{N}(\langle X,\boldsymbol{\beta}^*\rangle,\sigma^2)}\left[\tanh\left(\frac{\langle X, \boldsymbol{\beta}_t\rangle}{\sigma^2}Y\right)Y\right]\right)X\right]. \tag{4}$$

The above expression follows from taking the limit $n \to \infty$ in the EM update formula (2) and simplifying the result using the symmetry of the distribution of Y given X. We refer to Balakrishnan et al. (2017) for the details of this standard derivation.

We focus on one iteration of the population EM which yields the next iterate $\boldsymbol{\beta}'$. It is convenient to change the basis by choosing $\boldsymbol{v}_1 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$ in the direction of the current iterate and $\boldsymbol{v}_2$ to be the orthogonal complement of $\boldsymbol{v}_1$ in $\mathrm{span}\{\boldsymbol{\beta}, \boldsymbol{\beta}^*\}$. We expand them to an orthonormal basis $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_d\}$ in $\mathbb{R}^d$. Introduce the shorthand $b_1 := \langle \boldsymbol{\beta}, \boldsymbol{v}_1 \rangle = \|\boldsymbol{\beta}\|$, $b_1^* := \langle \boldsymbol{\beta}^*, \boldsymbol{v}_1 \rangle$ and $b_2^* := \langle \boldsymbol{\beta}^*, \boldsymbol{v}_2 \rangle$. Using the spherical symmetry of the distribution of X, we may write the next iterate $\boldsymbol{\beta}'$ as

$$\boldsymbol{\beta}' = \mathbb{E}_{\alpha_i}\left[\mathbb{E}_{y|\alpha_i}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}y\right)y\right]\sum_i \alpha_i \boldsymbol{v}_i\right], \tag{5}$$

where the expectation is taken over $\alpha_i \sim \mathcal{N}(0,1)$ and $y|\alpha_i \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)$. Without loss of generality, we assume $b_1, b_1^*, b_2^* \geq 0$. The lemma below plays a key role in our later development. It provides an explicit expression of $\boldsymbol{\beta}'$ in terms of the above basis system, which, among other things, implies that $\boldsymbol{\beta}' \in \mathrm{span}\,(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ (and hence $\boldsymbol{\beta}_t \in \mathrm{span}\,(\boldsymbol{\beta}_0, \boldsymbol{\beta}^*)$ for all t).

Lemma 1 *Define $\sigma_2^2 := \sigma^2 + b_2^{*2}$. We can write $\boldsymbol{\beta}' = b_1'\boldsymbol{v}_1 + b_2'\boldsymbol{v}_2$, where b_1' and b_2' satisfy*

$$b_1' = b_1^* S + R, \quad \text{and} \quad b_2' = b_2^* S, \tag{6}$$

where $S \geq 0$ and $R > 0$ are given explicitly in (19) in Appendix 1.1. Moreover, $S = 0$ iff $b_1^* = 0$.

2.3. Structural Properties of the Population EM

Note that the quantities b_1' and b_2' in Lemma 1 represent the projections of $\boldsymbol{\beta}'$ in and orthogonal to the direction of $\boldsymbol{\beta}$. From the expression of b_2', we immediately deduce the following structural property of the population EM update:

1. **Decreasing angle**: $\boldsymbol{\beta}'$ forms a smaller angle with $\boldsymbol{\beta}^*$ compared to $\boldsymbol{\beta}$. To see this, note that $0 \leq \tan\angle(\boldsymbol{\beta}', \boldsymbol{\beta}) = \frac{b_2'}{b_1'} \leq \frac{b_2^*}{b_1^*} = \tan\angle(\boldsymbol{\beta}^*, \boldsymbol{\beta})$. When $\frac{b_2'}{b_1'} > 0$, the angle strictly decreases; when $\frac{b_2'}{b_1'} = 0$, the angle remains the same. In particular, $\frac{b_2'}{b_1'} = 0$ holds iff $b_2' = 0$, that is, either $b_2^* = 0$ (i.e., $\boldsymbol{\beta} \in \mathrm{span}(\boldsymbol{\beta}^*)$) or $S = 0$ (i.e., $\boldsymbol{\beta} \perp \boldsymbol{\beta}^*$).

From the expression of b_1', we deduce the following (cf. Lemma 5):

2. **Contraction along $\boldsymbol{\beta}$**: In the direction of $\boldsymbol{v}_1$ (equivalently, $\boldsymbol{\beta}$), $\boldsymbol{\beta}'$ moves towards a unique fixed point $E(\boldsymbol{v}_1)$; i.e., $|b_1' - E(\boldsymbol{v}_1)| \leq |b_1 - E(\boldsymbol{v}_1)|$ with equality holds iff $b_1 = E(\boldsymbol{v}_1)$.

It is also easy to see that the iterates remain bounded: $\|\boldsymbol{\beta}'\| \leq 3\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}$ (cf. Lemma 22).

Interestingly, it can be shown that the population EM update is equivalent to applying *gradient ascent* with a fixed step size to the population log likelihood function of MLR. Building on the above structural properties, we obtain the following complete characterization of the fixed points of the population EM as well as the stationary points of the population log likelihood.

Theorem 1 (Population EM and Log-likelihood) *For each nonzero $\boldsymbol{\beta}$ not parallel to $\boldsymbol{\beta}^*$, in span$(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, the set of fixed points of the population EM is equal to the set of stationary points of the log-likelihood. This set contains exactly five elements: (i) $\boldsymbol{\beta}^*$ and $-\boldsymbol{\beta}^*$, which are global maxima; (ii) $\mathbf{0}$, which is a local minimum; (iii) $E(\boldsymbol{v})\boldsymbol{v}$ and $-E(\boldsymbol{v})\boldsymbol{v}$, where $\boldsymbol{v} \perp \boldsymbol{\beta}_*$. Moreover, a stationary point in the orthogonal space is a saddle point whose Hessian has a strictly positive eigenvalue.*

As $\pm\boldsymbol{\beta}^*$ are the only local maxima, it becomes less surprising that the population EM (equivalent to gradient ascent) converges to them from a random initialization. On the other hand, with the existence of saddle points, it is easy to see that the ℓ_2 distance to $\boldsymbol{\beta}^*$ *cannot* contract globally; that is, $\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| > \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$ for some $\boldsymbol{\beta}$. Note that GMM does not have such saddle points, and the ℓ_2 distance does decrease globally as is established in a previous work (Daskalakis et al., 2017).

3. Main Results on the Population EM

In this section, we provide our main results on the global convergence of the population EM. We adopt a new strategy for the convergence analysis to get around the aforementioned challenge based on the contraction of the ℓ_2 distance. We first prove a rapid decrease in angle and then show a geometric decrease in distance. The convergence result in three phases is summarized below:

1. **Increasing Cosine:** Starting from a randomly initialized vector in $\mathbb{R}^d$, after $O(\max(1, \eta^{-2}) \log d)$ iterations, EM outputs a vector whose angle with $\boldsymbol{\beta}^*$ is less than $\pi/3$.
2. **Decreasing Sine:** Starting from a vector whose angle with $\boldsymbol{\beta}^*$ is less than $\pi/3$, after $O(\max(1, \eta^{-2}))$ iterations, EM outputs a vector whose angle with $\boldsymbol{\beta}^*$ is less than $\pi/8$.
3. **Convergence in ℓ_2:** Starting from a vector whose angle with $\boldsymbol{\beta}^*$ is less than $\pi/8$, after $O(\max(1, \eta^{-2}) \log(1/\epsilon))$ iterations, EM outputs an estimate of $\boldsymbol{\beta}^*$ whose ℓ_2 error is $O(\epsilon)$.

All the above results hold for an arbitrary SNR, thus improving on previous results that are only established in the high SNR regime.

3.1. Convergence of Cosine

Recall that θ_0, θ and θ' denote the angles that $\boldsymbol{\beta}^*$ forms with $\boldsymbol{\beta}_0$ (initial iterate), $\boldsymbol{\beta}$ (current iterate), and $\boldsymbol{\beta}'$ (next iterate), respectively. By symmetry we may assume w.l.o.g. that $\cos\theta_0$ is positive. Note that $\cos\theta_0 = \Theta(1/\sqrt{d})$ with high probability. For the early stage of iterations, we focus on the cosine of the angle and show that it increases geometrically. Therefore, starting from $\cos\theta_0 = \Theta(1/\sqrt{d})$, a logarithmic number of iterations of EM is sufficient to guarantee $\cos\theta_t = O(1)$.

Theorem 2 (Cosine Convergence) *As long as $\frac{\pi}{2} > \theta \geq \frac{\pi}{3}$, each population EM iteration satisfies*

$$\cos\theta' \geq \kappa \cos\theta, \tag{7}$$

where $\kappa = \sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}} > 1$. Consequently, if $\cos\theta_0 = \Theta(1/\sqrt{d})$, after $T = O(\log(d)\max(1, \eta^{-2}))$ iterations, we get $\theta_T < \pi/3$ or $\cos\theta_T \geq \frac{1}{2}$.

The proof is in Appendix 2.2. From the proof, it shows that $\cos(\theta') \geq \cos(\theta)\sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}}$. However, the ratio between $\cos\theta'$ and $\cos\theta$ approaches 1 as θ goes to 0. In other words, cosine angles are not informative for establishing a constant convergence factor bounded away from 1 when θ is small. In the following subsection, we state a similar result for sine of the angle to complement this result.

3.2. Convergence of Sine

We next show that the sine of the angle converges geometrically to 0. This is reminiscent of the proof for Theorem 3 in Xu et al. (2016), where they used a similar logic to show *asymptotic* convergence. Here we provide an explicit rate of convergence by quantifying the amount of increase in sine, which is critical in order to port the population-level results to the finite sample setting.

Theorem 3 (Sine Convergence) *As long as $0 \le \theta < \frac{\pi}{2}$, each population EM iteration satisfies*

$$\sin\theta' \le \kappa \sin\theta, \tag{8}$$

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

It is proved in Appendix 2.1. Note that the speed of convergence increases as the angle decreases. This result is most useful when the angle is bounded away from $\pi/2$—complementary to the case covered by Theorem 2. We also remark that in a high SNR regime ($\eta \gg 1$), κ can be much smaller than 1 (depending on the initial angle); in a low SNR ($\eta \ll 1$) regime, however, the convergence rate cannot be faster than $1 - O(\eta^2)$, regardless of the initial angle.

3.3. Convergence of Distance

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

Theorem 4 (ℓ_2 Convergence) *Assume that $\theta < \pi/8$, and define $\sigma_2^2 = \sigma^2 + b_2^{*2}$. If $b_2^* < \sigma$ or $\frac{\sigma_2^2}{\sigma^2} b_1 < b_1^*$, then we have*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le \kappa \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \kappa(16\sin^3\theta)\|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}, \tag{9a}$$

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

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \le 0.6 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{9b}$$

In order to give a geometrically decaying error bound, we have an additional term in (9a) that depends on angle and SNR. When b_1 is close to b_1^* and σ is small, we get a better contraction (9b). The detailed proof is in Appendix 2.3.

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

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

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| < \kappa^T \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + T\kappa^T \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}. \tag{10}$$

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

$$0.6, \ \sqrt{\left(1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}\right)^{-1}}, \ \sqrt{1 - \frac{0.8\eta^2}{1+\eta^2}}. \tag{11}$$

The convergence rate depends on the SNR η and the norm of an initial guess. For different η's, the rate is either a constant or $1 - O(\eta^2)$, as was in the case of sine. Therefore, $T = O(\max(1, \eta^{-2}) \log(1/\epsilon))$ iterations is sufficient to achieve an ϵ-optimal solution. In the Appendix, we show that the convergence rate only becomes faster as the algorithm proceeds.

4. Finite Sample Analysis

We now turn to prove the convergence of the finite-sample EM given in Eq. (2). Along the way, we also prove the convergence of the Easy-EM algorithm given in Eq. (3). As we discuss in length below, Easy-EM is not only interesting on its own, but also useful in the setting where the "statistical fluctuation" $\epsilon_f \propto \sqrt{d/n}$ between the population and the finite-sample EM updates—which is determined by the sample size n—is $O(1)$ rather than $O(1/\sqrt{d})$.

In this section, we use $\boldsymbol{\beta}$ to denote our current iterate, $\boldsymbol{\beta}'$ for the output from one step of the *population* EM, and $\tilde{\boldsymbol{\beta}}'$ for the output from one step of the *finite-sample* EM. Accordingly, $\tilde{\theta}'$ denotes the angle between $\tilde{\boldsymbol{\beta}}'$ and $\boldsymbol{\beta}^*$. When we consider the sequence of iterates generated by the finite-sample EM, we use $\tilde{\boldsymbol{\beta}}_t$ for the t^{th} iterate and $\tilde{\theta}_t$ for its angle with $\boldsymbol{\beta}^*$. Similarly to the population EM discussed in previous section, we will show that the finite-sample EM converges in several phases:

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

2. **Increasing Cosine:** Starting from the vector obtained from the last step, run the finite-sample EM for $O(\min(\log d, \log(1/\epsilon_f)) \max(1, \eta^{-2}))$ iterations to get $\cos\tilde{\theta}_t \geq 1/2$.

3. **Decreasing Sine:** Starting from a vector with cosine of its angle $\cos\tilde{\theta}_0$ at least $1/2$, run the finite-sample EM for $O(\max(1, \eta^{-2}))$ iterations to get $\sin\tilde{\theta}_t \leq \sin(\pi/70)$.

4. **Convergence in ℓ_2:** Starting from $\tilde{\theta}_0 \leq \pi/70$, run the finite-sample EM for $O(\max(1, \eta^{-2}) \log(1/\epsilon))$ iterations to get $\|\tilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\| \leq O(\epsilon)$.

Collecting all the steps, we obtain the following overall guarantee:

Theorem 5 (The Finite Sample EM) *Suppose we start from an initial vector in $\mathbb{R}^d$ whose correlation with $\boldsymbol{\beta}^*$ is at least $\Omega\big(1/\sqrt{d}\big)$, with ℓ_2 norm at least $\|\boldsymbol{\beta}^*\|/10$. We run the sample-splitting finite-sample EM with $O(\max(1, poly(\eta^{-1}))\ (d/\epsilon^2) \log(T/\delta))$ fresh samples in each iteration. After $T = O(\max(1, \eta^{-2}) \max(\log d, \log(1/\epsilon)))$ iterations, we get*

$$\mathbb{P}(\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq \epsilon) \geq 1 - \delta.$$

4.1. Analysis of the Finite-Sample EM

We now provide the details for the four-phase convergence outlined above. We consider sample-splitting as an analysis technique, as it renders subsequent iterations of the EM algorithm independent. As with many other papers that have used this analysis technique, we believe it is an artifact of the analysis, but we are unable to find a way to remove it.

As discussed in the introduction, our approach is to couple the finite sample EM to the population EM. Work in Balakrishnan et al. (2017) establishes a bound on the ℓ_2 distance between the population

and the finite-sample EM in the form of $\|\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| = O\Big(\sqrt{d/n}\Big)$. This type of bound implies local contraction in distance; however it is not sufficient for us, as we need to control the angle outside of the local contraction region. Here we prove a more fine-grained result of the form $|\langle \tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle| = O\left(1/\sqrt{n} + d/n\right)$ (cf. Theorem 13 in Appendix 5.3). This allows us to show that the finite-sample EM decreases the angle up to a statistical fluctuation per iteration.

Theorem 6 *Suppose that $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$. Then, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon_f^2)$ samples for one finite-sample based EM iteration, we have*

$$\cos\tilde{\theta}' \geq \kappa(1 - 10\epsilon_f)\cos\theta - O\bigg(\max\Big(\frac{\epsilon_f}{\sqrt{d}}, \epsilon_f^2\Big)\bigg), \tag{12a}$$

$$\sin^2\tilde{\theta}' \leq \kappa' \sin^2\theta + O(\epsilon_f), \tag{12b}$$

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

The theorem implies that the cosine and sine of the angle improves, up to a quantity that depends on $\epsilon_f \propto \sqrt{d/n}$ (and hence on the sample size). We note the extra factor ϵ_f^2 in the bound. Technically, this arises from controlling the random fluctuation of the inverse sample covariance matrix $(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}$. We provide two sufficient conditions under which this term is negligible: (i) ϵ_f is small enough, namely, $< 1/\sqrt{d}$, or (ii) $\langle \boldsymbol{\beta}_0, \boldsymbol{\beta}^* \rangle > \epsilon_f$, in other words, the initialization is good (cf. proofs of Theorem 13 and Corollary 6 in the Appendix). In Section 4.2 we show that Easy-EM exhibits a very similar convergence behavior, without the appearance of the ϵ_f^2 term. Therefore, if $\epsilon_f > 1/\sqrt{d}$, one can simply run Easy-EM until the estimate has enough correlation with $\boldsymbol{\beta}^*$ (i.e., $\langle \boldsymbol{\beta}_t, \boldsymbol{\beta}^* \rangle > \epsilon_f$), and then switch to EM.

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

With the per-iteration bounds in (12a) and (12b), we can bound the angle after T steps of the finite-sample EM and thereby guarantee achieving a final error of ϵ. It will become clear that

$$\epsilon = \epsilon_f \max(1, \eta^{-2}) \tag{13}$$

(cf. Proof for Lemma 2, Lemma 3 and Lemma 7) since the final error has an accumulation of statistical fluctuations (quantified by ϵ_f) from $T = \tilde{O}(\max(1, \eta^{-2}))$ iterations.

Lemma 2 (Finite-Sample Cosine Convergence) *Assume $\|\tilde{\boldsymbol{\beta}}_0\| \geq \|\boldsymbol{\beta}^*\|/10$. Take $\epsilon_f > 0$ small enough to ensure $\kappa = (1 - 10\epsilon_f)\sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}} > 1$. We run the sample-splitting finite-sample EM, each step with $n/T = \tilde{O}\big(\max(1, \eta^{-2})d/\epsilon_f^2\big)$ samples and $T = O\big(\max(1, \eta^{-2})\log d\big)$ iterations. As long as $\tilde{\theta}_t > \pi/3$ for all $t \leq T$, we have with high probability*

$$\cos\tilde{\theta}_T \geq \kappa^T \cos\tilde{\theta}_0 - \frac{\kappa^T - 1}{\kappa - 1} O\Big(\frac{\epsilon_f}{\sqrt{d}}\Big). \tag{14}$$

In particular, when $\cos\tilde{\theta}_0 = \Theta\big(1/\sqrt{d}\big)$, we get $\cos\tilde{\theta}_T \geq \frac{1}{2} - O(\epsilon)$.

Lemma 3 (Finite-Sample Sine Convergence) *Suppose we get a $\tilde{\boldsymbol{\beta}}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/3$ from the previous phase. We run the sample-splitting sample-based EM, each step with $n/T = \tilde{O}(\max(1, \eta^{-2})d/\epsilon_f^2)$ samples. Then with high probability and a constant $\kappa = \left(\sqrt{1 + \frac{0.5\eta^2}{1+\eta^2}}\right)^{-1} < 1$, we have*

$$\sin^2 \tilde{\theta}_T \leq \kappa^{2T} \sin^2 \tilde{\theta}_0 + \frac{1}{1-\kappa^2} O(\epsilon_f). \tag{15}$$

After $T = O(\max(1, \eta^{-2}))$ iterations, we get $\sin^2 \tilde{\theta}_T \leq \sin^2 \frac{\pi}{70} + O(\epsilon)$.

Remark *We should take small ϵ_f such that $\tilde{\theta}'$ remains less than $\tilde{\theta}$ in each iteration with high probability. A sufficient condition is that $\epsilon_f < O(\min(1, \eta^2))$, which ensures that*

$$\left(1 + \frac{0.5\eta^2}{1+\eta^2}\right)^{-1} \sin^2 \tilde{\theta} + O(\epsilon_f) \leq \sin^2 \tilde{\theta},$$

Finally, suppose we have reached the angle below $\pi/70$. The following theorem, proved in Appendix 4, provides a convergence guarantee in ℓ_2 distance for sample based EM.

Theorem 7 (Finite-Sample Distance Convergence) *Suppose we get a $\tilde{\boldsymbol{\beta}}_0$ whose angle with $\boldsymbol{\beta}^*$ is less than $\frac{\pi}{70}$ from the previous phase. There exist a constant $C > 1$ for which the following holds.*

- *If $\eta < C$, sample-splitting finite-sample EM with $n/T = \tilde{O}(\eta^{-2} d/\epsilon_f^2)$ samples per iteration satisfies*

$$\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq \kappa^T \|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + T\kappa^T \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2} + O(\epsilon)\|\boldsymbol{\beta}^*\|, \tag{16}$$

  where κ is the maximum among (11) as in Corollary 1. After $T = O(\eta^{-2} \log(1/\epsilon))$ iterations, we estimate $\boldsymbol{\beta}^*$ with an ℓ_2 error bounded by $O(\epsilon)$.

- *If $\eta \geq C$, sample-splitting finite-sample EM with $n/T = \tilde{O}(d/\epsilon_f^2)$ samples per iteration satisfies*

$$\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq \kappa^T \|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + O(\epsilon)\sigma, \tag{17}$$

  where $\kappa = 0.95 + \epsilon_f < 1$. After $T = O(\log(1/(\sigma\epsilon)))$ iterations, we estimate $\boldsymbol{\beta}^*$ with an ℓ_2 error bounded by $\sigma O(\epsilon)$.

Note that the results for the low and high SNR cases are different and they actually require different proof techniques. For the low SNR regime, the bound (16) is obtained by coupling the population and the finite-sample EM updates as mentioned before. Since the statistical fluctuation between these two updates scales with $\|\boldsymbol{\beta}^*\| + \sigma$, the final estimation error depends on $\|\boldsymbol{\beta}^*\|$. For the high SNR regime, we in fact take a different approach and directly control $\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^*$ by using the sample covariance matrix $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ to our advantage. The resulting bound (17) scales with σ only and guarantees exact recovery when $\sigma = 0$.

4.2. Analysis of Easy-EM

In the results above we have assumed that the effect of the term ϵ_f^2 is negligible in equation (12a) . We believe this term is simply an artifact of our analysis. This motivates us to consider the Easy-EM update in (3), for which we can eliminate this ϵ_f^2 factor. We do so by proving a better concentration bound $|\langle \tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle| = O\big(1/\sqrt{n}\big)$ for Easy-EM using the fact that Easy-EM does not have the inverse sample covariance matrix, where we recall that $\tilde{\boldsymbol{\beta}}''$ denotes its next iterate. This bound allows us to establish the following theorem, which is a counterpart of Theorem 6 for EM.

Theorem 8 *Suppose that the norm of the current estimator $\|\boldsymbol{\beta}\|$ is larger than $\|\boldsymbol{\beta}^*\|/10$. Then, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon_f^2)$ samples for one Easy-EM iteration, we have*

$$\cos\tilde{\theta}'' \geq \kappa(1 - 10\epsilon_f)\cos\theta - O\left(\frac{\epsilon_f}{\sqrt{d}}\right), \tag{18a}$$

$$\sin^2\tilde{\theta}'' \leq \kappa' \sin^2\theta + O(\epsilon_f), \tag{18b}$$

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

The only difference between Theorems 6 and 8 is that equation (18a) does not has an extra factor ϵ_f^2. Thus, Easy-EM improves the angle in each step even without the assumption $\epsilon_f \leq 1/\sqrt{d}$, and therefore the multi-step bounds in Lemmas 2, 3 and (16) can be identically applied to Easy-EM.

4.3. Discussions

The overall sample complexity to achieve ϵ error is $n = \tilde{O}\big(\max(1, poly(\eta^{-1}))(d/\epsilon^2)\big)$. In the high SNR regime, this is $\tilde{O}(d/\epsilon^2)$, which is the minimax sample complexity up to log factors. Moreover, the final statistical error is $\tilde{O}\big(\sigma\sqrt{d/n}\big)$ which guarantees exact recovery as $\sigma \to 0$. In the low SNR regime, the sample complexity becomes $O(\eta^{-6}d/\epsilon^2)$. This high dependency on SNR arises because the convergence rates of sine and distance are $1 - O(\eta^2)$ in the low SNR regime, and the statistical fluctuation has to be smaller than η^2 in order to guarantee that every iteration improves the angle or distance. It seems to be the nature of EM algorithm as we have seen similarly high dependence on SNR in GMM settings (Daskalakis et al., 2017). Nevertheless, once enough number of samples are given to offset a low SNR, we achieve an statistical error of $\tilde{O}\big(\|\boldsymbol{\beta}^*\|\sqrt{d/n}\big)$.

5. Conclusion

We studied the EM algorithm for a mixture of two linear regression models. In the large sample limit, we showed that EM converges to true parameters globally without any specialized initialization. In finite sample case, we showed that EM enjoys the same convergences behavior, though it may need the aid of Easy-EM in the first few steps. It would be interesting to explore whether we can remove the dependency on Easy-EM. Extensions of this work could be analyzing the performance of EM when the weight of each component is not equal or there are more than two components.

Acknowledgement

J. Kwon and C. Caramanis are partially supported by NSF EECS-1609279, CCF-1302435, and CNS-1704778. W. Qian and Y. Chen are partially supported by NSF grants 1657420 and 1704828.

References

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

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

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

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

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

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

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

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

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 2019.

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

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

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

Martin J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley*, 2015.

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

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

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

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

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

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

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

1. Proofs for Population EM Update

1.1. Proof of Lemma 1

Lemma 1 *Define $\sigma_2^2 := \sigma^2 + b_2^{*2}$. We can write $\boldsymbol{\beta}' = b_1' \boldsymbol{v}_1 + b_2' \boldsymbol{v}_2$, where b_1' and b_2' satisfy*

$$b_1' = b_1^* S + R, \quad \text{and} \quad b_2' = b_2^* S, \tag{6}$$

where $S \geq 0$ and $R > 0$ are given explicitly in (19) in Appendix 1.1. Moreover, $S = 0$ iff $b_1^* = 0$.

$$\begin{aligned} S :=& \mathbb{E}_{\substack{\alpha_1 \sim \mathcal{N}(0,1), \\ y \sim \mathcal{N}(0, \sigma_2^2)}} \left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*) \right) + \frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*) \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*) \right) \right], \\ R :=& (\sigma^2 + \|\boldsymbol{\beta}^*\|^2) \mathbb{E}_{\substack{\alpha_1 \sim \mathcal{N}(0,1), \\ y \sim \mathcal{N}(0, \sigma_2^2)}} \left[\frac{\alpha_1^2 b_1}{\sigma^2} \tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y + \alpha_1 b_1^*) \right) \right]. \end{aligned} \tag{19}$$

For completeness of the proof, we repeat some arguments in the main text. Recall the EM update:

$$\boldsymbol{\beta}' = \mathbb{E}_{X \sim \mathcal{N}(0, I)} \left[\left(\mathbb{E}_{y | X \sim \mathcal{N}(\langle X, \boldsymbol{\beta}^* \rangle, \sigma^2)} \left[\tanh\left(\frac{\langle X, \boldsymbol{\beta} \rangle}{\sigma^2} y \right) y \right] \right) X \right].$$

We first change the basis by choosing $\boldsymbol{v}_1 = \boldsymbol{\beta} / \|\boldsymbol{\beta}\|$, the unit vector in the direction of the current estimator, and $\boldsymbol{v}_2$ to be the orthogonal complement of $\boldsymbol{v}_1$ in $\operatorname{span}\{\boldsymbol{\beta}, \boldsymbol{\beta}^*\}$. We let $\boldsymbol{v}_3, ..., \boldsymbol{v}_d$ be a completion to an orthonormal basis for the full parameter space, $\mathbb{R}^d$, along with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. By the spherical symmetry of the distribution of $\boldsymbol{x}$, we have

$$\boldsymbol{\beta}' = \mathbb{E}_{\alpha_i} \left[\mathbb{E}_{y | \alpha_i} \left[\tanh\left(\frac{b_1 \alpha_1}{\sigma^2} y \right) y \right] \sum_i \alpha_i \boldsymbol{v}_i \right], \tag{20}$$

where the expectation is taken over $\alpha_i \sim \mathcal{N}(0, 1)$, and $y | \alpha_i \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)$, and we defined $b_1 = \langle \boldsymbol{\beta}, \boldsymbol{v}_1 \rangle = \|\boldsymbol{\beta}\|$, $b_1^* = \langle \boldsymbol{\beta}^*, \boldsymbol{v}_1 \rangle$, and $b_2^* = \langle \boldsymbol{\beta}^*, \boldsymbol{v}_2 \rangle$. Without loss of generality, we assume $b_1, b_1^*, b_2^* \geq 0$. The inner expectation over y does not have any dependence on α_i for $i \geq 3$. Therefore, taking expectation over α_i for $i \geq 3$ yields 0, which implies $\boldsymbol{\beta}'$ is also on the plane spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2$. It enables us to rewrite it as $\boldsymbol{\beta}' = b_1' \boldsymbol{v}_1 + b_2' \boldsymbol{v}_2$ where

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

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

where the expectation is similarly taken over $\alpha_i \sim \mathcal{N}(0, 1)$, and $y | \alpha_i \sim \mathcal{N}(\alpha_1 b_1^* + \alpha_2 b_2^*, \sigma^2)$. In the following, we prove that b_1' and b_2' have a simplified representation as in Lemma 1.

Proof We start with second coordinate b_2'. We will occasionally omit variables for expectation when it is clear over which variable the expectation is taken. We can rewrite the equation (21b) as

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

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

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

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

For the first coordinate b_1', we take the similar strategy but we arrange it in a different way. First, we rewrite equation (21a) as

$$b_1' = \mathbb{E}_{\alpha_1\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{y\sim\mathcal{N}(0,\sigma_2^2))}\left[\tanh\left(\frac{b_1\alpha_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)(y+\alpha_1 b_1^*)\right]\alpha_1\right], \tag{22}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\alpha_1^2\right] \\
=&\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right) + \left(\frac{2b_1^* b_1}{\sigma^2}\alpha_1 + \frac{b_1}{\sigma^2}y\right)\alpha_1\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right] \\
=&\mathbb{E}\left[\tanh\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right) + \frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right] \\
&+ b_1^*\mathbb{E}\left[\frac{\alpha_1^2 b_1}{\sigma^2}\tanh'\left(\frac{\alpha_1 b_1}{\sigma^2}(y+\alpha_1 b_1^*)\right)\right].
\end{aligned} \tag{23}$$

On the other hand,

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

where we applied Stein's lemma for y this time. Plugging the above two equations into (22), we get

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

Finally, $R > 0$ since it is the expectation of positive values almost surely over the real line. Lemma 10 in the appendix shows that $S \geq 0$. $S = 0$ iff $b_1 = 0$ or $b_1^* = 0$. Since we only consider the case where $b_1 \neq 0$, the proof is complete. ■

1.2. Proof of Theorem 1

Theorem 1 (Population EM and Log-likelihood) *For each nonzero $\boldsymbol{\beta}$ not parallel to $\boldsymbol{\beta}^*$, in span$(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, the set of fixed points of the population EM is equal to the set of stationary points of the log-likelihood. This set contains exactly five elements: (i) $\boldsymbol{\beta}^*$ and $-\boldsymbol{\beta}^*$, which are global maxima; (ii) $\mathbf{0}$, which is a local minimum; (iii) $E(\boldsymbol{v})\boldsymbol{v}$ and $-E(\boldsymbol{v})\boldsymbol{v}$, where $\boldsymbol{v} \perp \boldsymbol{\beta}_*$. Moreover, a stationary point in the orthogonal space is a saddle point whose Hessian has a strictly positive eigenvalue.*

Proof From the correspondence between the EM update and gradient step as in Lemma 7, it is easy to see that $\boldsymbol{\beta}' = \boldsymbol{\beta}$ iff the current gradient is $\mathbf{0}$. Therefore, the set of fixed points of population EM is equal to the set of stationary points of the log-likelihood.
To characterize the fixed points, there are two key components as follows. First of all, we show in Lemma 4 that every 2 dimensional space that includes $\boldsymbol{\beta}^*$ contains exactly 5 fixed points, $\mathbf{0}$, $\boldsymbol{\beta}^*$, $-\boldsymbol{\beta}^*$ and two other symmetric points in the orthogonal direction to $\boldsymbol{\beta}^*$. Secondly, we need to some understanding of the property about the Hessian of those fixed points. For $\boldsymbol{\beta}^*$ and $-\boldsymbol{\beta}^*$, they are the global maxima as they are the optimal parameters. For $\mathbf{0}$, a simple calculation shows that the Hessian is positive definite, thus, it is a local minima. For any fixed point $\boldsymbol{v}$ in the orthogonal direction to $\boldsymbol{\beta}^*$, we use Stein's lemma to show that in the direction of $\boldsymbol{\beta}^*$, $\langle \boldsymbol{\beta}^*, \mathcal{H}\boldsymbol{\beta}^* \rangle$ is strictly positive, where $\mathcal{H}$ is the Hessian matrix (Proposition 9). On the other hand, we show that any fixed point $\boldsymbol{v}$ is a local maxima in span$(\boldsymbol{v})$. These two facts allow us to conclude that the fixed points in the orthogonal space are indeed saddle points. To illustrate the second point, we utilize two observations: (1) in the first part of the proof, we have demonstrated span $(\boldsymbol{v})$ is an invariant subspace for the EM operator and $\boldsymbol{v}$ is a contracting point; (2) the monotonicity property of the EM algorithm says that the log-likelihood of the EM iterate does not decrease. Therefore, any fixed point in the direction orthogonal to $\boldsymbol{\beta}^*$ is a local maxima in span$(\boldsymbol{v})$. ■

Lemma 4 *For each unit vector $\boldsymbol{v}$ satisfying $\boldsymbol{v} \perp \boldsymbol{\beta}^*$, the population EM starting at $\boldsymbol{\beta} \in$ span$(\boldsymbol{\beta}^*, \boldsymbol{v})$ has exactly five fixed points: 0, $\boldsymbol{\beta}^*$, $-\boldsymbol{\beta}^*$, $E(\boldsymbol{v})\boldsymbol{v}$, and $-E(\boldsymbol{v})\boldsymbol{v}$ for some $E(\boldsymbol{v}) > 0$.*

Proof Let $\boldsymbol{\beta}'$ be the EM update as in the standard notation. When $\boldsymbol{\beta} = 0$, we have $\boldsymbol{\beta}' = 0$ and thus $\boldsymbol{\beta} = 0$ is a fixed point. For the other cases, we will use a few facts established in Lemma 1:

- if $\langle \boldsymbol{\beta}, \boldsymbol{\beta}^* \rangle = 0$ (i.e, $b_1^* = 0$), it follows that $S = 0$ and $b_2' = 0$. In other words, if the current iterate $\boldsymbol{\beta}$ is orthogonal to the ground truth $\boldsymbol{\beta}_*$, the population EM update remains orthogonal to $\boldsymbol{\beta}^*$.
- if $\langle \boldsymbol{\beta}^\perp, \boldsymbol{\beta}^* \rangle = 0$ (i.e, $b_2^* = 0$), it follows that $b_2' = 0$. In other words, if the current iterate $\boldsymbol{\beta}$ is in the direction of $\boldsymbol{\beta}^*$ (or $-\beta_*$), the population EM iterate remains in that direction.
- if $\langle \boldsymbol{\beta}, \boldsymbol{\beta}^* \rangle > 0$ (or $\langle \boldsymbol{\beta}, -\boldsymbol{\beta}^* \rangle > 0$), it follows that $b_2^* > 0, S > 0$, and thus $b_2' > 0$. In other words, if the current iterate has an acute angle with $\boldsymbol{\beta}^*$ (or $-\boldsymbol{\beta}^*$), $\angle(\boldsymbol{\beta}', \boldsymbol{\beta}_*)$ (or $\angle(\boldsymbol{\beta}', \boldsymbol{\beta}_*)$ will strictly decrease and no fixed point can exist in this region.

Therefore, we deduce that the fixed points of the population EM lies either in span$(\boldsymbol{\beta}^*)$ or in the subspace orthogonal to $\boldsymbol{\beta}^*$. They are the invariant subspaces of the population EM operator. In Lemma 5, it is shown that for each unit direction of $\boldsymbol{\beta}$, there exists a unique contraction point. We

thus conclude that if $\boldsymbol{\beta}$ is in the direction of $\boldsymbol{\beta}^*(-\boldsymbol{\beta}^*)$, the fixed point is $\boldsymbol{\beta}^*(-\boldsymbol{\beta}^*)$; if $\boldsymbol{\beta}$ is in the direction of $\boldsymbol{v}(-\boldsymbol{v})$, the fixed point is $E(\boldsymbol{v})(-E(\boldsymbol{v}))$ for some $E(\boldsymbol{v}) > 0$. ■

Lemma 5 *Suppose $\langle \boldsymbol{\beta}, \boldsymbol{\beta}^* \rangle \geq 0$. Let $\boldsymbol{v}_1$ be the unit vector of $\boldsymbol{\beta}$ and b_1' be the notation used in Lemma 1, denoting the the projection of the EM update onto span($\boldsymbol{v}_1$). There exists a unique non-zero number $E(\boldsymbol{v}_1)$ satisfying*

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

Proof When $\boldsymbol{v}_1$ is fixed, b_1' only depends on $\|\boldsymbol{\beta}\|$. We thus use $f(\|\boldsymbol{\beta}\|)$ for b_1' in the following to emphasize it is a function of $\|\boldsymbol{\beta}\|$.

$$\begin{aligned} f(\|\boldsymbol{\beta}\|) :=& \mathbb{E}_{X,y}\left[y\langle X, \boldsymbol{v}_1\rangle \tanh\left(\frac{\|\boldsymbol{\beta}\| y \langle X, \boldsymbol{v}_1\rangle}{\sigma^2} \right) \right] \\ =& \mathbb{E}_{X\sim\mathcal{N}(0,I), y\sim\mathcal{N}(\langle \boldsymbol{\beta}^*, X\rangle, \sigma^2)}\left[y\langle X, \boldsymbol{v}_1\rangle \tanh\left(\frac{\|\boldsymbol{\beta}\| y \langle X, \boldsymbol{v}_1\rangle}{\sigma^2} \right) \right]. \end{aligned}$$

Let us check a few properties of f:

1. f is smooth (obvious).
2. f is strictly increasing and concave. Note that its derivative with respect to $\|\boldsymbol{\beta}\|$

   $$f'(\|\boldsymbol{\beta}\|) = \mathbb{E}_{X\sim\mathcal{N}(0,I), y\sim\mathcal{N}(\langle \boldsymbol{\beta}^*, X\rangle, \sigma^2)}\left[\frac{(y\langle X, \boldsymbol{v}_1\rangle)^2}{\sigma^2} \tanh'\left(\frac{\|\boldsymbol{\beta}\| y \langle X, \boldsymbol{v}_1\rangle}{\sigma^2} \right) \right]$$

   is positive and is strictly decreasing with respect to $\|\boldsymbol{\beta}\|$.
3. $f(0) = 0$ and $f'(0) > 1$ since

   $$\begin{aligned} f'(0) =& \mathbb{E}_{X\sim\mathcal{N}(0,I), y\sim\mathcal{N}(\langle \boldsymbol{\beta}^*, X\rangle, \sigma^2)}\left[\frac{(y\langle X, \boldsymbol{v}_1\rangle)^2}{\sigma^2} \right] \\ =& \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|^2(3\cos^2(\angle(\boldsymbol{\beta}, \boldsymbol{\beta}_*)) + \sin^2(\angle(\boldsymbol{\beta}, \boldsymbol{\beta}_*)))}{\sigma^2}. \end{aligned}$$

4. f is bounded (cf. Lemma 22)

Let $g(\|\boldsymbol{\beta}\|) := f(\|\boldsymbol{\beta}\|) - \|\boldsymbol{\beta}\|$, it is a strictly concave and smooth function from Property 2. Moreover, $g(0) = 0$, $g'(0) > 0$ from Property 3 and $\lim_{\|\boldsymbol{\beta}\|\to\infty} g(\|\boldsymbol{\beta}\|) = -\infty$ from Property 4. Lemma 6 shows that there exists a unique $E(\boldsymbol{v}_1) > 0$ for g such that $g(E(\boldsymbol{v}_1)) = 0$. Moreover when $\|\boldsymbol{\beta}\| < E(\boldsymbol{v}_1)$, $g(\|\boldsymbol{\beta}\|) > 0$ and when $\|\boldsymbol{\beta}\| > E(\boldsymbol{v}_1)$, $g(\|\boldsymbol{\beta}\|) < 0$. Equivalently, it means that

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

■

Lemma 6 *Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a smooth and concave function, with strictly decreasing derivative, satisfying $f(0) = 0, f'(0) > 0$, and $\lim_{x\to\infty} f(x) = -\infty$. Then there exists a unique $t > 0$ such that $f(t) = 0$ and $f'(t) < 0$. Moreover, $f(x) > 0$ if $x \in (0, t)$ and $f(x) < 0$ if $x \in (t, \infty)$.*

Proof Since f has a continuous gradient at 0 with $f'(0) > 0$, there exists $t_1 > 0$ such that $f'(x) > 0$ for all $x \le t_1$. We thus conclude that

$$f(x) > 0 \; \forall x \in (0, t_1]$$

by the Fundamental theorem of Calculus. By the continuity of f and the condition that $\lim_{x\to\infty} f(x) = -\infty$, there exists $t_2 > 0$ such that $f(t_2) < 0$. Rolle's theorem tells us that there exists $t \in (t_1, t_2)$ such that $f(t) = 0$. Since $f(0) = 0$, the mean value theorem tells us that there exists $t_3 \in (0, t)$ such that $f'(t_3) = 0$. By assumption f is strictly decreasing derivative, we have $f'(x) \le 0$ for all $x \ge t_3$ and $f'(x) > 0$ for all $x \in (0, t_3)$. It follows that $f(x)$ is increasing on $(0, t_3)$ and it is decreasing on (t_3, ∞). The statement follows. ■

Lemma 7 (Correspondence between EM and GD) *In the basic set up of the 2MLR problem, let $\mathcal{L}$ denote the log-likelihood for the population 2MLR as follows:*

$$\mathcal{L} = \mathbb{E}_{X,y} \log\Big(\sum_{z\in\{-1,1\}} f(X, y, z; \boldsymbol{\beta})\Big).$$

There is a correspondence between the gradient (with respect to $\boldsymbol{\beta}$) of the log-likelihood and the EM operator:

$$\boldsymbol{\beta}' = \boldsymbol{\beta} + \sigma^2 \nabla_{\boldsymbol{\beta}} \mathcal{L}.$$

Therefore, the set of fixed points of the population EM iterate $\boldsymbol{\beta}'$ is equal to the set of stationary points of the population log-likelihood of $\mathcal{L}$

Proof The log-likelihood function $\mathcal{L}$ of the population MLR with the optimal parameter $\boldsymbol{\beta}^*$ is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) =& \mathbb{E}_{X,y}\left[\log\left(\frac{1}{2}\cdot\Phi\left(y;\langle X,\boldsymbol{\beta}\rangle,\sigma^2\right)+\frac{1}{2}\cdot\Phi\left(y;-\langle X,\boldsymbol{\beta}\rangle,\sigma^2\right)\right)\right]\\ =& \mathbb{E}_{X,y}\left[\log\left(\frac{1}{2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\langle X,\boldsymbol{\beta}\rangle)^2}{2\sigma^2}\right)+\frac{1}{2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y+\langle X,\boldsymbol{\beta}\rangle)^2}{2\sigma^2}\right)\right)\right],\end{aligned} \tag{24}$$

where Φ denotes the pdf for the Gaussian distribution. The gradient of the population log-likelihood functions with respect to $\boldsymbol{\beta}$ has the following expression:

$$\nabla_{\boldsymbol{\beta}}\mathcal{L} = \frac{1}{\sigma^2}\left[-\boldsymbol{\beta} + \mathbb{E}_{X,y}\left[yX\tanh\left(\frac{y\langle X,\boldsymbol{\beta}\rangle}{\sigma^2}\right)\right]\right] = \frac{1}{\sigma^2}(-\boldsymbol{\beta}+\boldsymbol{\beta}'). \tag{25}$$

where the last line follows from (6). ■

Proposition 9 (Positive eigenvalue along $\boldsymbol{\beta}^*$) *In the basic set of the MLR problem, the Hessian matrix of the population log-likelihood is*

$$\mathcal{H} = \frac{1}{\sigma^2}\left(-I + \mathbb{E}_{X,y}\left[\frac{1}{\sigma^2} y^2 X X^T \tanh'\left(\frac{y\langle X, \boldsymbol{\beta}\rangle}{\sigma^2}\right)\right]\right).$$

Moreover, let $\hat{\boldsymbol{\beta}}_*$ be the unit vector in the direction of $\boldsymbol{\beta}_*$. The following holds for every fixed point $\boldsymbol{\beta}$ that is orthogonal to $\boldsymbol{\beta}^*$:

$$\langle \hat{\boldsymbol{\beta}}_*, \mathcal{H}\hat{\boldsymbol{\beta}}_*\rangle \geq \frac{1}{\sigma^2}\frac{\|\boldsymbol{\beta}^*\|^2}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}.$$

Proof Using the correspondence between the population EM update and gradient of the log-likelihood function of MLR,$\mathcal{L}$ (cf. Lemma 7):

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

Hence, the Hessian matrix is:

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

Recall the EM update:

$$\begin{aligned}\boldsymbol{\beta}' =& \mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[\mathbb{E}_{y|X\sim\mathcal{N}(\langle X,\boldsymbol{\beta}^*\rangle,\sigma^2)}\left[yX\tanh\left(\frac{y\langle X,\boldsymbol{\beta}\rangle}{\sigma^2}\right)\right]\right]\\ =& \sigma\mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[\mathbb{E}_{y|X\sim\mathcal{N}(\langle X,\frac{\boldsymbol{\beta}^*}{\sigma}\rangle,1)}\left[yX\tanh\left(y\langle X,\frac{\boldsymbol{\beta}}{\sigma}\rangle\right)\right]\right] \text{ (rescaling)}. \end{aligned} \tag{26}$$

The gradient with respect to $\boldsymbol{\beta}$ is:

$$\begin{aligned}\nabla_{\boldsymbol{\beta}'}\boldsymbol{\beta} =& \mathbb{E}_{X,y}\left[\frac{1}{\sigma^2}y^2 X X^T \tanh'\left(\frac{y\langle X,\boldsymbol{\beta}\rangle}{\sigma^2}\right)\right]\\ =& \mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[\mathbb{E}_{y|X\sim\mathcal{N}(\langle X,\frac{\boldsymbol{\beta}^*}{\sigma}\rangle,1)}\left[y^2 X X^\top \tanh'\left(y\langle X,\frac{\boldsymbol{\beta}}{\sigma}\rangle\right)\right]\right] \text{ (rescaling)}. \end{aligned} \tag{27}$$

The first part of the claim is proved. For the second part of the claim, it suffices to prove the case for $\sigma = 1$ due to the equivalent representation by rescaling in (26) and (27). If we can show the following relation

$$\langle \hat{\boldsymbol{\beta}}_*, \mathcal{H}\hat{\boldsymbol{\beta}}_*\rangle \geq \frac{\|\boldsymbol{\beta}_*\|^2}{\|\boldsymbol{\beta}_*\|^2 + 1} \tag{28}$$

holds when $\sigma = 1$, we can easily conclude that for general σ,

$$\langle \hat{\boldsymbol{\beta}}_*, \mathcal{H}\hat{\boldsymbol{\beta}}_*\rangle \geq \frac{1}{\sigma^2}\frac{\|\boldsymbol{\beta}_*\|^2}{\|\boldsymbol{\beta}_*\|^2 + \sigma^2}.$$

In the following, our effort is devote to proving (28) assuming $\sigma = 1$. The EM update is now simplified to the following:

$$\boldsymbol{\beta}' = \mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[\mathbb{E}_{y|X\sim\mathcal{N}(\langle X,\boldsymbol{\beta}^*\rangle,1)}\left[yX\tanh\left(y\langle X,\boldsymbol{\beta}\rangle\right)\right]\right].$$

As before, we use the following orthonormal basis with $\boldsymbol{v}_1 = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{v}_2 = \hat{\boldsymbol{\beta}}^*$, where $\hat{\boldsymbol{\beta}}$ is the unit vector of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}^*$ is the unit vector of $\boldsymbol{\beta}^*$. Note that $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$ because we assume $\boldsymbol{\beta} \perp \boldsymbol{\beta}_*$. Since $\boldsymbol{\beta}$ is a fixed point, it follows that $\boldsymbol{\beta} = \mathbb{E}_{X,y} yX \tanh(y\langle X, \boldsymbol{\beta}\rangle)$. A necessary condition is:

$$b_1 = \|\boldsymbol{\beta}\| = \mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon\right) \alpha_1 \tanh\left(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1\right) = b_1', \tag{29}$$

where we integrate over $\alpha_1 \sim \mathcal{N}(0,1)$, $\alpha_2 \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,1)$. Using Stein's Lemma for b_1' with respect to α_1, we have

$$\begin{aligned}
&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1 \tanh\left(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1\right)\right] \\
=&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \nabla_{\alpha_1} \left[\left(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon\right) \tanh(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1)\right] \\
=&\|\boldsymbol{\beta}\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\left(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon\right)^2 \tanh'(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1)\right].
\end{aligned} \tag{30}$$

We obtain a first relation by substituting (30) back to (29):

$$1 = \mathbb{E}_{\alpha_2,\alpha_1,\epsilon}[(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh'(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1)]. \tag{31}$$

Note that we can write $\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon = \sqrt{1 + \|\boldsymbol{\beta}^*\|^2} Z$ for $Z \sim \mathcal{N}(0,1)$, where the equality holds in the distribution sense. We can apply Stein's Lemma for b_1' again with respect to Z:

$$\begin{aligned}
&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1 \tanh(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1) \\
=&\mathbb{E}_{\alpha_2,Z} \sqrt{1 + \|\boldsymbol{\beta}^*\|^2} Z\alpha_1 \tanh(\|\boldsymbol{\beta}\| \sqrt{1 + \|\boldsymbol{\beta}^*\|^2} Z\alpha_1) \\
=&\sqrt{1 + \|\boldsymbol{\beta}^*\|^2}\mathbb{E}_{\alpha_2,Z} \nabla_Z [\alpha_1 \tanh(\|\boldsymbol{\beta}\| \sqrt{1 + \|\boldsymbol{\beta}^*\|^2} Z\alpha_1)] \\
=&\|\boldsymbol{\beta}\|(1 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}_{\alpha_2,Z}[\alpha_1^2 \tanh'(\|\boldsymbol{\beta}\| \sqrt{1 + \|\boldsymbol{\beta}^*\|^2} Z\alpha_1)] \\
=&\|\boldsymbol{\beta}\|(1 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}[\alpha_1^2 \tanh'(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1)].
\end{aligned} \tag{32}$$

We thus obtain a second relation by substituting (32) back to (29):

$$1 = (1 + \|\boldsymbol{\beta}^*\|^2)\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}[\alpha_1^2 \tanh'(\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\alpha_1)]. \tag{33}$$

The quantity of interest is the following:

$$\langle \boldsymbol{v}_2, \mathcal{H}\boldsymbol{v}_2 \rangle = -1 + \mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\alpha_2^2 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right]. \tag{34}$$

Let us apply Stein's Lemma with respect to α_2 to simplify the expression in (34):

$$\begin{aligned}
&-1 + \mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\alpha_2^2 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right] \\
=&-1 + \mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\nabla_{\alpha_2} \left[\alpha_2 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right]\right] \\
=&-1 + \mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right] \\
&+ 2\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\alpha_2 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon) \tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right] \\
&+ \|\boldsymbol{\beta}^*\|\|\boldsymbol{\beta}\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\alpha_2\alpha_1 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh''\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right] \\
=&2\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\alpha_2 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon) \tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right] \\
&+ \|\boldsymbol{\beta}^*\|\|\boldsymbol{\beta}\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon} \left[\alpha_2\alpha_1 (\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)^2 \tanh''\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2 + \epsilon)\right)\right],
\end{aligned} \tag{35}$$

where (35) follows from relation (31). In addition, if we use Stein's Lemma again for the following expression with respect to α_1, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\nabla_{\alpha_1}\left[\alpha_2\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
&+\|\boldsymbol{\beta}\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)^2\tanh''\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]. \qquad (36)
\end{aligned}$$

Substitute this relation (36) back to (35), we have

$$\begin{aligned}
&\langle \boldsymbol{v}_2, \mathcal{H}\boldsymbol{v}_2\rangle\\
=&\underbrace{\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]}_{A}\\
&+\underbrace{\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]}_{B}. \qquad (37)
\end{aligned}$$

We bound A and B separately. In Lemma 8, we show A is non-negative. In Lemma 9, we show B is at least $\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}$, thus completing the proof. ■

Lemma 8 (Bounding A) *We have*

$$\begin{aligned}
&\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)^2\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right].
\end{aligned}$$

Proof Apply Stein's lemma with respect to α_2 to the left hand side of the equation:

$$\begin{aligned}
&\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\nabla_{\alpha_2}\left[\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\right]\\
=&\|\boldsymbol{\beta}^*\|^2\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^2\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
&+\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^3(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh''\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}+\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|^2\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^3(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh''\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]. \qquad (38)
\end{aligned}$$

In (38), we use relation (33) for the first summand. Let us rewrite

$$\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon=\sqrt{\|\boldsymbol{\beta}^*\|^2+1}z_1,$$

where $z_1 \sim N(0,1)$ and z_1 is independent of α_1. The following relation holds by applying Stein's lemma:

$$\begin{aligned}
&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)^2\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&(\|\boldsymbol{\beta}^*\|^2+1)\mathbb{E}_{z_1,\alpha_1}\left[\alpha_1^2 z_1^2\tanh'\left(\|\boldsymbol{\beta}\|\sqrt{\|\boldsymbol{\beta}^*\|^2+1}\alpha_1 z_1\right)\right]\\
=&(\|\boldsymbol{\beta}^*\|^2+1)\mathbb{E}_{z_1,\alpha_1}\left[\nabla_{z_1}\left[\alpha_1^2 z_1\tanh'\left(\|\boldsymbol{\beta}\|\sqrt{\|\boldsymbol{\beta}^*\|^2+1}\alpha_1 z_1\right)\right]\right]\\
=&(\|\boldsymbol{\beta}^*\|^2+1)\mathbb{E}_{z_1,\alpha_1}\left[\alpha_1^2\tanh'\left(\|\boldsymbol{\beta}\|\sqrt{\|\boldsymbol{\beta}^*\|^2+1}\alpha_1 z_1\right)\right]\\
&+\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|^2+1)^{1.5}\mathbb{E}_{z_1,\alpha_1}\left[\alpha_1^3 z_1\tanh''\left(\|\boldsymbol{\beta}\|\sqrt{\|\boldsymbol{\beta}^*\|^2+1}\alpha_1 z_1\right)\right]\\
=&1+\|\boldsymbol{\beta}\|(\|\boldsymbol{\beta}^*\|^2+1)\mathbb{E}_{\alpha_2,\alpha_1}\left[\alpha_1^3(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh''\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]. \qquad (39)
\end{aligned}$$

Substituting equation (39) in (38), we have:

$$\begin{aligned}
&\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}+\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}\left[\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)^2\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]-1\right]\\
=&\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)^2\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right].
\end{aligned}$$

■

Lemma 9 (Bounding B) *We have*

$$\|\boldsymbol{\beta}^*\|\mathbb{E}\left[\alpha_2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]=\frac{\|\boldsymbol{\beta}^*\|^2}{1+\|\boldsymbol{\beta}^*\|^2}.$$

Proof On the one hand, we can use Stein's lemma with respect to α_1 for the following quantity:

$$\begin{aligned}
&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1\tanh\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\nabla_{\alpha_1}\left[\alpha_2\tanh\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\right]\\
=&\|\boldsymbol{\beta}\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]. \qquad (40)
\end{aligned}$$

On the other hand, we can use Stein's lemma with respect to α_2:

$$\begin{aligned}
&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_2\alpha_1\tanh\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\nabla_{\alpha_2}\left[\alpha_1\tanh\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\right]\\
=&\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|\mathbb{E}_{\alpha_2,\alpha_1,\epsilon}\left[\alpha_1^2\tanh'\left(\|\boldsymbol{\beta}\|\alpha_1(\|\boldsymbol{\beta}^*\|\alpha_2+\epsilon)\right)\right]\\
=&\frac{\|\boldsymbol{\beta}\|\|\boldsymbol{\beta}^*\|}{1+\|\boldsymbol{\beta}^*\|^2}. \qquad (41)
\end{aligned}$$

By setting (40)=(41), we are done. ■

2. Proofs for Main Results on Population EM

2.1. Proof of Theorem 3

Theorem 3 (Sine Convergence) *As long as* $0 \le \theta < \frac{\pi}{2}$, *each population EM iteration satisfies*

$$\sin\theta' \le \kappa \sin\theta, \tag{8}$$

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

Proof From equation (6), we can compute cosine and sine at the next iteration,

$$\cos\theta' = \frac{\langle \boldsymbol{\beta}^*, \boldsymbol{\beta}' \rangle}{\|\boldsymbol{\beta}^*\|\|\boldsymbol{\beta}'\|} = \frac{S\|\boldsymbol{\beta}^*\|^2 + Rb_1^*}{\|\boldsymbol{\beta}^*\|\sqrt{R^2 + S^2\|\boldsymbol{\beta}^*\|^2 + 2SRb_1^*}}, \tag{42}$$

$$\begin{aligned}
\sin\theta' &= \frac{Rb_2^*}{\|\boldsymbol{\beta}^*\|\sqrt{R^2 + S^2\|\boldsymbol{\beta}^*\|^2 + 2SRb_1^*}} \\
&= \sin\theta \frac{1}{\sqrt{1 + (S/R)^2\|\boldsymbol{\beta}^*\|^2 + 2(S/R)b_1^*}} \\
&\le \sin\theta \frac{1}{\sqrt{1 + 2(S/R)b_1^*}}.
\end{aligned} \tag{43}$$

Now we are left with proving $\frac{S}{R}b_1^* \ge \frac{b_1^{*2}}{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}$, which gives us the claimed result by plugging it into (43). To see that, we first observe

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

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

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

We can rewrite the left side as

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

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

2.2. Proof of Theorem 2

Theorem 2 (Cosine Convergence) *As long as $\frac{\pi}{2} > \theta \geq \frac{\pi}{3}$, each population EM iteration satisfies*

$$\cos\theta' \geq \kappa \cos\theta, \tag{7}$$

where $\kappa = \sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}} > 1$. Consequently, if $\cos\theta_0 = \Theta(1/\sqrt{d})$, after $T = O(\log(d)\max(1, \eta^{-2}))$ iterations, we get $\theta_T < \pi/3$ or $\cos\theta_T \geq \frac{1}{2}$.

Proof Recall that from the proof in Theorem 3, we have

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

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

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

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

$$\begin{aligned}
\frac{z\|\boldsymbol{\beta}^*\|^2 + b_1^*}{\|\boldsymbol{\beta}^*\|\sqrt{1 + 2zb_1^* + z^2\|\boldsymbol{\beta}^*\|^2}} &= \sqrt{\frac{z^2\|\boldsymbol{\beta}^*\|^2 + 2zb_1^* + b_1^{*2}/\|\boldsymbol{\beta}^*\|^2}{1 + 2zb_1^* + z^2\|\boldsymbol{\beta}^*\|^2}} \\
&= \sqrt{1 - \frac{b_2^{*2}/\|\boldsymbol{\beta}^*\|^2}{1 + 2zb_1^* + z^2\|\boldsymbol{\beta}^*\|^2}},
\end{aligned}$$

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

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

$$\cos\theta\sqrt{1 + \frac{b_2^{*2}}{k(\sigma^2)^{-1} + b_1^{*2}}} \geq \cos\theta\sqrt{1 + \frac{\sin^2\theta}{\cos^2\theta + \frac{1}{2}(1+\eta^{-2})}} \tag{44}$$

$$\geq \cos\theta\sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}}, \tag{45}$$

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

2.3. Proof of Theorem 4

Before we prove Theorem 4, we state two lemmas that are essential in our proof. Let all the symbols be as defined in Section 2. Recall that

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

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

Lemma 10 $1-\left(\sqrt{1+\frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1,b_1^*)b_1^*}{\sigma_2^2}}\right)^{-1} \le S \le 1.$

Proof From equation (23) in proof of lemma 1, we get

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

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

For the lower bound of S, we can apply the lemmas 1, 2 from Daskalakis et al. (2017).

Lemma 1 in Daskalakis et al. (2017) *Let $\alpha,\beta\ge 0$ and $X\sim\mathcal{N}(\alpha,\sigma^2)$, then* $\mathbb{E}[\tanh'(\beta X/\sigma^2)X]\ge 0$.

Lemma 2 in Daskalakis et al. (2017) *Let $\alpha,\beta\ge 0$ and $X\sim\mathcal{N}(\alpha,\sigma^2)$, then* $\mathbb{E}[\tanh(\beta X/\sigma^2)]\ge 1-\exp[-\frac{\min(\alpha,\beta),\alpha}{2\sigma^2}]$.

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

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

■

Lemma 11 *b_1' is increasing in b_1. Furthermore, in the limit $b_1\to\infty$,*

$$\lim_{b_1\to\infty} b_1' = \frac{2}{\pi}(b_1^*\tan^{-1}\left(\frac{b_1^*}{\sigma_2}\right)+\sigma_2). \tag{46}$$

Proof First, we show that b_1' is increasing in b_1. From (21a), differentiate it with respect to b_1 yields

$$\frac{db_1'}{db_1} = \mathbb{E}\left[\tanh'(\frac{b_1\alpha_1}{\sigma^2}y)y^2\alpha_1^2\right] \geq 0. \tag{47}$$

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

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

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

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

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

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

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

Combining the results, we get the desired lemma. ■

Now we are ready to prove Theorem 4.

Theorem 4 (ℓ_2 Convergence) *Assume that $\theta < \pi/8$, and define $\sigma_2^2 = \sigma^2 + b_2^{*2}$. If $b_2^* < \sigma$ or $\frac{\sigma_2^2}{\sigma^2}b_1 < b_1^*$, then we have*

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \leq \kappa\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \kappa(16\sin^3\theta)\|\boldsymbol{\beta}^*\|\frac{\eta^2}{1+\eta^2}, \tag{9a}$$

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

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\| \leq 0.6\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \tag{9b}$$

Proof [Proof of Theorem 4] First, difference in second coordinate is easily bounded.

$$(b_2^* - b_2') = (1-S)b_2^* \leq \left(\sqrt{1 + \frac{\min(\frac{\sigma_2^2}{\sigma^2}b_1, b_1^*)b_1^*}{\sigma_2^2}}\right)^{-1} b_2^*. \tag{48}$$

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

We start from the following observation. Suppose $b_1 = \frac{\sigma^2}{\sigma_2^2}b_1^*$. From equation (22), we have

$$b_1' = \mathbb{E}_{\alpha_1}[\mathbb{E}_{y\sim\mathcal{N}(\alpha_1 b_1^*, \sigma_2^2)}[\tanh(\frac{\alpha_1 b_1^*}{\sigma_2^2}y)y]\alpha_1] = \mathbb{E}_{\alpha_1}[\alpha_1^2 b_1^*] = b_1^*. \tag{49}$$

Also from Lemma 11, b_1' is increasing in b_1. We will separate the cases based on this point.

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

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

where in (a) we used mean-value theorem, and in (b) we applied lemma 1, 2 in Daskalakis et al. (2017). In turn, we have

$$b_1^* - b_1' \leq \kappa^3\left(b_1^* - \frac{\sigma_2^2}{\sigma^2}b_1\right) \leq \kappa^3(b_1^* - b_1), \tag{50}$$

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

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

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

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

$$b_1' - b_1^* \leq \kappa^3(\frac{\sigma_2^2}{\sigma^2}b_1 - b_1^*) = \kappa^3(b_1 - b_1^*) + \kappa^3\frac{b_2^{*2}}{\sigma^2}b_1. \tag{51}$$

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

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

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

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

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

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

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

(ii) $b_1 < 2b_1^*$. We will assume $b_1\frac{b_2^{*2}}{\sigma^2} \ge (\frac{1}{\kappa^2} - 1)(b_1 - b_1^*)$. Otherwise, we can easily get $b_1' - b_1^* \le \kappa(b_1 - b_1^*)$ similarly by plugging it into equation (51).

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

We bound B. We rearrange terms as below:

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

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

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

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

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

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

Case III. $b_1 > \frac{\sigma_2^2}{\sigma_2^2} b_1^*$, $\sigma < b_2^*$*:* This condition leads us to a special analysis, a constant rate of contraction in local region with high SNR.

First note that, $b_1' \ge b_1^*$ and its difference $(b_1' - b_1^*)$ is increasing in b_1. Therefore, invoking lemma 11 yields

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

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

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

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

On the other side,

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

Combining the result, we get

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

as claimed. ■

Proof of Corollary 1

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

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| < \kappa^T \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + T\kappa^T \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2}. \tag{10}$$

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

$$0.6,\ \sqrt{\left(1 + \frac{\|\boldsymbol{\beta}_0\|^2}{\sigma^2}\right)^{-1}},\ \sqrt{1 - \frac{0.8\eta^2}{1+\eta^2}}. \tag{11}$$

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

If $\frac{\sigma_2^2}{\sigma^2} b_1$ is larger than b_1^*, b_1' becomes larger than b_1^* as we can conclude from Lemma 11 and (49). If $\frac{\sigma_2^2}{\sigma^2} b_1$ were less than b_1^*, then the corresponding $\frac{\sigma_2^2}{\sigma^2} b_1$ at the next iteration is larger than it, as it is inferred from (50). The fact that $b_1^* = \|\boldsymbol{\beta}^*\| \cos\theta_t$ is increasing is obvious from the fact that angle is always decreasing.

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

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

each of which can be rewritten as

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

Since we start from $\theta_0 < \pi/8$, we can plug $\theta_0 = \pi/8$ above and simplify the candidates as (11). We will pick the maximum among these values and fix κ.

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

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

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

3. Proofs for Finite-Sample Based EM

3.1. Proof for Theorem 6

Theorem 6 *Suppose that* $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$. *Then, with* $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon_f^2)$ *samples for one finite-sample based EM iteration, we have*

$$\cos\tilde{\theta}' \geq \kappa(1 - 10\epsilon_f)\cos\theta - O\bigg(\max\Big(\frac{\epsilon_f}{\sqrt{d}}, \epsilon_f^2\Big)\bigg), \tag{12a}$$

$$\sin^2\tilde{\theta}' \leq \kappa'\sin^2\theta + O(\epsilon_f), \tag{12b}$$

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

Proof We start from the end of the proof for Theorem 13. We now replace statistical errors in terms of ϵ_f using the sample complexity $n = \tilde{O}((1+\eta^{-2})d/\epsilon_f^2)$. Recall the way we compute cosine,

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

where the last two inequalities follows from the Lemma 23 in Appendix 5.2, and equation (44) in the proof of Theorem 2.

Now for sine, we have that

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

where the last inequality comes from Theorem 3. ■

3.2. Proof of Theorem 8

Theorem 8 *Suppose that the norm of the current estimator $\|\boldsymbol{\beta}\|$ is larger than $\|\boldsymbol{\beta}^*\|/10$. Then, with $n = \tilde{O}(\max(1, \eta^{-2})d/\epsilon_f^2)$ samples for one Easy-EM iteration, we have*

$$\cos\tilde{\theta}'' \geq \kappa(1 - 10\epsilon_f)\cos\theta - O\left(\frac{\epsilon_f}{\sqrt{d}}\right), \tag{18a}$$

$$\sin^2\tilde{\theta}'' \leq \kappa'\sin^2\theta + O(\epsilon_f), \tag{18b}$$

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

Proof From bounding A in the proof of Theorem 13, we can directly see

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

for some constant c_1. For bounding the norm, standard covering set argument tells that we can take union bound over 1/2-covering set of unit sphere to bound $P(\sup_{v\in\mathbb{S}^d}|(\tilde{\boldsymbol{\beta}}'' - \boldsymbol{\beta}')^\top v| \geq t)$, from which we can conclude

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

with probability at least $1 - \delta$.

Bound for cosine and sine can be derived by the exactly same procedure used in the proof of Theorem 6. ■

Now we are ready to prove lemmas on finite-sample based EM in three convergence phases. We will use the concentration results that with probability $1 - \delta/T$ in each EM iteration, $\|\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\| \leq \epsilon_f$ from Balakrishnan et al. (2017) as well as Theorem 13 in Appendix 5.3.

Proof of Lemma 2

Lemma 2 (Finite-Sample Cosine Convergence) *Assume $\|\tilde{\boldsymbol{\beta}}_0\| \geq \|\boldsymbol{\beta}^*\|/10$. Take $\epsilon_f > 0$ small enough to ensure $\kappa = (1 - 10\epsilon_f)\sqrt{1 + \frac{\eta^2}{\frac{2}{3}+\eta^2}} > 1$. We run the sample-splitting finite-sample EM, each step with $n/T = \tilde{O}\big(\max(1, \eta^{-2})d/\epsilon_f^2\big)$ samples and $T = O\big(\max(1, \eta^{-2})\log d\big)$ iterations. As long as $\tilde{\theta}_t > \pi/3$ for all $t \leq T$, we have with high probability*

$$\cos\tilde{\theta}_T \geq \kappa^T\cos\tilde{\theta}_0 - \frac{\kappa^T - 1}{\kappa - 1}O\Big(\frac{\epsilon_f}{\sqrt{d}}\Big). \tag{14}$$

In particular, when $\cos\tilde{\theta}_0 = \Theta\big(1/\sqrt{d}\big)$, we get $\cos\tilde{\theta}_T \geq \frac{1}{2} - O(\epsilon)$.

Proof From equation (12a) with sufficiently small ϵ_f, we have

$$\begin{aligned}\cos\tilde{\theta}_T &\geq \kappa\cos\tilde{\theta}_{T-1} - O(\frac{\epsilon_f}{\sqrt{d}}) \\ &\geq \kappa^2\cos\tilde{\theta}_{T-2} - (1+\kappa)O(\frac{\epsilon_f}{\sqrt{d}}) \\ &... \\ &\geq \kappa^T\cos\tilde{\theta}_0 - (1+\kappa+\kappa^2+...+\kappa^{T-1})O(\frac{\epsilon_f}{\sqrt{d}}) \\ &\geq \kappa^T\cos\tilde{\theta}_0 - \frac{\kappa^T-1}{\kappa-1}O(\frac{\epsilon_f}{\sqrt{d}}),\end{aligned}$$

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

Proof of lemma 3

Lemma 3 (Finite-Sample Sine Convergence) *Suppose we get a $\tilde{\boldsymbol{\beta}}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/3$ from the previous phase. We run the sample-splitting sample-based EM, each step with $n/T = \tilde{O}(\max(1,\eta^{-2})d/\epsilon_f^2)$ samples. Then with high probability and a constant $\kappa = \left(\sqrt{1+\frac{0.5\eta^2}{1+\eta^2}}\right)^{-1} < 1$, we have*

$$\sin^2\tilde{\theta}_T \leq \kappa^{2T}\sin^2\tilde{\theta}_0 + \frac{1}{1-\kappa^2}O(\epsilon_f). \tag{15}$$

After $T = O(\max(1,\eta^{-2}))$ iterations, we get $\sin^2\tilde{\theta}_T \leq \sin^2\frac{\pi}{70} + O(\epsilon)$.

Proof Similarly,

$$\begin{aligned}\sin^2\tilde{\theta}_T &\leq \kappa^2\sin^2\tilde{\theta}_{T-1} + O(\epsilon_f) \\ &\leq \kappa^4\sin^2\tilde{\theta}_{T-2} + (1+\kappa^2)O(\epsilon_f) \\ &... \\ &\leq \kappa^{2T}\sin^2\tilde{\theta}_0 + (1+\kappa^2+\kappa^4+...+\kappa^{2(T-1)})O(\epsilon_f) \\ &\leq \kappa^{2T}\sin^2\tilde{\theta}_0 + \frac{1}{1-\kappa^2}O(\epsilon_f),\end{aligned}$$

with probability $1-\delta$.

Finally,

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

which yields the desired result. ■

4. Proof of Theorem 7

Theorem 7 (Finite-Sample Distance Convergence) *Suppose we get a $\tilde{\boldsymbol{\beta}}_0$ whose angle with $\boldsymbol{\beta}^*$ is less than $\frac{\pi}{70}$ from the previous phase. There exist a constant $C > 1$ for which the following holds.*

- *If $\eta < C$, sample-splitting finite-sample EM with $n/T = \tilde{O}(\eta^{-2} d/\epsilon_f^2)$ samples per iteration satisfies*

$$\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq \kappa^T \|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + T\kappa^T \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2} + O(\epsilon)\|\boldsymbol{\beta}^*\|, \tag{16}$$

  where κ is the maximum among (11) as in Corollary 1. After $T = O(\eta^{-2}\log(1/\epsilon))$ iterations, we estimate $\boldsymbol{\beta}^*$ with an ℓ_2 error bounded by $O(\epsilon)$.

- *If $\eta \geq C$, sample-splitting finite-sample EM with $n/T = \tilde{O}(d/\epsilon_f^2)$ samples per iteration satisfies*

$$\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq \kappa^T \|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + O(\epsilon)\sigma, \tag{17}$$

  where $\kappa = 0.95 + \epsilon_f < 1$. After $T = O(\log(1/(\sigma\epsilon)))$ iterations, we estimate $\boldsymbol{\beta}^*$ with an ℓ_2 error bounded by $\sigma O(\epsilon)$.

The first part of the theorem is proved in Section 4.1 and the second part of the theorem is proved in Section 4.2.

4.1. Statistical Bound depending on the optimal parameter

Lemma 12 (Convergence of Distance in Finite-Sample EM) *Suppose we get $\tilde{\boldsymbol{\beta}}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\pi/8$ from previous phase. We run sample-splitting EM with $n/T = \tilde{O}(\max(1, \eta^{-2})d/\epsilon^2)$, getting*

$$\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| \leq \kappa^T \|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + T\kappa^T \|\boldsymbol{\beta}^*\| \frac{\eta^2}{1+\eta^2} + O(\epsilon)\|\boldsymbol{\beta}^*\|, \tag{52}$$

where κ is the maximum among (11) as in Corollary 1.

After $T = O(\max(1, \eta^{-2})\log(1/\epsilon_1))$ iterations, we get $\boldsymbol{\beta}_*$ within $O(\epsilon)$ error.

Proof We assume $||\boldsymbol{\beta}^*|| = 1$ for the sake of simplicity in the proof. We start from Theorem 4. Note that the chosen κ satisfies $\sin^3\tilde{\theta}_T \leq \kappa^T \sin^3\tilde{\theta}_0 + \frac{1}{1-\kappa}O(\epsilon_f)$, which can be shown similarly as

Lemma 3.

$$\begin{aligned}
\|\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\| &\leq \kappa\|\tilde{\boldsymbol{\beta}}_{T-1} - \boldsymbol{\beta}^*\| + O(\epsilon_f) + \kappa(16\sin^3\tilde{\theta}_{T-1})\frac{\eta^2}{1+\eta^2} \\
&\leq \kappa^2\|\tilde{\boldsymbol{\beta}}_{T-2} - \boldsymbol{\beta}^*\| + (1+\kappa)O(\epsilon_f) + \frac{16\eta^2}{1+\eta^2}(\kappa^2\sin^3\tilde{\theta}_{T-2} + \kappa\sin^3\tilde{\theta}_{T-1}) \\
&\ldots \\
&\leq \kappa^T\|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + \frac{1}{1-\kappa}O(\epsilon_f) + \frac{16\eta^2}{1+\eta^2}(\kappa^T\sin^3\tilde{\theta}_0 + \kappa^{T-1}\sin^3\tilde{\theta}_1 + \ldots + \kappa\sin^3\tilde{\theta}_{T-1}) \\
&\leq \kappa^T\|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + \frac{1}{1-\kappa}O(\epsilon_f) + \frac{16\eta^2}{1+\eta^2}(T\kappa^T\sin^3\tilde{\theta}_0 + \frac{\kappa+\kappa^2+\ldots+\kappa^T}{1-\kappa}O(\epsilon_f)) \\
&\leq \kappa^T\|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + \frac{1}{1-\kappa}O(\epsilon_f) + T\kappa^T\frac{\eta^2}{1+\eta^2} + \frac{16\eta^2}{1+\eta^2}\frac{1}{(1-\kappa)^2}O(\epsilon_f) \\
&= \kappa^T\|\tilde{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\| + T\kappa^T\frac{\eta^2}{1+\eta^2} + \frac{1}{1-\kappa}O(\epsilon_f) + \frac{1}{(1-\kappa)^2}\frac{\eta^2}{1+\eta^2}O(\epsilon_f).
\end{aligned}$$

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

4.2. Statistical Bound independent of the optimal parameter

The statistical error we have seen in the previous result is proportional to $\|\boldsymbol{\beta}^*\|$. This unsatisfactory result, especially in the high SNR regime, has been often ignored in literature as if EM algorithm does not guarantee an exact recovery. However, this is in contrast to the result in Yi et al. (2014) where they guaranteed exact recovery in the noiseless setting. In other words, the existing statistical guarantees are not *tight*. In this section, we provide a more refined analysis of the (standard) EM algorithm in the finite sample case. The main difference from previous technique is that instead of coupling $\tilde{\boldsymbol{\beta}}'$ and $\boldsymbol{\beta}'$ directly, we utilize the sample covariance matrix ($\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$) to decompose the error between $\tilde{\boldsymbol{\beta}}'$ and $\boldsymbol{\beta}^*$ so that the additive statistical error does not depend on $\|\boldsymbol{\beta}^*\|$. One implication of our results is that in the high SNR regime, the l_2 estimation error of the EM iterate does not scale linearly with $\|\boldsymbol{\beta}^*\|$, but only with σ.

Recall that the $1-d$ EM update α' for GMM with two symmetric components Daskalakis et al. (2017), with the current parameter α and the optimal parameter is the following:

$$\alpha' = \mathbb{E}_{X\sim\mathcal{N}(\alpha^*,\sigma^2)} X\tanh\left(\frac{\alpha X}{\sigma^2}\right).$$

The consistency property guarantees that:

$$\alpha^* = \mathbb{E}_{X\sim\mathcal{N}(\alpha^*,\sigma^2)} X\tanh\left(\frac{\alpha^* X}{\sigma^2}\right).$$

It follows that for each $i = 1,\ldots,n$:

$$\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,\sigma^2)} y_i\boldsymbol{x}_i\tanh\left(\frac{y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle}{\sigma^2}\right) = \boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\beta}^*.$$

This allows us to decompose the difference of $\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^*$ in the following way:

$$\begin{aligned}
&\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^* \\
&= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle}{\sigma^2}\right) - \mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle}{\sigma^2}\right)\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\underbrace{\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle}{\sigma^2}\right) - \mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle}{\sigma^2}\right)\right)}_{\text{term 1}} \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\underbrace{\left(\mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle}{\sigma^2}\right) - \mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle}{\sigma^2}\right)\right)}_{\text{term 2}}.
\end{aligned}$$

For term 1, we can apply a standard concentration result for function of Gaussian random variables by conditioning on the event that the covariance matrix $(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top)$ is close to 1 in spectral norm. Specifically, it has been proved that this term is $O(\sqrt{\frac{d}{n}})$ in ℓ_2 norm, which is independent of $\|\boldsymbol{\beta}^*\|$. For term 2, we observe that for each i, the following difference

$$\mathbb{E}_{y_i} y_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle}{\sigma^2}\right) - \mathbb{E}_{y_i} y_i \tanh\left(\frac{y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle}{\sigma^2}\right)$$

is the difference between a 1-d population EM iterate and the optimal parameter in the GMM problem with the current iterate being $\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle$ and the optimal parameter being $\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle$. We are able to adapt the sensitivity analysis technique in Daskalakis et al. (2017) here to show that this term is a contraction term when SNR is large.

The main result is summarized in the following theorem:

Theorem 10 (Improved Convergence of distance in Finite-Sample EM in high SNR) *There exists constants $C > 1$ such that for all $\eta > C$, the following holds: suppose we get $\boldsymbol{\beta}_0$ whose angle formed with $\boldsymbol{\beta}^*$ is less than $\frac{\pi}{70}$ and $\|\boldsymbol{\beta}_0\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$, we run the sample-splitting finite-sample EM with $n/T = \tilde{O}(d/\epsilon_f^2)$, getting:*

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| \leq (0.95 + \epsilon_f)^\top \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + O(\epsilon_f)\sigma. \tag{53}$$

After $T = O(\log(1/\epsilon_f))$ iterations, we estimate $\boldsymbol{\beta}_*$ with an ℓ_2 error bounded by $O(\epsilon_f)$.

4.3. Proof of Theorem 10

Proof Consider the scaling: $\boldsymbol{\beta} \to \frac{\boldsymbol{\beta}}{\sigma}$, $\boldsymbol{\beta}^* \to \frac{\boldsymbol{\beta}^*}{\sigma}$, and $y \to \frac{y}{\sigma}$. We can without loss of generality assume that $\sigma = 1$. In the following proof, we omit the appearance of σ for simplicity. As we have

shown before, we can decompose the difference $\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^*$ in the following way:

$$\begin{aligned}
&\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^* \\
&= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right) - \mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle\right)\right) \\
&= \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}}_{A}\underbrace{\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right) - \mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right)\right)}_{B} \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\underbrace{\left(\mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right) - \mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle\right)\right)}_{C}.
\end{aligned}$$

We provide bounds for A, B and C in the following:

- $\|A\| = 1 + O\left(\sqrt{\frac{d}{n}}\right)$ (standard concentration result),
- Conditioning on the sample covariance matrix has bounded spectral norm, $\|B\| = O\left(\sqrt{\frac{d}{n}}\right)$ (cf. Proposition 11),
- If $\eta \geq c$ for some constant $c > 1$, $C \leq \left(0.95 + O(1/\sqrt{d})\right)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|$ (cf. Proposition 12).

Therefore, for one step of EM, with $n/T = \tilde{O}(d/\epsilon_f^2)$ samples,

$$\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}^* = (1+\epsilon_f)\epsilon_f + (1+\epsilon_f)(0.95+\epsilon_f)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq 2\epsilon_f + (0.95 + 2\epsilon_f)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|.$$

Since all the future iterate remains lower bounded by $\frac{\|\boldsymbol{\beta}^*\|}{10}$ (cf. Lemma 23) in ℓ_2 norm and the angle increases (cf. Lemma 3). We can use induction to show that:

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*\| \leq (0.95 + 2\epsilon_f)^\top\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\| + \frac{2}{0.2 - \epsilon_f}\epsilon$$

The result follows by picking small enough ϵ_f. ■

Proposition 11 *(Controlling B) For each fixed $\boldsymbol{\beta}$, with probability at least $1 - \exp(-cn) - 6^d\exp\left(-\frac{nt^2}{72}\right)$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{y_i}\left[y_i\boldsymbol{x}_i \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right)\right]\right\| \leq t$$

for some absolute constant $c > 0$.

Proof We will use the standard epsilon-net argument. Let $\boldsymbol{v} \in \mathbb{R}^d$, define

$$f_{\boldsymbol{v}}(y) := \frac{1}{n}\sum_{i=1}^{n} y_i \langle \boldsymbol{x}_i, \boldsymbol{v}\rangle \tanh\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right).$$

Suppose that $\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\| \le 2$, we show in the following that $f_{\boldsymbol{v}}(y) - \mathbb{E}_y f_{\boldsymbol{v}}(y)$ is $\frac{9}{n}$-subgaussian. The tool is a standard concentration result for function of the Gaussian random variables summarized in Lemma 13:

Lemma 13 (Lemma 2.1 of Wainwright (2015)) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, then for every convex $\phi : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}\left(\phi\left(f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})\right)\right) \le \mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(\boldsymbol{X}), \boldsymbol{Y}\rangle\right)\right],$$

where $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, I_n)$ are standard Gaussians and are independent.

Note that for each i, the derivative of $f_{\boldsymbol{v}}(y)$ with respect to y can be computed explicitly:

$$\frac{\partial f_{\boldsymbol{v}}(y)}{\partial y_i} = \langle \boldsymbol{x}_i, \boldsymbol{v}\rangle(\tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) + y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle \tanh'(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)) \ \forall i.$$

Here we abuse the notation y. We actually take the derivative with respect to the noise in y_i, which is distributed as a standard Gaussian. The following numerical inequality for $g(z) := \tanh(z) + z\tanh'(z)$ will be used:

$$|\tanh(z) + z\tanh'(z)| \le 2.$$

For any $\lambda \in \mathbb{R}$, we have:

$$\begin{aligned}
&\mathbb{E}_y\left[\exp\left(\lambda\left(f_{\boldsymbol{v}}(y) - \mathbb{E}\left[f_{\boldsymbol{v}}(y)\right]\right)\right)\right] \\
\le&\mathbb{E}_{y,\boldsymbol{z}}\left[\exp\left(\lambda\left(\frac{\pi}{2}\langle \nabla f'_{\boldsymbol{v}}(y), \boldsymbol{z}\rangle\right)\right)\right] && \text{(Lemma 13)} \\
=&\mathbb{E}_y\mathbb{E}_{\boldsymbol{z}}\left[\exp\left(\lambda\left(\frac{\pi}{2}\frac{1}{n}\sum_{i=1}^{n} z_i\langle \boldsymbol{x}_i, \boldsymbol{v}\rangle g(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)\right)\right)\right] \\
=&\mathbb{E}_y \exp\left(\lambda^2\frac{\pi^2}{8n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\langle \boldsymbol{x}_i, \boldsymbol{v}\rangle g\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle\right)\right)^2\right]\right) && \text{(independence of } z_i s'\text{)} \\
\le&\mathbb{E}_y \exp\left(\lambda^2\frac{\pi^2}{2n}\left[\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i, \boldsymbol{v}\rangle^2\right]\right) && \text{(numerical bound on } g\text{)} \\
=&\mathbb{E}_y \exp\left(\lambda^2\frac{\pi^2}{2n}\boldsymbol{v}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right]\boldsymbol{v}\right) \\
\le&\mathbb{E}_y \exp\left(\frac{\pi^2\lambda^2}{n}\right) \le \exp\left(\frac{18\lambda^2}{n}\right),
\end{aligned}$$

where the last line comes from our assumption that $\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\| \leq 2$. Using the standard $\frac{1}{2}$-net argument for the norm, we can show that

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\left[y_i\boldsymbol{x}_i\tanh\left(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle\right)-\mathbb{E}_{y_i}y_i\boldsymbol{x}_i\tanh\left(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle\right)\right]\right\|>t\;\middle|\;\left\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|\leq 2\right)\\ \leq&6^d\exp\left(-\frac{nt^2}{72}\right).\end{aligned}$$

To finish the proof, we note the event $\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\| \leq 2$ holds with probability at least $1-\exp(-cn)$ for some absolute constant $c>0$. ■

Proposition 12 *(Controlling C) There exists an absolute constant $c>1$ such that in the regime where $\eta > c$ the following holds: for each fixed $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$, and its angle with $\boldsymbol{\beta}^*$, θ is less than $\frac{\pi}{70}$, we run a finite-sample EM with $n=\tilde{O}\left(\frac{d}{\epsilon_f^2}\right)$, getting:*

$$\begin{aligned}&\left\|\mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_y\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle)\right\|\\ \leq&(0.95+\epsilon_f/\sqrt{d})\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|\end{aligned}$$

Proof For each i, we observe that

$$\mathbb{E}_{y_i}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_{y_i}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle)$$

is the difference between the 1-d population EM iterate and the optimal parameter for the GMM problem with the current iterate being $\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle$ and the optimal parameter being $\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle$. In Daskalakis et al. (2017), they have developed the sensitivity analysis technique to bound the difference with the restriction that the current iterate has the same sign as the optimal parameter. In our case, we note that covariance vector $\boldsymbol{x}_i$ can possibly cause $\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle$ and $\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle$ to have opposite signs despite the fact that $\boldsymbol{\beta}$ has an acute angle with $\boldsymbol{\beta}^*$. We get around this issue by performing a more refined sensitivity analysis in both regions: (1)$\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle\geq 0$; (2) $\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle<0$.

The key element of the sensitivity analysis is to use the following decomposition:

$$\begin{aligned}&\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle)\\ =&\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)\\ &+\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle)\\ =&\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)\\ &+\boldsymbol{\beta}-\boldsymbol{\beta}^*\end{aligned}$$

where the last step follows from the consistency property of the EM operator. The mean value theorem tells us that:

$$\begin{aligned}&\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)-\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle,1)}y_i\tanh(y_i\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)\\ =&\int_{t=0}^1\mathbb{E}_{y_i\sim\mathcal{N}(\langle\boldsymbol{x}_i,Z(t)\rangle,1)}\Delta_i(y)\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{\beta}^*-\boldsymbol{\beta})dt\end{aligned}$$

with

$$Z(t) = \boldsymbol{\beta} + t(\boldsymbol{\beta}^* - \boldsymbol{\beta}),$$
$$\Delta_i(y) = \frac{\partial[y_i \boldsymbol{x}_i \tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)]}{\partial y_i} = \tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) + y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle \tanh'(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle).$$

Therefore, the original difference is equivalent to:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{y_i\sim\mathcal{N}(\langle \boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)} y_i \tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{y_i\sim\mathcal{N}(\langle \boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)} y_i \tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle)\\ =&\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top \left(1 - \int_0^1 \mathbb{E}_{y\sim\mathcal{N}(\langle \boldsymbol{x}_i, Z(t)\rangle,1)}\Delta_i(y))dt\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \end{aligned} \tag{54}$$

Since $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ is fixed, we can assume that the Gaussians $\{\boldsymbol{x}_i\}$s have the orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^n$ satisfying $\boldsymbol{v}_1 = \hat{\boldsymbol{\beta}}_*$ and span$(\boldsymbol{v}_1, \boldsymbol{v}_2)$ =span$(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$. Therefore, to bound the difference term (54), it suffices to understand the spectral norm of the 2×2 submatrix:

$$\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top \left(1 - \int_0^1 \mathbb{E}_{y\sim\mathcal{N}(\langle \boldsymbol{x}_i, Z(t)\rangle,1)}\Delta_i(y))dt\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right]_{:2,:2} \tag{55}$$

so that we can bound:

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{y_i\sim\mathcal{N}(\langle \boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)} y_i \tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{y_i\sim\mathcal{N}(\langle \boldsymbol{x}_i,\boldsymbol{\beta}^*\rangle,1)} y_i \tanh(y_i\langle \boldsymbol{x}_i, \boldsymbol{\beta}^*\rangle)\right\|\\ \le&\left\|\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top \left(1 - \int_0^1 \mathbb{E}_{y\sim\mathcal{N}(\langle \boldsymbol{x}_i, Z(t)\rangle,1)}\Delta_i(y))dt\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right]_{:2,:2}\right\| \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|. \end{aligned} \tag{56}$$

We will provide an explicit bound for $1 - \int_0^1 \mathbb{E}_{y\sim\mathcal{N}(\langle \boldsymbol{x}_i, Z(t)\rangle,1)}\Delta_i(y))dt$ in Lemma 15. For now, we just need to use the fact that they are bounded by a constant. This implies that each entry of the matrix (55) is a sub-exponential random variable and it is close to its expectation with statistical error $O(1/\sqrt{n})$. Furthermore, we can deduce the spectral norm of this 2×2 submatrix is close to the spectral norm of the expectated submatrix, with a statistical error $O(1/\sqrt{n})$. The problem is thus further reduced to bound the spectral norm of the following 2×2 matrix:

$$\mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[1 - \int_0^1 \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle,1)}\Delta(y))dt)\right]_{:2,:2} \tag{57}$$

In Lemma 14, it is proved that the spectral norm of (57) is bounded by

$$\frac{1}{2}\left(1 + \frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{\pi}\max\left(1 - \frac{1}{(1+0.5\tau^2)^2}, \sin(\theta)\right) + 2.6\sin(\theta),$$

where $\tau = \min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)$. Since $\|\boldsymbol{\beta}\| \ge \frac{\|\boldsymbol{\beta}^*\|}{10}$, it follows that $\tau \ge \frac{\|\boldsymbol{\beta}^*\|}{10}$, and

$$1 - \frac{1}{(1+0.5\tau^2)^2} \ge 1 - \frac{1}{(1+0.05\|\boldsymbol{\beta}^*\|^2)^2}.$$

As long as $\|\boldsymbol{\beta}^*\|$ is sufficiently large, $1-\frac{1}{(1+0.05\|\boldsymbol{\beta}^*\|^2)^2}$ will dominate $\sin(\theta)\leq\frac{\pi}{8}$. Therefore, a loose bound for the above spectral norm is:

$$\frac{1}{2}+\frac{1}{\pi}+(\frac{1}{2}-\frac{1}{\pi})\frac{1}{(1+0.05\|\boldsymbol{\beta}^*\|^2)^2}+2.6\sin(\theta)$$

We note that as $\|\boldsymbol{\beta}^*\|\to\infty$ and let $\phi=\frac{\pi}{70}$, the above term converges to $\frac{1}{2}+\frac{1}{\pi}+4\sin(\theta)<0.95$. Thus, there exists c, such that whenever $\|\boldsymbol{\beta}^*\|\geq c$, the above ratio is bounded by 0.95 for all $\theta<\frac{\pi}{70}$. Now we can conclude that the matrix (55) has spectral norm $0.95+O(1/\sqrt{n})$ with high probability and the proof follows from (56). ■

Lemma 14 *Let $\tau=\max(\|\boldsymbol{\beta}^*\|,\|\boldsymbol{\beta}\|)$, and θ be the angle between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Suppose that $\theta\leq\frac{\pi}{8}$ and the orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^d$ satisfy $\boldsymbol{v}_1=\hat{\boldsymbol{\beta}}^*$ and span$(\boldsymbol{v}_1,\boldsymbol{v}_2)$ =span$(\boldsymbol{\beta},\boldsymbol{\beta}^*)$, the following inequality holds:*

$$\begin{aligned}&\left\|\left[\mathbb{E}_{X\sim\mathcal{N}(0,I)}XX^\top\left(1-\int_0^1\mathbb{E}_{y\sim\mathcal{N}(\langle X,Z(t)\rangle,1)}\Delta(t))dt\right)\right]_{:2,:2}\right\|\\ \leq&\frac{1}{2}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right)+\frac{1}{\pi}\max\left(1-\frac{1}{(1+0.5\tau^2)^2},\sin(\theta)\right)+C\sin(\theta).\end{aligned}$$

for some absolute constant $0<C\leq 2.6$.

Proof We first provide an elementary bound for $1-\int_0^1\mathbb{E}_{y\sim\mathcal{N}(\langle X,Z(t)\rangle,1)}\Delta(t))dt)$ in Lemma 15. Using symmetry and Lemma 15, the following holds for the 2 by 2 submatrix of $\mathbb{E}_{X\sim\mathcal{N}(0,I)}XX^\top(1-\int_0^1\mathbb{E}_{y\sim\mathcal{N}(\langle X,Z(t)\rangle,1)}\Delta(t))dt)$:

$$\begin{aligned}&\left[\mathbb{E}_{X\sim\mathcal{N}(0,I)}XX^\top\left(1-\int_0^1\mathbb{E}_{y\sim\mathcal{N}(\langle X,Z(t)\rangle,1)}\Delta(t))dt\right)\right]_{:2,:2}\\ \preceq&\left[2\mathbb{E}_{X\sim\mathcal{N}(0,I)}1_{\langle X,\boldsymbol{\beta}\rangle>0,\langle X,\boldsymbol{\beta}^*\rangle>0}XX^\top\exp\left(-\frac{\min(\langle X,\boldsymbol{\beta}\rangle,\langle X,\boldsymbol{\beta}^*\rangle)^2}{2}\right)+2.25\mathbb{E}_{X\sim\mathcal{N}(0,I)}1_{\langle X,\boldsymbol{\beta}\rangle\langle X,\boldsymbol{\beta}^*\rangle<0}XX^\top\right]_{:2,:2}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\|\left[\mathbb{E}_{X\sim\mathcal{N}(0,I)}XX^\top(1-\int_0^1\mathbb{E}_{y\sim\mathcal{N}(\langle X,Z(t)\rangle,1)}\Delta(t))dt)\right]_{:2,:2}\right\|\\ \leq&2\Bigg\|\underbrace{\left[\mathbb{E}_{X\sim\mathcal{N}(0,I)}1_{\langle X,\boldsymbol{\beta}\rangle>0,\langle X,\boldsymbol{\beta}^*\rangle>0}XX^\top\exp\left(-\frac{\min(\langle X,\boldsymbol{\beta}\rangle,\langle X,\boldsymbol{\beta}^*\rangle)^2}{2}\right)\right]_{:2,:2}}_{M_1}\Bigg\|\\ &+2.25\Bigg\|\underbrace{\left[\mathbb{E}_{X\sim\mathcal{N}(0,I)}1_{\langle X,\boldsymbol{\beta}\rangle\langle X,\boldsymbol{\beta}^*\rangle<0}XX^\top\right]_{:2,:2}}_{M_2}\Bigg\|.\end{aligned}$$

We use the polar coordinates (r, ϕ) for the first two components: X_1, X_2.

$$\begin{aligned}(X_1, X_2) &= (r\cos\phi, r\sin\phi)\\ \langle X, \boldsymbol{\beta}\rangle &= \|\boldsymbol{\beta}\|X_1\cos(\phi) + \|\boldsymbol{\beta}\|X_2\sin(\phi)\\ &= \|\boldsymbol{\beta}\|r\cos(\phi-\theta)\\ \langle X, \boldsymbol{\beta}^*\rangle &= \|\boldsymbol{\beta}^*\|X_1 = \|\boldsymbol{\beta}^*\|r\cos(\phi).\end{aligned}$$

It is seen that the region $S_1 := \{X : \langle X, \boldsymbol{\beta}\rangle > 0, \langle X, \boldsymbol{\beta}^*\rangle > 0\}$ corresponds to $S_1 = \{(r,\phi) : r > 0, \phi \in (-\frac{\pi}{2}+\theta, \frac{\pi}{2})\}$ using the polar coordinates. Similarly, the region $S_2 := \{X : \langle X, \boldsymbol{\beta}\rangle\langle X, \boldsymbol{\beta}^*\rangle < 0\}$ corresponds to $S_1 = \{(r,\phi) : r > 0, \phi \in (-\frac{\pi}{2}, -\frac{\pi}{2}+\theta)\cup(\frac{\pi}{2}, \frac{\pi}{2}+\theta)\}$ using the polar coordinates. This helps us to get an explicit formula for each of the entry in M_1 and M_2. Before providing bounds for each entry, we use the following relation for $\min(\langle X, \boldsymbol{\beta}\rangle, \langle X, \boldsymbol{\beta}^*\rangle)$:

$$\begin{aligned}\cos(\phi-\theta) \geq \cos(\phi) \quad \phi \in \left(\frac{\theta}{2}, \frac{\pi}{2}\right),\\ \cos(\phi-\theta) \leq \cos(\phi) \quad \phi \in \left(-\frac{\pi}{2}+\theta, \frac{\theta}{2}\right).\end{aligned} \tag{58}$$

Therefore,

$$\min(\langle X, \boldsymbol{\beta}\rangle, \langle X, \boldsymbol{\beta}^*\rangle) \in \begin{cases}(\tau r\cos(\theta), \tau r\cos(\phi-\theta) & \phi \in \left(\frac{\theta}{2}, \frac{\pi}{2}\right)\\ (\tau r\cos(\phi-\theta), \tau r\cos(\theta)) & \phi \in \left(-\frac{\pi}{2}+\theta, \frac{\theta}{2}\right).\end{cases} \tag{59}$$

In Lemma 16 and Lemma 20, we show that:

- $(1,1)$th entry of M_1: $0 < M_1^{1,1} \leq \frac{1}{4}\left(1 + \frac{1}{(1+0.5\tau^2)^2}\right) + \frac{\sin(\theta)}{2\pi}$,
- $(2,2)$th entry of M_1: $0 < M_1^{2,2} \leq \frac{1}{4}\left(1 + \frac{1}{(1+0.5\tau^2)}\right) + \frac{1}{2\pi}\left(1 - \frac{1}{(1+0.5\tau^2)^2}\right)$,
- $(1,2)$th entry of M_1: $|M_1^{1,2}| \leq \max\left(\sin^2(\phi)\left[\frac{1}{2\pi(1+\cos^2(\theta)\tau^2)} + \frac{1}{2\pi(1+\tau^2\sin^2(\theta))}\right], \frac{1}{\pi}\sin(\theta) + \frac{1}{2\pi}\frac{\sin^2(\theta)}{1+\tau^2\sin^2(\theta)}\right)$,
- $(1,1)$th entry of M_2: $M_2^{1,1} = \frac{\theta}{\pi} - \frac{\sin(2\theta)}{2\pi}$,
- $(2,2)$th entry of M_2: $M_2^{2,2} = \frac{\theta}{\pi} + \frac{\sin(2\theta)}{2\pi}$,
- $(1,2)$th entry of M_2: $M_2^{1,2} = -\frac{\sin^2(\theta)}{\pi}$.

Now we can apply Lemma 21 to bound the spectral norm of M_1 and M_2:

$$\begin{aligned}\|M_1\| \leq& \frac{1}{4}\left(1 + \frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{2\pi}\max\left(\sin(\theta), 1 - \frac{1}{(1+0.5\tau^2)^2}\right) + |M_1^{1,2}|\\ \leq& \frac{1}{4}\left(1 + \frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{2\pi}\max\left(\sin(\theta), 1 - \frac{1}{(1+0.5\tau^2)^2}\right) + C_1\sin(\theta)\end{aligned}$$

for some absolute constant $0 < C_1 < 0.4$. Similarly,

$$\|M_2\| \leq \frac{\theta}{\pi} + \frac{\sin(2\theta)}{2\pi} + \frac{\sin^2(\theta)}{\pi} \leq C_2\sin(\theta).$$

for some absolute constant $0 < C_2 < 0.8$. In the last step, we use the fact that when $\theta \in (0, \frac{\pi}{8})$, $\theta \leq 1.1\sin(\theta)$. We thus obtain a compact bound for the spectral norm of the 2×2 matrix, $\left\| \left[\mathbb{E}_{X\sim\mathcal{N}(0,I)} XX^\top (1 - \int_0^1 \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} \Delta(t)) dt)\right]_{2\times 2} \right\|$:

$$2\|M_1\| + 2.25\|M_2\| \leq \frac{1}{2}\left(1 + \frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{\pi}\max\left(\sin(\theta), 1 - \frac{1}{(1+0.5\tau^2)^2}\right) + C_3\sin(\theta) \tag{60}$$

for some absolute constant $0 < C_3 \leq 2.6$. ■

Lemma 15 *When* $\langle X, \boldsymbol{\beta}\rangle > 0, \langle X, \boldsymbol{\beta}^*\rangle > 0$,

$$\int_{t=0}^1 [1 - \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} \Delta(t)] dt \leq \exp\left(-\frac{\min(\langle X, \boldsymbol{\beta}\rangle, \langle X, \boldsymbol{\beta}^*\rangle)^2)}{2}\right).$$

When $\langle X, \boldsymbol{\beta}\rangle < 0, \langle X, \boldsymbol{\beta}^*\rangle < 0$,

$$\int_{t=0}^1 [1 - \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} \Delta(t)] dt \leq \exp\left(-\frac{\min(-\langle X, \boldsymbol{\beta}\rangle, -\langle X, \boldsymbol{\beta}^*\rangle)^2}{2}\right).$$

When $\langle X, \boldsymbol{\beta}\rangle$ *and* $\langle X, \boldsymbol{\beta}^*\rangle$ *have different sign,*

$$\int_{t=0}^1 [1 - \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} \Delta(t)] dt \leq 2.25.$$

Proof We first show the bound for $\langle X, \boldsymbol{\beta}\rangle > 0$, $\langle X, \boldsymbol{\beta}^*\rangle > 0$, and the bound for $\langle X, \boldsymbol{\beta}\rangle < 0$, $\langle X, \boldsymbol{\beta}^*\rangle < 0$ can be proved in the same way.

$$\begin{aligned}
&\int_{t=0}^1 \left[1 - \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} \Delta(t)\right] dt \\
&= \int_{t=0}^1 \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} \left[1 - \tanh(y\langle X, \boldsymbol{\beta}\rangle)\right] dt - \int_{t=0}^1 \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)} y\langle X, \boldsymbol{\beta}\rangle \tanh'(y\langle X, \boldsymbol{\beta}\rangle) dt \qquad (61) \\
&\leq \int_{t=0}^1 \exp\left(-\frac{Z(t)\min(Z(t), \langle X, \boldsymbol{\beta}\rangle)}{2}\right) dt \qquad (62) \\
&\leq \int_{t=0}^1 \exp\left(-\frac{\min(\langle X, \boldsymbol{\beta}\rangle, \langle X, \boldsymbol{\beta}^*\rangle)^2}{2}\right) dt \qquad (63) \\
&= \exp\left(-\frac{\min(\langle X, \boldsymbol{\beta}\rangle, \langle X, \boldsymbol{\beta}^*\rangle)^2}{2}\right).
\end{aligned}$$

Inequality (62) follows since the second summand in (61) is non-negative (c.f Lemma 2.3) and inequality (63) follows from Lemma 2.3, with the condition $\langle X, \boldsymbol{\beta}\rangle Z(t) \geq 0$ satisfied. To establish the bound for $\langle X, \boldsymbol{\beta}^*\rangle\langle X, \boldsymbol{\beta}\rangle < 0$, we again use the following numerical inequality:

$$|\tanh(z) + z\tanh'(z)| \leq 1.25.$$

Therefore,

$$\int_{t=0}^{1} [1 - \mathbb{E}_{y\sim\mathcal{N}(\langle X, Z(t)\rangle, 1)}\Delta(t)]dt \leq 2.25$$

■

Lemma 16 (M_1) *Let $\tau = \min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)$, and let θ be the angle between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, the following bounds hold for each entry of the symmetric 2×2 matrix M_1:*

$$\begin{aligned}
0 < M_1^{1,1} &\leq \frac{1}{4}\left(1 + \frac{1}{(1+0.5\tau^2)^2}\right) + \frac{\sin(\theta)}{2\pi}, \\
0 < M_1^{2,2} &\leq \frac{1}{4}\left(1 + \frac{1}{(1+0.5\tau^2)}\right) + \frac{1}{2\pi}\left(1 - \frac{1}{(1+0.5\tau^2)^2}\right), \\
M_1^{1,2} &\leq \sin^2(\theta)\left[\frac{1}{2\pi(1+\cos^2(\theta)\tau^2)} + \frac{\tau^2}{(1+\tau^2\sin^2(\theta))(1+\tau^2)}\right], \\
-M_1^{1,2} &\leq \frac{1}{\pi}\sin(\theta) + \frac{1}{2\pi}\frac{\sin^2(\theta)}{1+\tau^2\sin^2(\theta)}.
\end{aligned}$$

Proof We need to go through a very careful integration. $M_1^{1,1}$ and $M_1^{2,2}$ are clearly non-negative.

$$\begin{aligned}
M_1^{1,1} &= \left[\mathbb{E}_{X\sim\mathcal{N}(0,I)} 1_{\langle X,\boldsymbol{\beta}\rangle>0, \langle X,\boldsymbol{\beta}^*\rangle>0} XX^\top \exp\left(-\frac{\min(\langle X,\boldsymbol{\beta}\rangle, \langle X,\boldsymbol{\beta}^*\rangle)^2}{2}\right)\right]_{11} \\
&= \int_{r=0}^{\infty}\int_{\phi=-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}} r^2\cos^2(\phi)\exp\left(-\frac{\min(\|\boldsymbol{\beta}\|r\cos(\phi-\theta), \|\boldsymbol{\beta}^*\|r\cos(\phi))^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&\leq \int_{r=0}^{\infty}\int_{\phi=-\frac{\pi}{2}+\theta}^{\frac{\theta}{2}} r^2\cos^2(\phi)\exp\left(-\frac{r^2\cos^2(\phi-\theta)\min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&+ \int_{r=0}^{\infty}\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} r^2\cos^2(\phi)\exp\left(-\frac{r^2\cos^2(\phi)\min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \qquad (64) \\
&= \int_{\phi=-\frac{\pi}{2}+\theta}^{\frac{\theta}{2}} \frac{1}{2\pi}\cos^2(\phi)\frac{2}{(1+\cos^2(\phi-\theta)\min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)^2)^2}d\phi \\
&+ \int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \frac{1}{2\pi}\cos^2(\phi)\frac{2}{(1+\cos^2(\phi)\min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)^2)^2}d\phi \qquad (65) \\
&= \int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \frac{1}{2\pi}[\cos^2(\phi-\theta)+\cos^2(\phi)]\frac{2}{(1+\cos^2(\phi)\min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)^2)^2}d\phi, \qquad (66)
\end{aligned}$$

where step (64) follows from the bound on $\min(\langle X,\boldsymbol{\beta}\rangle, \langle X,\boldsymbol{\beta}^*\rangle)$ in (59), and step (65) holds by integrating over r. Finally, step (66) holds by change of variable. In a similar fashion, we can bound

$M_1^{2,2}$.

$$
\begin{aligned}
M_1^{2,2} &= \left[\mathbb{E}_{X\sim\mathcal{N}(0,I)} 1_{\langle X,\boldsymbol{\beta}\rangle>0,\langle X,\boldsymbol{\beta}^*\rangle>0} XX^\top \exp\left(-\frac{\min(\langle X,\boldsymbol{\beta}\rangle,\langle X,\boldsymbol{\beta}^*\rangle)^2}{2}\right)\right]_{22} \\
&= \int_{r=0}^{\infty}\int_{\phi=-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}} r^2\sin^2(\phi)\exp\left(-\frac{\min(\|\boldsymbol{\beta}\|r\cos(\phi-\theta),\|\boldsymbol{\beta}^*\|r\cos(\phi))^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&\le \int_{r=0}^{\infty}\int_{\phi=-\frac{\pi}{2}+\theta}^{\frac{\theta}{2}} r^2\sin^2(\phi)\exp\left(-\frac{r^2\cos^2(\phi)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&+ \int_{r=0}^{\infty}\int_{\phi=-\frac{\theta}{2}}^{\frac{\pi}{2}} r^2\sin^2(\phi)\exp\left(-\frac{r^2\cos^2(\phi-\theta)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&= \int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}}\frac{1}{2\pi}[\sin^2(\phi-\theta)+\sin^2(\phi)]\frac{2}{(1+\cos^2(\phi)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2)^2}d\phi. \qquad (67)
\end{aligned}
$$

Finally, to obtain a bound for $|M_1^{1,2}|$, we upper bound both $M_1^{1,2}$ and $-M_1^{1,2}$:

$$
\begin{aligned}
M_1^{1,2} &= \left[\mathbb{E}_{X\sim\mathcal{N}(0,I)} 1_{\langle X,\boldsymbol{\beta}\rangle>0,\langle X,\boldsymbol{\beta}^*\rangle>0} XX^\top \exp\left(-\frac{\min(\langle X,\boldsymbol{\beta}\rangle,\langle X,\boldsymbol{\beta}^*\rangle)^2}{2}\right)\right]_{12} \\
&= \int_{r=0}^{\infty}\int_{\phi=-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}} r^2\sin(\phi)\cos(\phi)\exp\left(-\frac{\min(\|\boldsymbol{\beta}\|r\cos(\phi-\theta),\|\boldsymbol{\beta}^*\|r\cos(\phi))^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&\le \int_{r=0}^{\infty}\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} r^2\sin(\phi)\cos(\phi)\exp\left(-\frac{r^2\cos^2(\phi)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \qquad (68) \\
&+ \int_{r=0}^{\infty}\int_{\phi=0}^{\frac{\theta}{2}} r^2\sin(\phi)\cos(\phi)\exp\left(-\frac{r^2\cos^2(\phi-\theta)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&+ \int_{r=0}^{\infty}\int_{\phi=-\frac{\pi}{2}+\theta}^{0} r^2\sin(\phi)\cos(\phi)\exp\left(-\frac{r^2\cos^2(\phi)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2}{2}\right)\frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)rdrd\phi \\
&= \int_{\phi\in(-\frac{\pi}{2}+\theta,0)\cup(\frac{\theta}{2},\frac{\pi}{2})}\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2)^2}\sin(\phi)\cos(\phi)d\phi \\
&+ \int_{\phi=0}^{\frac{\theta}{2}}\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi-\theta)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2)^2}\sin(\phi)\cos(\phi)d\phi. \qquad (69)
\end{aligned}
$$

Note that in the above bound, the sign of $\sin(\phi)\cos(\phi)$ differs between region $(-\frac{\pi}{2},0)$ and $(0,\frac{\pi}{2})$. Similarly, for $-M_1^{1,2}$, we have

$$
\begin{aligned}
-M_1^{1,2} \le= &\int_{\phi\in(-\frac{\pi}{2}+\theta,0)\cup(\frac{\theta}{2},\frac{\pi}{2})} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi-\theta)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2)^2}\sin(\phi)\cos(\phi)d\phi \\
&- \int_{\phi=0}^{\frac{\theta}{2}}\frac{1}{2\pi}\frac{2}{(1+\cos^2(\theta)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2)^2}\sin(\phi)\cos(\phi)d\phi. \qquad (70)
\end{aligned}
$$

The next step is to provide upper bounds for those integrals, (66),(67), (69) and (70). The final bounds for $M_1^{1,1}$, $M_1^{2,2}$ and $M_1^{1,2}$ are established in Lemma 17, Lemma 18 and Lemma 19 respectively. ■

Lemma 17 (M_1, $(1,1)^{\text{th}}$ entry) *Let $\tau = \min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)$. Suppose $\theta \leq \frac{\pi}{8}$, the following holds:*

$$\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \cos^2(\phi-\theta)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \leq \frac{\sin(\theta)}{2} + \frac{\pi}{8}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right),$$
$$\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \cos^2(\phi)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \leq \frac{\pi}{8}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right).$$

Hence, $M_1^{1,1} \leq \frac{1}{4}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right) + \frac{\sin(\theta)}{2\pi}$.

Proof We divide the region $(\frac{\theta}{2}, \frac{\pi}{2})$ into two parts, $(\frac{\theta}{2}, \gamma) \cup (\gamma, \frac{\pi}{2})$ for some $\gamma > \frac{\theta}{2}$. In the first part, we bound $\frac{1}{(1+\cos^2(\phi)\tau^2)^2}$ by $\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}$, and in the second part, we bound $\frac{1}{(1+\cos^2(\phi)\tau^2)^2}$ by 1. It then follows that:

$$\begin{aligned}
&\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \cos^2(\phi-\theta)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \\
\leq &\int_{\phi=\frac{\theta}{2}}^{\gamma} \cos^2(\phi-\theta)\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}d\phi + \int_{\phi=\gamma}^{\frac{\pi}{2}} \cos^2(\phi-\theta)d\phi \\
= &\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}\left[\frac{1}{2}(\gamma-\frac{\theta}{2}) + \frac{1}{4}(\sin(2\gamma-2\theta)+\sin(\theta))\right] + \left[\frac{1}{2}(\frac{\pi}{2}-\gamma) + \frac{1}{4}(\sin(\pi-2\theta)-\sin(2\gamma-2\theta))\right] \\
\leq &\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}\frac{1}{2}\gamma + \frac{1}{2}(\frac{\pi}{2}-\gamma) + \frac{1}{4}\sin(2\theta) + \frac{1}{4}\sin(\theta)
\end{aligned}$$

The last step holds since $\sin(2\gamma-2\theta) > 0$. By picking $\gamma = \frac{\pi}{4}$, the above bound becomes:

$$\frac{\pi}{8}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{4}\sin(2\theta) + \frac{1}{4}\sin(\theta).$$

Similarly,

$$\begin{aligned}
&\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \cos^2(\phi)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \\
\leq &\int_{\phi=\frac{\theta}{2}}^{\gamma} \frac{1}{(1+\cos^2(\gamma)\tau^2)^2}\cos^2(\phi)d\phi + \int_{\phi=\gamma}^{\frac{\pi}{2}} \cos^2(\phi)d\phi \\
= &\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}\left[\frac{1}{2}(\gamma-\frac{\phi}{2}) + \frac{1}{4}(\sin(2\gamma)-\sin(\phi))\right] + \left[\frac{1}{2}(\frac{\pi}{2}-\gamma) + \frac{1}{4}(\sin(\pi)-\sin(2\gamma))\right] \\
\leq &\frac{\pi}{8}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right) - \frac{1}{4}\sin(\phi),
\end{aligned}$$

where we again pick $\gamma = \frac{\pi}{4}$ in the last step.

Therefore, by adding the above two bounds together, we show that

$$\begin{aligned} M_1^{1,1} &= \frac{1}{\pi}\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} [\cos^2(\phi-\theta)+\cos^2(\phi)]\frac{1}{(1+\cos^2(\phi)\min(\|\boldsymbol{\beta}\|,\|\boldsymbol{\beta}^*\|)^2)^2}d\phi \\ &\leq \frac{1}{4}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right)+\frac{1}{2\pi}\sin(\theta). \end{aligned}$$

■

Lemma 18 (M_1, $(2,2)^{\text{th}}$ entry) *Let $\tau = \min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)$. Suppose $\theta \leq \frac{\pi}{8}$, the following holds:*

$$\begin{aligned} &\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \sin^2(\phi-\theta)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \\ \leq &\frac{\pi}{8}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right)+\frac{1}{4}\left(1-\frac{1}{(1+0.5\tau^2)^2}\right)\cos(2\theta)-\frac{1}{4}\sin(2\theta)-\frac{1}{4(1+0.5\tau^2)^2}\sin(\theta), \\ &\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \sin^2(\phi)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \\ \leq &\frac{\pi}{8}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right)+\frac{1}{4}\left(1-\frac{1}{(1+0.5\tau^2)^2}\right)+\frac{1}{4}\frac{1}{(1+0.5\tau^2)^2}\sin(\theta). \end{aligned}$$

Hence, $M_1^{2,2} \leq \frac{1}{4}\left(1+\frac{1}{(1+0.5\tau^2)}\right)+\frac{1}{2\pi}\left(1-\frac{1}{(1+0.5\tau^2)^2}\right)$.

Proof The method is similar as before where we divide the region $(\frac{\theta}{2}, \frac{\pi}{2})$ into two parts, $(\frac{\theta}{2}, \gamma)\cup(\gamma, \frac{\pi}{2})$ for some $\gamma > \frac{\theta}{2}$.

$$\begin{aligned} &\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \sin^2(\phi-\theta)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \\ \leq &\int_{\phi=\frac{\theta}{2}}^{\gamma} \sin^2(\phi-\theta)\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}d\phi+\int_{\phi=\gamma}^{\frac{\pi}{2}}\sin^2(\phi-\theta)d\phi \\ = &\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}\left[\frac{1}{2}(\gamma-\frac{\theta}{2})-\frac{1}{4}(\sin(2\gamma-2\theta)+\sin(\theta))\right]+\frac{1}{2}\left(\frac{\pi}{2}-\gamma\right)-\frac{1}{4}\left[\sin(\pi-2\theta)-\sin(2\gamma-2\theta)\right] \\ = &\frac{\pi}{4}\frac{1}{2}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right)+\frac{1}{4}\left(1-\frac{1}{(1+0.5\tau^2)^2}\right)\cos(2\theta)-\frac{1}{4}\sin(2\theta)-\frac{1}{4(1+0.5\tau^2)^2}\sin(\theta), \end{aligned}$$

where we pick $\gamma = \frac{\pi}{4}$ in the last step.
Similarly,

$$
\begin{aligned}
&\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \sin^2(\phi)\frac{1}{(1+\cos^2(\phi)\tau^2)^2}d\phi \\
\leq &\int_{\phi=\frac{\theta}{2}}^{\gamma} \sin^2(\theta)\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}d\phi + \int_{\phi=\gamma}^{\frac{\pi}{2}} \sin^2(\phi)d\phi \\
= &\frac{1}{(1+\cos^2(\gamma)\tau^2)^2}\left[\frac{1}{2}(\gamma-\frac{\theta}{2}) - \frac{1}{4}(\sin(2\gamma)-\sin(\theta))\right] + \frac{1}{2}(\frac{\pi}{2}-\gamma) - \frac{1}{4}\left[\sin(\pi)-\sin(2\gamma)\right] \\
= &\frac{\pi}{4}\frac{1}{2}\left(1+\frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{4}\left(1-\frac{1}{(1+0.5\tau^2)^2}\right) + \frac{1}{4}\frac{1}{(1+0.5\tau^2)^2}\sin(\theta),
\end{aligned}
$$

where we again pick $\gamma = \frac{\pi}{4}$ in the last step. Therefore,

$$
\begin{aligned}
M_1^{2,2} = &\int_{\phi=\frac{\theta}{2}}^{\frac{\pi}{2}} \frac{1}{2\pi}[\sin^2(\phi-\theta)+\sin^2(\phi)]\frac{2}{(1+\cos^2(\phi)\min(\|\beta\|,\|\beta_*\|)^2)^2}d\phi. \\
&\frac{1}{4}\left(1+\frac{1}{(1+0.5\tau^2)}\right) + \frac{1}{2\pi}\left(1-\frac{1}{(1+0.5\tau^2)^2}\right).
\end{aligned}
$$

■

Lemma 19 ($M_1(1,2)^{\text{th}}$ entry) *Let $\tau = \min(\|\boldsymbol{\beta}\|, \|\boldsymbol{\beta}^*\|)$. Suppose $\theta \leq \frac{\pi}{8}$, the following holds:*

$$
\begin{aligned}
M_1^{1,2} \leq &\sin^2(\theta)\left[\frac{1}{2\pi(1+\cos^2(\theta)\tau^2)} + \frac{1}{2\pi(1+\tau^2\sin^2(\theta))}\right], \\
-M_1^{1,2} \leq &\frac{1}{\pi}\sin(\theta) + \frac{1}{2\pi}\frac{\sin^2(\theta)}{1+\tau^2\sin^2(\theta)}.
\end{aligned}
$$

Proof In (69), we know that

$$
\begin{aligned}
M_1^{1,2} \leq &\int_{\phi\in(-\frac{\pi}{2}+\theta,0)\cup(\frac{\theta}{2},\frac{\pi}{2})} \frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\sin(\phi)\cos(\phi)d\phi \\
&+ \int_{\phi=0}^{\frac{\theta}{2}} \frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi-\theta)\tau^2)^2}\sin(\phi)\cos(\phi)d\phi.
\end{aligned}
$$

It remains to bound the right hand side.

$$\begin{aligned}
&\int_{\phi\in(-\frac{\pi}{2}+\theta,0)\cup(\frac{\theta}{2},\frac{\pi}{2})} \frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\sin(\phi)\cos(\phi)d\phi + \int_{\phi=0}^{\frac{\theta}{2}} \frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi-\theta)\tau^2)^2}\sin(\phi)\cos(\phi)d\phi \\
&\leq \frac{1}{\pi(1+\cos^2(\theta)\tau^2)}\int_{\phi=0}^{\frac{\theta}{2}}\sin(\phi)\cos(\phi)d\phi + \frac{1}{2\pi}\left[\frac{-\cos^2(\theta)}{(1+\tau^2)(1+\tau^2\sin^2(\theta))} + \frac{\cos^2(\theta)}{1+\cos^2(\theta)\tau^2}\right] \\
&\leq \frac{1}{\pi(1+\cos^2(\theta)\tau^2)}\int_{\phi=0}^{\frac{\theta}{2}}\sin(\phi)\cos(\phi)d\phi + \frac{1}{2\pi}\left[\frac{-\cos^2(\theta)}{(1+\tau^2)(1+\tau^2\sin^2(\theta))} + \frac{1}{1+\tau^2}\right] \\
&\leq \frac{\sin^2(\theta)}{2\pi(1+\cos^2(\theta)\tau^2)} + \frac{\sin^2(\theta)}{2\pi(1+\tau^2\sin^2(\theta))} \\
&\leq \sin^2(\theta)\left[\frac{1}{2\pi(1+\cos^2(\theta)\tau^2} + \frac{1}{2\pi(1+\tau^2\sin^2(\theta))}\right].
\end{aligned}$$

Next, let us look at the bound for $-M_1^{1,2}$ in (70). There are two terms, one is

$$T_1 := \int_{\phi\in(-\frac{\pi}{2}+\theta,0)\cup(\frac{\theta}{2},\frac{\pi}{2})} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi-\theta)\tau^2)^2}\sin(\phi)\cos(\phi)d\phi$$

and the other is:

$$T_2 := -\int_{\phi=0}^{\frac{\theta}{2}} \frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2}\sin(\phi)\cos(\phi)d\phi$$

When $\phi \in (0, \pi/8)$, $T_2 < 0$. For T_1, let us use change of variable and write the integral as:

$$\begin{aligned}
&\int_{\phi\in(-\frac{\pi}{2},-\theta)\cup(-\frac{\theta}{2},\frac{\pi}{2}-\theta)} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2}\sin(\phi+\theta)\cos(\phi+\theta)d\phi \\
=&\underbrace{\int_{\phi\in(-\frac{\pi}{2},-\theta)\cup(-\frac{\theta}{2},\frac{\pi}{2}-\theta)} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\sin(\phi)\cos(\phi)\cos(2\theta)d\phi}_{\text{Part1}} \\
&+\underbrace{\int_{\phi\in(-\frac{\pi}{2},-\theta)\cup(-\frac{\theta}{2},\frac{\pi}{2}-\theta)} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\sin(\theta)\cos(\theta)\cos(2\phi)d\phi}_{\text{Part2}}.
\end{aligned}$$

Note that the first part can be computed exactly as before,

$$\begin{aligned}
&\int_{\phi\in(-\frac{\pi}{2},-\theta)\cup(-\frac{\theta}{2},\frac{\pi}{2}-\theta)} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\sin(\phi)\cos(\phi)\cos(2\theta)d\phi \\
\leq&\int_{\phi\in(-\frac{\pi}{2},-\theta)\cup(-\theta,\frac{\pi}{2}-\theta)} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\sin(\phi)\cos(\phi)\cos(2\theta)d\phi \\
=&\frac{1}{2\pi}\left[\frac{\cos^2(\theta)}{1+\cos^2(\theta)\tau^2} - \frac{\cos^2(\theta)-\sin^2(\theta)}{(1+\tau^2\sin^2(\theta))(1+\tau^2\cos^2(\theta))}\right]\cos(2\theta) \\
\leq&\frac{1}{2\pi}\frac{\sin^2(\theta)}{1+\tau^2\sin^2(\theta)}.
\end{aligned} \tag{71}$$

The second part contains the factor $\sin(\theta)\cos(\theta)$, and it remains to bound:

$$\int_{\phi\in(-\frac{\pi}{2},-\theta)\cup(-\frac{\theta}{2},\frac{\pi}{2}-\theta)} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\cos(2\phi)d\phi.$$

Note that the intergrand is an even function in ϕ. Moreover, when $|\phi|<\frac{\pi}{4}$, the intergrand is negative, and when $|\phi|\in(\frac{\pi}{4},\frac{\pi}{2})$, the integrand is positive. Thus the integral can be further upper bounded by:

$$\begin{aligned}
&2\int_{\phi\in(\frac{\pi}{4},\frac{\pi}{2})} -\frac{1}{2\pi}\frac{2}{(1+\cos^2(\phi)\tau^2)^2}\cos(2\phi)d\phi\\
\leq&\frac{2}{\pi}\int_{\phi=\frac{\pi}{4}}^{\frac{\pi}{2}} -\cos(2\phi)d\phi\\
=&\frac{1}{\pi}.
\end{aligned}$$

Combining the bound on two parts, we obtain:

$$\begin{aligned}
-M_1^{1,2} =&T_1+T_2\\
\leq&\frac{1}{2\pi}\frac{\sin^2(\theta)}{1+\tau^2\sin^2(\theta)}+\frac{1}{\pi}\sin(\theta).
\end{aligned}$$

■

Lemma 20 (M_2) *The entries of the symmetric 2×2 matrix M_2 are the following:*

$$\begin{aligned}
M_2^{1,1} =&\frac{\theta}{\pi}-\frac{\sin(2\theta)}{4\pi}\\
M_2^{2,2} =&\frac{\theta}{\pi}+\frac{\sin(2\theta)}{4\pi}\\
M_2^{1,2} =&-\frac{\sin^2(\theta)}{\pi}.
\end{aligned}$$

Proof It is a simple calculation.

$$\begin{aligned}
M_2^{1,1} =&\frac{1}{\pi}\int_{r>0}\int_{\phi\in(-\frac{\pi}{2},-\frac{\pi}{2}+\theta)} r^2\cos^2(\phi)\exp\left(-\frac{r^2}{2}\right)rdrd\phi\\
=&\frac{2}{\pi}\int_{\phi\in(-\frac{\pi}{2},-\frac{\pi}{2}+\theta)}\cos^2(\phi)d\phi=\frac{\theta}{\pi}-\frac{\sin(2\theta)}{2\pi}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
M_2^{2,2} =&\frac{1}{\pi}\int_{r>0}\int_{\phi\in(-\frac{\pi}{2},-\frac{\pi}{2}+\theta)} r^2\sin^2(\phi)\exp\left(-\frac{r^2}{2}\right)rdrd\phi\\
=&\frac{2}{\pi}\int_{\phi\in(-\frac{\pi}{2},-\frac{\pi}{2}+\theta)}\sin^2(\phi)d\phi=\frac{\theta}{\pi}+\frac{\sin(2\theta)}{2\pi}.
\end{aligned}$$

For the cross term,

$$\begin{aligned} M_2^{1,2} =& \frac{1}{\pi}\int_{r>0}\int_{\phi\in(-\frac{\pi}{2},-\frac{\pi}{2}+\theta)} r^2\sin(\theta)\cos(\theta)\exp\left(-\frac{r^2}{2}\right) r dr d\phi \\ =& \frac{1}{\pi}\int_{\phi\in(-\frac{\pi}{2},-\frac{\pi}{2}+\theta)} \sin(2\phi) d\phi = \frac{\sin^2(\theta)}{\pi}. \end{aligned}$$

■

Lemma 21 (Bound on Spectral norm of a 2×2 matrix) *Let M be a symmetric 2×2 matrix*

$$M = \begin{bmatrix} a & c \\ c & b \end{bmatrix}.$$

Suppose $a, b > 0$, the spectral norm of M is bounded by $\max(a,b) + |c|$.

Proof The characteristic polynomial for the matrix is:

$$p(x) = x^2 - (a+b)x + ab - c^2.$$

It has two roots: $x_1 = \frac{a+b+\sqrt{(a-b)^2+4c^2}}{2}$ and $x_2 = \frac{a+b-\sqrt{(a-b)^2+4c^2}}{2}$. When $a, b > 0$, the larger root is upper bounded by $\frac{a+b+|a-b|+2|c|}{2}$, which is dominated by $\max(a,b)+|c|$. ■

5. Proofs for Auxiliary Results

5.1. Upper Bound for Norm

Lemma 22 (Bounded population EM iterates) *For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have*

$$\|\boldsymbol{\beta}'\| \le 3\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}. \tag{72}$$

Proof From Lemma 11, we know that $b_1' \le b_1^* + \frac{2}{\pi}\sqrt{\sigma^2 + b_2^{*2}}$. On the other side, from lemma 10 we have $b_2' \le b_2^*$. Therefore,

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

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

5.2. Lower Bound for Norm

Lemma 23 *If $\|\boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}^*\|/10$, then after one finite-sample EM update with $n = O(\max(1, poly(\eta^{-2}))(d/\epsilon^2))$ samples, $\|\tilde{\boldsymbol{\beta}}'\| \geq \|\boldsymbol{\beta}^*\|/10$.*

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

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

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

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

However,

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

Simiarly,

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

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

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

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

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

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

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

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

$$\begin{aligned}
&\frac{2}{\pi}\sigma_2 \int_0^\infty \int_0^\infty \alpha_1 y \tanh(\frac{b_1}{\sigma}\alpha_1 y) e^{-\frac{y^2}{2}} e^{-\frac{\alpha_1^2}{2}} d\alpha_1 dy \\
&\quad \geq \frac{2}{\pi}\sigma_2 \int_0^\infty \int_0^\infty \alpha_1 y (\frac{b_1}{\sigma}\alpha_1 y - \frac{1}{3}(\frac{b_1}{\sigma}\alpha_1 y)^3) e^{-\frac{y^2}{2}} e^{-\frac{\alpha_1^2}{2}} d\alpha_1 dy \\
&\quad = b_1 \frac{\sigma_2}{\sigma}(1 - 3\frac{b_1^2}{\sigma^2}) \geq b_1 \sqrt{1+\frac{24}{25}\eta^2}(1 - 3\frac{b_1^2}{\sigma^2}).
\end{aligned} \tag{73}$$

If $\eta = \frac{\|\boldsymbol{\beta}^*\|}{\sigma} \geq 5$, then since we assumed $\frac{b_1}{\sigma} < 1/2$, we have $b_1\sqrt{1+\frac{24}{25}\eta^2}(1-3\frac{b_1^2}{\sigma^2}) \geq \frac{5}{4}b_1$. Otherwise, suppose $b_1 = \|\boldsymbol{\beta}^*\|/10$, then we have $b_1' \geq b_1\sqrt{1+\frac{24}{25}\eta^2}(1-\frac{3}{100}\eta^2)$. When $1 \leq \eta \leq 5$, we have $b_1' \geq \frac{5}{4}b_1$. When $0 \leq \eta \leq 1$, we have $b_1' \geq b_1(1+0.3\eta^2)$. Since by (47) we know b_1' is increasing as b_1 increases, and $\|\boldsymbol{\beta}'\| \geq b_1'$. Therefore, we conclude that sufficiently ϵ guarantees $\|\tilde{\boldsymbol{\beta}}'\| \geq \frac{\|\boldsymbol{\beta}^*\|}{10}$.

$\sin\theta \leq 0.2$*:* Assume $b_1 = \frac{\|\boldsymbol{\beta}^*\|}{10} < \frac{\sigma^2}{\sigma_2^2}b_1^*$. Otherwise we can do as in the first case. From equation (50), we have

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

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

5.3. Concentration Result in One Direction

Theorem 13 *Consider one iteration of sample-based EM algorithm. There exist absolute constants $c_1, c_2 > 0$, such that statistical error in a fixed direction $\boldsymbol{\beta}^*$ can be bounded with probability at least $1-\delta$, by*

$$|\langle \tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', \boldsymbol{\beta}^* \rangle| \leq \sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\left(c_1\sqrt{\frac{1}{n}\log(1/\delta)} + c_2\frac{d}{n}\log(1/\delta)\right). \tag{74}$$

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

$$\tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}' = \underbrace{(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}}_{\hat{\Sigma}^{-1}} \underbrace{(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i \tanh(y_i\langle \boldsymbol{x}_i, \beta\rangle/\sigma^2))}_{\hat{\mu}} - \underbrace{\mathbb{E}[yX\tanh(y\langle X,\beta\rangle/\sigma^2)]}_{\mu}. \tag{75}$$

Now we fix some $v \in R^d$ such that $\|v\| = 1$, and give a bound for $|\langle \tilde{\boldsymbol{\beta}}' - \beta, v\rangle|$. First observe that,

$$\begin{aligned}|\langle \tilde{\boldsymbol{\beta}}' - \boldsymbol{\beta}', v\rangle| &= |(\hat{\Sigma}^{-1}\hat{\mu} - \mu)^\top v| \\ &= |(\hat{\mu} - \mu)^\top v + \mu^\top(\hat{\Sigma}^{-1} - I)v + (\hat{\mu} - \mu)^\top(\hat{\Sigma}^{-1} - I)v| \\ &\leq |\underbrace{(\hat{\mu} - \mu)^\top v}_{A}| + |\underbrace{\mu^\top(\hat{\Sigma}^{-1} - I)v}_{B}| + |\underbrace{(\hat{\mu} - \mu)^\top(\hat{\Sigma}^{-1} - I)v}_{C}|.\end{aligned}$$

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

Bounding A: The product of two sub-Gaussian random variables is sub-exponential, which can be easily shown with the notion of sub-Gaussian norm and sub-exponential norm Vershynin (2010).

The random variable $y_i\langle \boldsymbol{x}_i, v\rangle \tanh(y_i\langle \boldsymbol{x}_i, \beta\rangle/\sigma^2))$ is sub-exponential with parameter $C\sqrt{\sigma^2 + \|\beta^*\|^2}$ for some constant C, since $|\tanh(\cdot)| \leq 1$, $\langle \boldsymbol{x}_i, v\rangle$ is sub-Gaussian with parameter 1, and y_i is sub-Gaussian with parameter at most $\sqrt{\sigma^2 + \|\beta^*\|^2}$.

Applying the concentration inequality for sub-exponential random variable, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i y_i\langle \boldsymbol{x}_i, v\rangle \tanh\left(y_i\frac{\langle \boldsymbol{x}_i, \beta\rangle}{\sigma^2}\right) - \mathbb{E}\left[yX\tanh\left(y\frac{\langle X, \beta\rangle}{\sigma^2}\right)\right]\right| \geq t\right) \leq \exp\left(-\frac{nt^2}{K(\sigma^2 + \|\beta^*\|^2)}\right),$$

for some absolute constant K.

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

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

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

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

from where we can see $\mu^\top(\hat{\Sigma}^{-1} - I)v = \mu^\top(I - \hat{\Sigma})v + \|\mu\|\tilde{O}(\frac{d}{n})$.

For simplicity, let us define $u = \frac{\mu}{\|\mu\|}$ and derive a bound for $u^\top(I - \hat{\Sigma})v$. Now we are left with bounding $u^\top(I - \hat{\Sigma})v = u^\top v - \frac{1}{n}\sum_i(\boldsymbol{x}_i^\top u)(\boldsymbol{x}_i^\top v)$. Let two random variables $Z_1 = X^\top u$, $Z_2 = X^\top v$. Since Z_1, Z_2 are sub-Gaussian random variables with parameter 1, Z_1Z_2 is sub-exponential with parameter at most 2. Thus, we get the sub-exponential concentration bound $u^\top(I - \hat{\Sigma})v \leq \tilde{O}(\sqrt{\frac{1}{n}})$. This yields $B \leq c_2\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}(\sqrt{\frac{1}{n}\log(1/\delta)} + \frac{d}{n})$ since $\|\mu\| \leq O(\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2})$ due to Lemma 22.

Bounding C: We have $\|\hat{\mu} - \mu\| \leq c_5\sqrt{\sigma^2 + \|\boldsymbol{\beta}^*\|^2}\sqrt{\frac{d}{n}\log(1/\delta)}$ from Balakrishnan et al. (2017) with probability at least $1 - \delta$, as well as $\|\hat{\Sigma}^{-1} - I\|_{op} \leq c_2\sqrt{\frac{d}{n}\log(1/\delta)}$. Therefore, we get

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

This gives a bound for C.

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