Supplement to "On the Minimax Optimality of the EM Algorithm for Learning Two-Component Mixed Linear Regression"

In the supplementary material, we collect proofs and results deferred from the main text. In Appendix A, we denote key notation used in the proofs. We provide proof for Theorem 1 in Appendix B while the proof for Theorem 2 is presented in Appendix C. The proofs of auxiliary lemmas used to prove the main theorems are in Appendices D and E. Finally, in Appendix F, we provide a result justifying the initialization with spectral methods in Theorem 1, the proof for super-linear convergence of population EM operator under very high SNR regime, and the proof for the convergence rate of EM algorithm when the variance of regression noise is unknown.

A Additional Notations

We sometimes use the transformed coordinate where the first two coordinate spans θ and θ^*. That is, let $\{v_1, ..., v_d\}$ be standard basis in the transformed coordinate such that $v_1 = \theta/\|\theta\|$, and $\mathbf{span}(v_1, v_2) = \mathbf{span}(\theta, \theta^*)$. Since Gaussian distribution is invariant to rotation, we often work on the transformed space in the proofs. Let $\alpha = \angle(\theta, \theta^*)$, $\eta = \|\theta^*\|/\sigma^*$, and $\sigma_2^2 = 1 + \|\theta^*\|^2 \sin^2\alpha$.

We define a few more quantities to simplify the notations throughout the proofs. Let x_1, x_2 be $X^\top v_1, X^\top v_2$ respectively. Following the notation in Kwon et al. (2019), we denote $b_1^* = {\theta^*}^\top v_1 = \|\theta^*\| \cos\angle(\theta, \theta^*)$, and $b_2^* = {\theta^*}^\top v_2 = \|\theta^*\| \sin\angle(\theta, \theta^*)$. Note that in this transformed coordinate, due to the symmetry of the distribution, $M_{mlr}(\theta)^\top v_j = 0$ for all $j \geq 3$. Hence we focus on bounding the values in first two coordinates.

Using the coordinate transformation and new notations defined here, we can write the population operator in new coordinate as:

$$\begin{aligned} M_{mlr}(\theta) &= \mathbb{E}_{X,Y}\left[\tanh(YX^\top\theta)YX\right] \\ &= \mathbb{E}_{x_1,x_2,y}\left[\tanh(yx_1\|\theta\|)x_1 y\right] v_1 + \mathbb{E}_{x_1,x_2,y}\left[\tanh(yx_1\|\theta\|)x_2 y\right] v_2, \end{aligned} \tag{10}$$

where $y|(x_1, x_2) \sim \mathcal{N}(x_1 b_1^* + x_2 b_2^*, 1)$. Note that we simplify y as a single Gaussian due to the symmetry in the signs of y and Gaussian noise.

B Proof of Theorem 1

We first consider middle-to-high SNR regimes and then we consider low SNR regimes. In middle-to-high SNR regimes, we assume that we start from the initialization where $\cos\alpha \geq 0.95$. We note that the additional requirement $\|\theta_n^0\| \geq 0.9\|\theta^*\|$ is to prevent the analysis to become over-complicated (see Appendix C.3 for the arguments for starting from well-aligned small estimators).

We will frequently use the fact that $\|\theta^*\| \sin\alpha \leq \|\theta - \theta^*\|$. We can check that θ remains in this good initialization region using the convergence property of angles (see the arguments for sine values in Appendix C.3). Before getting into the detailed proof, we state some useful lemmas from previous work. We need the following lemma for the contraction rate of the population EM operator (5):

Lemma 3 (Theorem 4 in Kwon et al. (2019))**.** *Assume $\alpha < \pi/8$. Then, we have*

$$\|M_{mlr}(\theta) - \theta^*\| \leq \max\{\kappa, 0.6\}\|\theta - \theta^*\| + \kappa(16\sin^3\alpha)\|\theta^*\|\frac{\eta^2}{1+\eta^2}, \tag{11}$$

where $\kappa = \left(\sqrt{1 + \min\{\sigma_2^2\|\theta\|, \|\theta^*\|\cos\alpha\}^2/\sigma_2^2}\right)^{-1}$.

B.1 High SNR Regime

First, we arrange the sample operator as the following:

$$\begin{aligned}
M_{n,mlr}(\theta) - \theta^* &= \left(\frac{1}{n}\sum_i X_i X_i^\top\right)^{-1}\left(\frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta)\right) - \theta^* \\
&= \left(\frac{1}{n}\sum_i X_i X_i^\top\right)^{-1}\left(\frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta) - \frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta^*)\right. \\
&\qquad \left. + \frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta^*) - \frac{1}{n}\sum_i X_i X_i^\top \theta^*\right) \\
&= \left(\frac{1}{n}\sum_i X_i X_i^\top\right)^{-1}\left(\underbrace{\mathbb{E}_{X,Y}[XY\Delta_{(X,Y)}(\theta)]}_{:=A_1} + \underbrace{\frac{1}{n}\sum_i X_i Y_i \Delta_{(X_i,Y_i)}(\theta) - \mathbb{E}_{X,Y}[XY\Delta_{(X,Y)}(\theta)]}_{:=A_2}\right. \\
&\qquad \left. + \underbrace{\frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta^*) - \mathbb{E}_{Y_i|X_i}\left[\frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta^*)\right]}_{:=A_3}\right), \qquad (12)
\end{aligned}$$

where $\Delta_{(X,Y)}(\theta) := \tanh(YX^\top\theta) - \tanh(YX^\top\theta^*)$. In the term A_3, the expectation is taken over $Y_i|X_i \sim \frac{1}{2}\mathcal{N}(X_i^\top\theta^*, 1) + \frac{1}{2}\mathcal{N}(-X_i^\top\theta^*, 1)$, letting X_i fixed. Note that the true parameters are fixed points of the EM operators, and it is easy to check that the expectation in A_3 is equivalent to $\frac{1}{n}\sum_i X_i X_i^\top \theta^*$.

Now, we claim the following bounds with A_1, A_2, and A_3 in equation (12):

$$A_1 < 0.9\|\theta - \theta^*\|, \qquad (13)$$

$$A_2 \leq (\|\theta - \theta^*\| + 1)\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}, \qquad (14)$$

$$A_3 \leq C\sqrt{d\log(1/\delta)/n}, \qquad (15)$$

with probability at least $1 - 5\delta$. Here, C is some universal constant.

Assume that the above claims are given at the moment, we proceed to finish the proof of the convergence of EM algorithm under high SNR regime. In fact, plugging the results from equations (13), (14), and (15) into equation (12), we find that

$$\begin{aligned}
\|M_{n,mlr}(\theta) - \theta^*\| &\leq \left(0.9 + \sqrt{d\log^2(n\|\theta^*\|/\delta)/n}\right)\|\theta - \theta^*\| + C_1\sqrt{d\log^2(n\|\theta^*\|/\delta)/n} \\
&\leq \gamma\|\theta - \theta^*\| + C_1\sqrt{d\log^2(n\|\theta^*\|/\delta)/n},
\end{aligned}$$

for some $\gamma < 1$. From here, let $\epsilon_n := C_1\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}$ and we iterate over t to bound the estimation error in t^{th} step:

$$\begin{aligned}
\|\theta_n^{t+1} - \theta^*\| &\leq \gamma\|\theta_n^t - \theta^*\| + \epsilon_n \leq \gamma^2\|\theta_n^{t-1} - \theta^*\| + (1+\gamma)\epsilon_n \\
&\leq \ldots \leq \gamma^t\|\theta_n^0 - \theta^*\| + \frac{1}{1-\gamma}\epsilon_n.
\end{aligned}$$

After $t \geq c_1\log(n\|\theta^*\|/d)$ iterations, we have $\|\theta_n^t - \theta^*\| \leq c_2\sqrt{d/n}$ where c_1 and c_2 are universal constants. As a consequence, we reach the conclusion of the theorem for high SNR regime.

Proof of claim (13): In order to bound A_1, we can use the result of Corollary 1 in Appendix B.2. Observe that

$$\mathbb{E}\left[XY\tanh(YX^\top\theta^*)\right] = \theta^*,$$

$$\mathbb{E}\left[XY\tanh(YX^\top\theta)\right] = M_{mlr}(\theta).$$

From Corollary 1, we conclude that

$$A_1 = \mathbb{E}_{X,Y}[XY\Delta_{(X,Y)}(\theta)] < 0.9\|\theta-\theta^*\|.$$

Therefore, we reach the conclusion of claim (13).

Proof of claim (14): Next, we bound A_2. We first discretize the parameter space for θ as the following:

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\theta\in\mathbb{B}(\theta^*,r)}\|\frac{1}{n}\sum_{i=1}^n X_iY_i\Delta_i(\theta) - \mathbb{E}[XY\Delta(\theta)]\| \geq t\Big)\\
&\quad= \underbrace{\mathbb{P}\left(\sup_{j\in[\mathcal{N}_\epsilon]}\|\frac{1}{n}\sum_{i=1}^n X_iY_i\Delta_i(\theta_j) - \mathbb{E}[X_iY_i\Delta_i(\theta_j)]\| \geq t/2\right)}_{\text{finite-sample error}}\\
&\quad+ \underbrace{\mathbb{P}\left(\sup_{\|\theta-\theta'\|\leq\epsilon}\|\frac{1}{n}\sum_{i=1}^n X_iY_i(\Delta_i(\theta)-\Delta_i(\theta')\| + \|\mathbb{E}\left[XY(\Delta(\theta)-\Delta(\theta'))\right]\| \geq t/2\right)}_{\text{discretization error}},
\end{aligned}$$

where $\Delta_i(\theta)$ is a shorthand for $\Delta_i(\theta) := \tanh(Y_iX_i^\top\theta) - \tanh(Y_iX_i^\top\theta^*)$, $\Delta(\theta)$ is a shorthand for $\Delta(\theta) = \tanh(YX^\top\theta) - \tanh(YX^\top\theta^*)$, $\mathcal{N}_\epsilon$ is ϵ-covering number of $\mathbb{B}(\theta^*, r)$, and $\{\theta_j, j\in[\mathcal{N}_\epsilon]\}$ is the corresponding ϵ-covering set.

The discretization error can be bounded by the Lipschitz continuity of the function Δ_i, namely, $|\Delta_i(\theta)-\Delta_i(\theta')| \leq |Y_i||X^\top\theta - X^\top\theta'|$ for all θ, θ'. It follows that

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n X_iY_i(\Delta_i(\theta)-\Delta_i(\theta')\| &\leq \|\frac{1}{n}\sum_{i=1}^n Y_i^2X_iX_i^\top(\theta-\theta')\|\\
&\leq \epsilon\|\!|\frac{1}{n}\sum_{i=1}^n Y_i^2X_iX_i^\top\|\!|_{\text{op}}.
\end{aligned}$$

Note that $\mathbb{E}[Y^2XX^\top] = I + 2\theta^*\theta^{*\top}$, hence $\|\!|\mathbb{E}[Y^2XX^\top]\|\!|_{\text{op}} \leq 2\|\theta^*\|^2+1$. Furthermore, from Lemma 10, we have $\|\!|\frac{1}{n}\sum_{i=1}^n Y_i^2X_iX_i^\top\|\!|_{\text{op}} \leq 3\|\theta^*\|^2$ with probability at least $1-\delta$. We conclude that

$$\text{discretization error} \leq 6\epsilon\|\theta^*\|^2$$

with probability at least $1-\delta$.

In order to bound the finite-sample error for each fixed θ_j, we adopt the per-sample decomposition argument used in the previous works (Kwon and Caramanis, 2020b) and (Kwon and Caramanis, 2020a). In order to simplify the notation, let Z_i be the noise such that $Y_i = \nu_iX_i^\top\theta^* + Z_i$ where ν_i is an independent Rademacher variable. We define good events as follows:

$$\begin{aligned}
\mathcal{E}_1 &= \{2|X^\top(\theta^*-\theta)| \leq |X^\top\theta^*|\},\\
\mathcal{E}_2 &= \{|X^\top\theta^*| \geq 2\tau\},\\
\mathcal{E}_3 &= \{|Z| \leq \tau\},
\end{aligned}$$

where we decide τ later. Let the good event $\mathcal{E}_{good} := \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$. Then we have a following lemma:

Lemma 4. *Under the event $\mathcal{E}_{good}$, we have*

$$|\Delta_{(X,Y)}(\theta)| \leq \exp(-\tau^2).$$

Proof. Without loss of generality, let $\nu = +1$. We can check that

$$YX^\top\theta = (\nu X^\top\theta^* + Z)(X^\top\theta^*) + (\nu X^\top\theta^* + Z)(X^\top(\theta-\theta^*))$$

$$= (\nu X^\top \theta^* + Z)(X^\top \theta^* + X^\top(\theta - \theta^*))$$
$$\geq \tau \cdot \tau = \tau^2.$$

Since $\tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)} \geq 1 - \exp(-x)$ for $x \geq 0$, we have $\tanh(YX^\top\theta) \geq 1 - \exp(-\tau^2)$. Similarly, $\tanh(YX^\top\theta^*) \geq 1 - \exp(-\tau^2)$. On the other hand, $\tanh(x) \leq 1$ for all x. We can conclude that $\Delta_{(X,Y)}(\theta) \leq \exp(-\tau^2)$. For the other sign $\nu = -1$, we can show it similarly. □

To simplify the notation, we denote $W_i := \nu_i X_i X_i^\top \theta^* \Delta_i(\theta)$. Then, we can decompose A_2 as follows:

$$A_2 = \underbrace{\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i \Delta_i(\theta) - \mathbb{E}[XZ\Delta(\theta)]\right)}_{:=T_1} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^n W_i - \mathbb{E}[W]\right)}_{:=T_2}. \tag{16}$$

We first claim the following high probability bound with T_1:

$$\mathbb{P}\left(\|T_1\| \geq t\right) \leq \exp\left(-\frac{nt^2}{K_0} + K_0' d\right), \tag{17}$$

for some universal constants $K_0, K_0' > 0$, where we assumed $n \gg d$ to ignore sub-exponential tail part. The proof of claim (17) is deferred to the end of the proof of high SNR regime.

For the term T_2 in equation (16), we apply per-sample decomposition.

$$\frac{1}{n}\sum_i W_i - \mathbb{E}[W] = \frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_{good}} - \mathbb{E}[W 1_{\mathcal{E}_{good}}]) + \frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_1^c} - \mathbb{E}[W 1_{\mathcal{E}_1^c}])$$
$$+ \frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_1 \cap \mathcal{E}_2^c} - \mathbb{E}[W 1_{\mathcal{E}_1 \cap \mathcal{E}_2^c}]) + \frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3^c} - \mathbb{E}[W 1_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3^c}]).$$

In the sequel, we will show that

$$\mathbb{P}\left(\|\frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_{good}} - \mathbb{E}[W 1_{\mathcal{E}_{good}}])\| \geq t\right) \leq \exp\left(-\frac{nt^2}{K_1\|\theta^*\|^2 \exp(-2\tau^2)} + K_1' d\right), \tag{18}$$

$$\mathbb{P}\left(\|\frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_1^c} - \mathbb{E}[W 1_{\mathcal{E}_1^c}])\| \geq t\right) \leq \exp\left(-\frac{nt^2}{K_2\|\theta - \theta^*\|^2} + K_2' d\right), \tag{19}$$

$$\mathbb{P}\left(\|\frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_1 \cap \mathcal{E}_2^c} - \mathbb{E}[W 1_{\mathcal{E}_1 \cap \mathcal{E}_2^c}])\| \geq t\right) \leq \exp\left(-\frac{nt^2}{K_3\tau^2} + K_3' d\right), \tag{20}$$

$$\mathbb{P}\left(\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\frac{1}{n}\sum_i (W_i 1_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3^c} - \mathbb{E}[W 1_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3^c}])\| = 0\right) \geq 1 - \delta, \tag{21}$$

where $K_{(\cdot)}$ are all some universal constants. The last probability is due to our choice $\tau = \Theta(\sqrt{\log(n\|\theta^*\|/\delta)})$ such that no sample fall in the event $\mathcal{E}_3^c$ with probability at least $1 - \delta$. We set t and ϵ as follows:

$$t = O\left((\|\theta - \theta^*\| + 1)\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}\right),$$
$$\epsilon = O\left(\|\theta^*\|^{-2}\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}\right).$$

The overall finite-sample error term is bounded by taking union bound over ϵ-covering set. Note that $\log(\mathcal{N}_\epsilon) \leq c \cdot d\log(\|\theta^*\|)$ for some universal constant c. Hence the total probability of $\|T_2\| \geq t$ is dominated by

$$\exp\left(-\frac{nt^2}{K_2\|\theta - \theta^*\|^2} + K_2' d\log(n\|\theta^*\|/d)\right) + \exp\left(-\frac{nt^2}{K_3\tau^2} + K_3' d\log(n\|\theta^*\|/d)\right),$$

for some (new) constants $K_2, K_2', K_3, K_3' > 0$. Our choice of t gives 5δ total probability bound for the finite-sample error. We can conclude that $A_2 \leq t \leq (\|\theta - \theta^*\| + 1)\sqrt{d\log^2(n\|\theta^*\|/\delta)/n}$ with probability at least $1 - 5\delta$. Hence, we reach the conclusion of claim (14).

Proof of claim (15): Finally, for bounding A_3, we use Proposition 11 in Kwon et al. (2019) that exactly targets to bound this quantity.

Lemma 5 (Proposition 11 in Kwon et al. (2019)). *For each fixed θ, with probability at least $1 - \exp(-cn) - 6^d \exp(-nt^2/72)$,*

$$\|\frac{1}{n}\sum_i X_i Y_i \tanh(Y_i X_i^\top \theta) - \frac{1}{n}\sum_i \mathbb{E}_{Y_i|X_i}\left[Y_i X_i \tanh(Y_i X_i^\top \theta)\right]\| \le t, \tag{22}$$

for some absolute constant $c > 0$.

Applying the above lemma for $\theta = \theta^*$, we can show that $A_3 \le C\sqrt{d\log(1/\delta)/n}$ with probability at least $1-\delta$. As a consequence, we obtain claim (15).

Proof of Equation (17). We use the notion of sub-exponential Orcliz norm to bound (17). It is easy to see that $X_i Z_i \Delta_i$ is a sub-exponential random vector with Orcliz norm $O(1)$. Using the standard concentration result in Vershynin (2010), we get the result.

Proof of Equation (18). Similarly to the previous case, we need to bound the sub-exponential norm of the quantity:

$$\begin{aligned}
\left\|W_i 1_{\mathcal{E}_{good}}\right\|_{\psi_1} &= \sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\mathbb{E}\left[|(X_i^\top u)(X_i^\top\theta^*)\Delta_i 1_{\mathcal{E}_{good}}|^p\right]^{1/p} \\
&\le \exp(-\tau^2)\sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\mathbb{E}\left[|(X_i^\top u)(X_i^\top\theta^*)|^p\right]^{1/p} \\
&\le \exp(-\tau^2)\sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\sqrt{\mathbb{E}[(X^\top u)^{2p}]\mathbb{E}\left[(X_i^\top\theta^*)^{2p}\right]}^{1/p} \\
&\le K_0\|\theta^*\|\exp(-\tau^2).
\end{aligned}$$

We use the fact that $|\Delta_i(\theta)| \le \exp(-\tau^2)$ under the good event, Cauchy-Schwartz inequality, and p^{th}-order moments of Gaussian is $O((2p)^{p/2})$. Similarly using the result in Vershynin (2010), we have the equation (18).

Proof of Equation (19). We check the sub-exponential ψ_1-Orcliz norm again.

$$\begin{aligned}
\left\|W_i 1_{\mathcal{E}_1^c}\right\|_{\psi_1} &= \sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\mathbb{E}\left[|(X_i^\top u)(X_i^\top\theta^*)\Delta_i 1_{\mathcal{E}_1^c}|^p\right]^{1/p} \\
&\le \sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\mathbb{E}\left[|(X_i^\top u)(X_i^\top(\theta^*-\theta))|^p\right]^{1/p} \\
&\le K_1\|\theta^*-\theta\|,
\end{aligned}$$

from which we again use the standard result to get (19).

Proof of Equation (20).

$$\begin{aligned}
\left\|W_i 1_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right\|_{\psi_1} &= \sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\mathbb{E}\left[|(X_i^\top u)(X_i^\top\theta^*)\Delta_i 1_{\mathcal{E}_1\cap\mathcal{E}_2^c}|^p\right]^{1/p} \\
&\le \sup_{u\in\mathbb{S}^{d-1}}\sup_{p\ge 1} p^{-1}\mathbb{E}\tau\left[|(X_i^\top u)|^p\right]^{1/p} \\
&\le K_2\tau,
\end{aligned}$$

getting the desired result.

Proof of Equation (21). For this quantity, note that

$$P(\forall i\in[n], |Z_i| \lesssim \log(n/\delta)) \ge 1 - n\exp(-\tau^2).$$

Hence it is very likely that no sample falls into this category. Meanwhile, we can bound the expectation term:

$$\begin{aligned}\sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}[W^\top u 1_{\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3^c}] &\le \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}[(W^\top u)1_{\mathcal{E}_1\cap\mathcal{E}_2}|\mathcal{E}_3^c]P(\cap\mathcal{E}_3^c)\\ &\le \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}[|(X_i^\top u)(X_i^\top\theta^*)1_{\mathcal{E}_1\cap\mathcal{E}_2}||\mathcal{E}_3^c]P(\mathcal{E}_3^c)\\ &\le \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}[|(X_i^\top u)(X_i^\top\theta^*)|]P(\mathcal{E}_3^c)\\ &\le K_4\|\theta^*\|\exp(-\tau^2).\end{aligned}$$

Since $\tau = \Theta(\log(n\|\theta^*\|/\delta))$, we have the result.

B.2 Middle SNR Regime

We consider two cases, when $\|\theta^*\| \ge 1$ and $\|\theta^*\| \le 1$.

Case (i) $1 \le \|\theta^*\| \le C$**:** Given the initialization conditions in Theorem 1, we can get the following corollary of Lemma 3.

Corollary 1. *When* $\|\theta^*\| \ge 1$ *and* $\sin\alpha < 0.1$*, we have*

$$\|M_{mlr}(\theta) - \theta^*\| < 0.9\|\theta - \theta^*\|.$$

The proof of Corollary 1 is in Appendix D.2.1. Furthermore, from the uniform concentration Lemma 11 in Appendix E, for all $\theta : \|\theta - \theta^*\| \le O(\|\theta^*\|)$, we have

$$\|M_{n,mlr}(\theta) - M_{mlr}(\theta)\| \le C\sqrt{d\log^2(n/\delta)/n}$$

with probability $1-\delta$. From here, we can check that

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

Case (ii) $C_0(d\log^2(n/\delta)/n)^{1/4} \le \|\theta^*\| \le 1$**:** In this case, the result of Lemma 3 shows that:

Corollary 2. *When* $\|\theta^*\| \le 1$ *and* $\sin\alpha < 0.1$*, we have*

$$\|M_{mlr}(\theta) - \theta^*\| \le \left(1 - \frac{1}{8}\|\theta^*\|^2\right)\|\theta - \theta^*\|. \tag{23}$$

In order to analyze the convergence of finite-sample EM operator, we first divide the iterations into several epochs. Let $\bar{C}_0 = \|\theta_n^0 - \theta^*\|$. We consider that in each l^{th} epoch, θ satisfies $\bar{C}_0 2^{-l-1} \le \|\theta - \theta^*\| \le \bar{C}_0 2^{-l}$. Note that such consideration of dividing into several epochs is only conceptual, and does not affect the implementation of the EM algorithm.

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

$$\frac{1}{8}\|\theta^*\|^2\|\theta - \theta^*\| \ge 2cr\sqrt{d\log^2(n/\delta)/n},$$

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

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

then we have

$$\|M_{n,\mathrm{mlr}}(\theta) - \theta^*\| \le \left(1 - \frac{1}{16}\|\theta^*\|^2\right)\|\theta - \theta^*\|.$$

Arranging the terms, we require that

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

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

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

Therefore, θ is guaranteed to move closer to θ^* as long as $\bar{C}_0 2^{-l} \leq c_2\|\theta^*\|^{-1}\sqrt{d\log^2(n/\delta)/n}$. Note that each epoch takes $O(\|\theta^*\|^{-2})$ iterations to enter the next epoch. We can conclude that after $l = O(\log(n/d))$ epochs, we enter the region where $\|\theta - \theta^*\| \leq c_2\|\theta^*\|^{-1}\sqrt{d\log^2(n/\delta)/n}$ for some absolute constant $c_2 > 0$.

For δ probability bound, we can replace δ with $\delta/\log(n/d)$ and take a union bound of the uniform deviation of finite-sample EM operators given in Lemma 11 for all epochs. This does not change the complexity in the final statistical error.

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

Remark 1. *After $O(\log(n/d))$ epochs, studying on the property of the Hessian in a very close neighborhood of $\|\theta^*\|$ may lead to a guarantee that EM indeed converges to the empirical MLE, see Section 6 in Wu and Zhou (2019) for example.*

B.3 Low SNR Regime

As mentioned in the main text, the core idea of the low SNR regime is that EM essentially cannot distinguish the cases between $\theta^* = 0$ and $\theta^* \neq 0$. Therefore, instead of studying the contraction of population EM operator to θ^*, we study its contraction to 0. Given that insight, we have the following result with the norm of population EM operator:

Lemma 6. *There exists some universal constants $c_u > 0$ such that,*

$$\|\theta\|(1 - 4\|\theta\|^2 - c_u\|\theta^*\|^2) \leq \|M_{mlr}(\theta)\| \leq \|\theta\|(1 - \|\theta\|^2 + c_u\|\theta^*\|^2).$$

The proof of the Lemma 6 is in Section D.1.1. The result of Lemma 6 shows that the contraction coefficient of the population operator M_{mlr} consists of two terms: the non-expansive term, which is at the order of $1 - \mathcal{O}(\|\theta\|^2)$, and the quadratic term $\|\theta^*\|^2$ (up to some constant). Since we are in low SNR regime, the contraction coefficient gets close to 1. It demonstrates that the updates from population EM operator suffers from sub-linear convergence rate, instead of geometric convergence rate as that in high SNR regime.

From Lemma 2, we immediately have that

$$\sup_{\|\theta\|\leq r}\|M_{n,\text{mlr}}(\theta) - M_{\text{mlr}}(\theta)\| \leq cr\sqrt{d\log^2(n/\delta)/n},$$

for some universal constant $c > 0$.

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

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

$$\epsilon_n \geq 4c_u\|\theta^*\|^2 + 4\sup_{\theta\in\mathbb{B}(\theta^*, r_l)}\|M_{n,\text{ind}}(\theta) - M_{\text{ind}}(\theta)\|/r_l,$$

with $r_l = \epsilon_n^{\alpha_l}$. During this period, from Lemma 6 on contraction of population EM, and Lemma 2 concentration of finite sample EM, we can check that

$$\begin{aligned}\|M_{n,\mathrm{ind}}(\theta)\| &\le \|\theta\| - 0.5\|\theta\|^3 + c_u\|\theta\|\|\theta^*\|^2 + \sup_{\theta\in\mathbb{B}(\theta^*,r)} \|M_{n,\mathrm{ind}}(\theta) - M_{\mathrm{ind}}(\theta)\| \\ &\le \|\theta\| - \frac{1}{2}\epsilon_n^{3\alpha_{l+1}} + \frac{1}{4}\epsilon_n^{\alpha_l+1}.\end{aligned}$$

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

$$\alpha_{l+1} = \frac{1}{3}(\alpha_l + 1). \tag{24}$$

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

$$\|M_{n,\mathrm{ind}}(\theta)\| \le \|\theta\| - \frac{1}{4}\epsilon_n^{\alpha_l+1}.$$

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

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

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

It is easy to check $\alpha_l = (1/3)^l(\alpha_0 - 1/2) + 1/2$ from (24). We can set $l = C\log(1/\beta)$ for some universal constant C such that α_l is $1/2 - \beta$ for arbitrarily small $\beta > 0$. In conclusion, $\|\theta_n^t\| \le \epsilon_n^{1/2-\beta} \le c\cdot(d\ln^2(n/\delta)/n)^{1/4-\beta/2}$ with high probability as long as $t \ge \epsilon_n^{-1}l \gtrsim \sqrt{d/n}\log(1/\beta)$ where c is some universal constant. Hence we can set $\beta = C/\log(d/n)$ to get a desired result $\|\theta_n^t\| \le c\cdot(d\ln^2(n/\delta)/n)^{1/4}$. Since $\|\theta^*\| \le C_0(d\ln^2(n/\delta)/n)^{1/4}$, it implies $\|\theta_n^t - \theta^*\| \le c_1(d\ln^2(n/\delta)/n)^{1/4}$ where c_1 is some universal constant.

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

C Global Convergence of the (Easy) EM

This appendix gives a full proof of Theorem 2. We prove the result for bounded instances with $\{\theta^* : \|\theta^*\| \le C\}$ for some universal constant $C > 0$. The global convergence property of the (Easy)-EM algorithm will be used for the initialization for Theorem 1, hence we will focus on the iterations that the estimator stays outside of the initialization region. While we start with Easy-EM when $\cos\angle(\theta_n^0, \theta^*)$ is in order $O(1/\sqrt{d})$, note that we can safely go back to the standard EM algorithm as soon as $\cos\angle(\theta_n^t, \theta^*)$ becomes $\Theta(1)$ (see Section 4 in Kwon et al. (2019) for more details).

C.1 Decreasing Norm with Large Initialization in Low SNR Regime

In low SNR regime, we require that $\|\theta_n^0\| \le 0.2$. Here, when we initialize with large norm such that $\|\theta_n^0\| \ge 0.2$, we show that in a finite number of steps it becomes that $\|\theta_n^0\| \le 0.2$. We remark that in low SNR regime we consider when $\|\theta^*\| \ll 1$.

First, suppose $\|\theta\| \ge 2/3$. Then,

$$\begin{aligned}\|M_{\mathrm{mlr}}(\theta)\| &\le \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}[(X^\top\theta^*)(X^\top u)\tanh(YX^\top\theta)] + \mathbb{E}[Z(X^\top u)\tanh(YX^\top\theta)] \\ &\le \sup_{u\in\mathbb{S}^{d-1}} \sqrt{\mathbb{E}[(X^\top\theta^*)^2]\mathbb{E}[(X^\top u)^2]} + \mathbb{E}[|Z(X^\top u)|],\end{aligned}$$

$$\leq \|\theta^*\| + \mathbb{E}[|Z(X^\top u)|] \leq \|\theta^*\| + 2/\pi.$$

where $Z \sim \mathcal{N}(0,1)$ such that $Y = X^\top \theta^* + Z$. Since the uniform deviation in Easy-EM is given by Lemma 11 as $\sqrt{d\log^2(n/\delta)/n}$, we can conclude that

$$\begin{aligned}\|M_{n,\mathrm{mlr}}(\theta)\| &\leq \|M_{\mathrm{mlr}}(\theta)\| + O\left(\sqrt{d\log^2(n/\delta)/n}\right)\\ &\leq \|\theta^*\| + 2/\pi + O\left(\sqrt{d\log^2(n/\delta)/n}\right) \leq 2/3.\end{aligned}$$

Next, suppose $0.2 \leq \|\theta\| \leq 2/3$. Following the notation in Appendix A, we recall equation (10),

$$M_{\mathrm{mlr}}(\theta) = \mathbb{E}[yx_1 \tanh(yx_1\|\theta\|)]v_1 + \mathbb{E}[yx_2 \tanh(yx_1\|\theta\|)]v_2,$$

where $y = X^\top\theta^* + z$ where $z \sim \mathcal{N}(0,1)$, $x_1 = X^\top v_1$ and $x_2 = X^\top v_2$. We will see in Appendix D.1.1 that $M_{\mathrm{mlr}}(\theta)^\top v_2 \leq \frac{1}{2}\|\theta\|\|\theta^*\|^2 \leq c_0\sqrt{d\log^2(n/\delta)/n}$ for some absolute constant $c_0 > 0$. Therefore, we focus on bounding the first term.

Let $a = 4$, and define event $\mathcal{E} := \{x_1^2 + z^2 \leq a\}$. We expand $M_{\mathrm{mlr}}(\theta)$ as follows:

$$\begin{aligned}M_{\mathrm{mlr}}(\theta)^\top v_1 &\leq \|\theta\|\mathbb{E}[y^2x_1^2 1_{\mathcal{E}}] + \mathbb{E}[|yx_1|1_{\mathcal{E}^c}]\\ &\leq \|\theta\|\mathbb{E}[z^2x_1^2 1_{\mathcal{E}}] + \mathbb{E}[|zx_1|1_{\mathcal{E}^c}] + O(\|\theta^*\|).\end{aligned}$$

By converting the above expression to Rayleigh distribution with $x_1 = r\cos w, z = r\sin w$, we can more explicitly find the values of the expectations in the above equation. That is,

$$\mathbb{E}[z^2x_1^2 1_{\mathcal{E}}] = \frac{1}{2\pi}\int_0^{2\pi}\cos^2 w \sin^2 w dw \int_0^4 r^5\exp(-r^2/2)dr \approx 1 - 0.013,$$

and

$$\mathbb{E}[|zx_1|1_{\mathcal{E}^c}] = \frac{1}{2\pi}\int_0^{2\pi}|\cos w \sin w|dw \int_4^\infty r^3\exp(-r^2/2)dr \leq 0.002,$$

Now using the condition that $\|\theta\| \leq 0.2$, we have

$$M_{\mathrm{mlr}}(\theta)^\top v_1 \leq \|\theta\|(1 - 0.003) + O(\|\theta^*\|) \leq \gamma\|\theta\| + O(\|\theta^*\|),$$

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

$$\|M_{\mathrm{mlr}}(\theta)\| \leq \gamma\|\theta\| + O\left(\sqrt{d\log^2(n/\delta)/n} + \|\theta^*\|\right).$$

Hence we can conclude that after $t = O(1)$ iterations, $\|\theta_n^t\| \leq 0.2$.

C.2 Angle Convergence in Middle-to-High SNR Regime

Now we work in the regime where $\|\theta^*\| = \eta \geq c_\eta(d\log(n/\delta)^2/n)^{1/4}$ for some sufficiently large constant $c_\eta > 0$. We first focus on the convergence of angle from random initialization.

Let us denote $\alpha_t := \angle(\theta_n^t, \theta^*)$. Note that since we initialize with a random vector sampled uniformly from the unit sphere, $\cos\alpha_0 = O(1/\sqrt{d})$. We bring the following lemma for the change in angles for a *fixed* estimator θ_n^t given in Kwon et al. (2019):

Lemma 7 (Theorem 8 in Kwon et al. (2019)). *Let $\epsilon_f := c_0\max(1,\eta^{-1})\sqrt{d/n}$ be the statistical fluctuation with some universal constant $c_0 > 0$ in one-step iteration of Easy-EM. Suppose the norm of the current estimator $\|\theta_n^t\|$ is larger than $\|\theta^*\|/10$. Then we have,*

$$\cos\alpha_{t+1} \geq \kappa_t(1 - 10\epsilon_f)\cos\alpha_t - \frac{\epsilon_f}{\sqrt{d}}, \tag{25}$$

$$\sin^2\alpha_{t+1} \le \kappa_t' \sin^2\alpha_t + \epsilon_f, \tag{26}$$

where $\kappa_t = \sqrt{1+\frac{\sin^2\alpha_t}{\cos^2\alpha_t+\frac{1}{2}(1+\eta^{-2})}} \ge 1$, *and* $\kappa_t' = \left(1+\frac{2\eta^2}{1+\eta^2}\cos^2\alpha_t\right)^{-1} < 1$.

Here, the κ_t comes from Theorem 2 in Kwon et al. (2019) for the convergence rate of the cosine values of the population EM operator. The key idea in the above lemma is that when we bound the statistical error of cosine value, we need to bound an error in one fixed direction $u := \theta^*/\|\theta^*\|$ instead of all directions in $\mathbb{R}^d$ to bound l_2 norm. More specifically, they show that

$$\left(\frac{1}{n}\sum_i (X_i^\top u) Y_i \tanh(Y_i X_i^\top \theta) - M_{\mathrm{mlr}}(\theta)^\top u\right) \lesssim (1+\|\theta^*\|)\sqrt{1/n} \lesssim (1+\|\theta^*\|)\epsilon_f/\sqrt{d}.$$

Remark 2. *Kwon et al. (2019) requires the sample-splitting scheme in which we draw a new batch of samples at every step. The main challenge when we try to remove the sample-splitting is to show that the above argument holds for all* $\theta : \|\theta\| \le r$ *where* $r = O(\max\{1, \|\theta^*\|\})$. *For large* $\|\theta^*\|$, *getting a right order of uniform statistical error is challenging: discretization of* θ *results in extra* $\sqrt{d}$ *factor, while the Ledoux-Talagrand type approach as in Lemma 11 results in extra* $O(\|\theta^*\|)$ *factor. Therefore, here we show only for bounded instances with* $\|\theta^*\| \le C$, *and leave the analysis for arbitrarily large* $\|\theta^*\|$ *as future work.*

Now we adopt their approach to work *without* sample-splitting, and get a right order of sample complexity. First, when we work with bounded θ^*, we follow the steps in Lemma 11, while we can skip the procedure in which we take a union bound over 1/2-covering set of the unit sphere to bound l_2 norm of a random vector. This yields that

$$\sup_{\|\theta\|\le r}\left|\frac{1}{n}\sum_i (X_i^\top u) Y_i \tanh(Y_i X_i^\top \theta) - M_{\mathrm{mlr}}(\theta)^\top u\right| \le cr\sqrt{\log^2(n/\delta)/n}, \tag{27}$$

for the absolute constant $c > 0$ given by Lemma 11. Let $\epsilon_f := c\sqrt{d\log^2(n/\delta)/n}$. The cosine value can be bounded as follows:

$$\begin{aligned}
\cos\alpha_{t+1} &= \frac{(\theta^*)^\top \theta_n^{t+1}}{\|\theta_n^{t+1}\|\|\theta^*\|} \\
&= \frac{u^\top(M_{\mathrm{ind}}(\theta_n^t) - \theta_n^{t+1})}{\|\theta_n^{t+1}\|} + \frac{u^\top M_{\mathrm{ind}}(\theta_n^t)}{\|M_{\mathrm{ind}}(\theta_n^t)\|}\frac{\|M_{\mathrm{ind}}(\theta_n^t)\|}{\|\theta_n^{t+1}\|}, \\
&\ge -\frac{\epsilon_f}{\sqrt{d}}\frac{r}{\|\theta_n^{t+1}\|} + \frac{u^\top M_{\mathrm{ind}}(\theta_n^t)}{\|M_{\mathrm{ind}}(\theta_n^t)\|}\frac{\|M_{\mathrm{ind}}(\theta_n^t)\|}{\|M_{\mathrm{ind}}(\theta_n^t)\| + r\epsilon_f} \\
&\ge \kappa_t \cos\alpha_t\left(1 - \frac{r\epsilon_f}{\|M_{\mathrm{ind}}(\theta_n^t)\|}\right) - \frac{\epsilon_f}{\sqrt{d}}\frac{r}{\|M_{\mathrm{mlr}}(\theta_n^t)\| - r\epsilon_f},
\end{aligned}$$

where the last inequality comes from Theorem 2 in Kwon et al. (2019).

Finally, we need to show that $r/\|M_{\mathrm{mlr}}(\theta_n^t)\| = O(1)$ such that we can set ϵ_f as some sufficiently small absolute constant (that does not depend on η). We first need the following lemma on the norm of the next estimator:

Lemma 8. *If* $\|\theta\| \le \|\theta^*\|/10$, *then*

$$\|M_{mlr}(\theta)\| \ge \|\theta\|(1 + d_1 \cdot \min\{1, \|\theta\|^2\}).$$

Otherwise, if $\|\theta\| \ge \|\theta^*\|/10$, *we have*

$$\|M_{mlr}(\theta)\| \ge \frac{\|\theta^*\|}{10}(1 + d_2 \cdot \min\{1, \|\theta^*\|^2\}).$$

for some universal constants $d_1, d_2 > 0$.

We defer the proof of this lemma to Appendix D.4.

We need the uniform concentration (27) for several values of $r = C_0, C_0 2^{-1}, ..., C_0 2^{-l+1}, C_0 2^{-l}$ where $C_0 = 3C$ and $l = O(\log(n/d))$. We can replace δ by $\delta/\log(n/d)$ for union bound, which does not change the order of statistical error. Pick k such that $C_0 2^{-k} \le \|\theta_n^t\| \le C_0 2^{-k+1} = r$.

When $\|\theta_n^t\| \le \|\theta^*\|/10$, we can apply the Lemma 8 to see

$$r/\|M_{\mathrm{mlr}}(\theta_n^t)\| \le C_0 2^{-k+1}/(C_0 2^{-k}) = 2,$$

where we used $r = 2^{-k+1}$. Therefore, $r/M_{\mathrm{mlr}}(\theta_n^t) = O(1)$. On the other hand, if $\|\theta_n^t\| \ge \|\theta^*\|/10$, then we divide the cases when $\|\theta^*\| \ge 1/\max(3, c_2)$ where $c_2 > 0$ satisfies the lower bound given in equation (32):

$$\|M_{\mathrm{mlr}}(\theta)\| \ge \|\theta\|(1 - 3\|\theta\|^2) - c_2\|\theta\|\|\theta^*\|^2.$$

When $\|\theta^*\| \ge 1/\max(3, c_2)$ and $\|\theta_n^t\| \ge \|\theta^*\|/10$, by Lemma 8 we have $r/M_{\mathrm{mlr}}(\theta) \le C_0 \max(3, c_2) = O(1)$ since all parameters here are universal constants. On the other hand, if $\|\theta^*\| \le 1/\max(3, c_2)$ and $\|\theta_n^t\| \ge \|\theta^*\|/10$, then from equation (32) we have

$$\|M_{\mathrm{mlr}}(\theta)\| \ge \|\theta\|(1 - 3\|\theta\|^2) - c_2\|\theta\|\|\theta^*\|^2 \ge \|\theta\|/2.$$

Therefore, $r/\|M_{\mathrm{mlr}}(\theta_n^t)\| \le C_0 2^{-k+1}/(C_0 2^{-k-1}) = 4 = O(1)$.

From the above case study, we have that

$$\cos\alpha_{t+1} \ge \kappa_t \cos\alpha_t(1 - c_4\epsilon_f) - c_5\frac{\epsilon_f}{\sqrt{d}},$$

for some absolute constants $c_4, c_5 > 0$. Now observe that as long as $\sin\alpha_t > c_\alpha$, $\kappa_t = 1 + c_6\min\{1, \eta^2\}$ for some sufficiently small constant $c_\alpha, c_6 > 0$. Also, recall that we are considering the middle-to-high SNR regime when $\eta^2 \ge c_\eta\sqrt{d\log^2(n/\delta)/n}$ for some sufficiently large constant $c_\eta > 0$, whereas $\epsilon_f \le c\sqrt{d\log^2(n/\delta)/n}$ for another fixed constant $c > 0$. Therefore, there exists a universal constant $c_7 > 0$ such that for all $\cos\alpha_t \ge 1/\sqrt{d}$, we have

$$\cos\alpha_{t+1} \ge (1 + c_7\min(1, \eta^2))\cos\alpha_t.$$

After $t = O(\eta^{-2}\log(d))$ iterations starting from $\cos\alpha_0 = 1/\sqrt{d}$, we have $\cos\alpha_t \ge 0.95$ or $\sin\alpha_t \le 0.1$.

C.3 Stability and Convergence in Middle-to-High SNR Regime after Alignment

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

Sine stays below some threshold. Once θ_n^t and θ^* are well-aligned, using $\sin^2\alpha_t = 1 - \cos^2\alpha_t$, similar arguments can be applied for sin values:

$$\begin{aligned} &\sin^2\alpha_{t+1} \le (1 - c_1\min(1, \eta^2))\sin^2\alpha_t, && \text{if} \quad \sin^2\alpha_t \ge c_2 \\ &\sin^2\alpha_{t+1} \le c_2, && \text{else} \quad \sin^2\alpha_t \le c_2, \end{aligned}$$

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

Initialization from small estimators after alignment. After the angle is aligned such that $\sin\alpha_t \le c_2$. We see how fast $\|\theta_n^t\|$ enters the desired initialization region that Theorem 1 requires, when $\|\theta_n^t\| \le 0.9\|\theta^*\|$.

Let us first consider the case $0.1\|\theta^*\| \le \|\theta_n^t\| \le 0.9\|\theta^*\|$. We recall Lemma 3 such that

$$\begin{aligned} \|\theta^* - M_{\mathrm{mlr}}(\theta_n^t)\| &\le \kappa\|\theta_n^t - \theta^*\| + \kappa 16\sin^2\alpha\|\theta_n^t - \theta^*\|\frac{\eta^2}{1+\eta^2} \\ &\le \kappa(1 + (16\sin^2\alpha)\eta^2)\|\theta_n^t - \theta^*\|, \end{aligned}$$

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

$$\|\theta^* - M_{\mathrm{mlr}}(\theta_n^t)\| \le (1 - c_4\min(1, \eta^2))\|\theta - \theta^*\|,$$

for some constant $c_4 > 0$. Since we are in the regime $\eta^2 \ge c_\eta\sqrt{d\log^2(n/\delta)/n}$ for sufficiently large c_η, by appropriately setting the constants we have $\|M_{n,\mathrm{mlr}}(\theta_n^t) - \theta^*\| \le (1 - c_5\min(1, \eta^2))\|\theta - \theta^*\|$ for some absolute

constant $c_5 > 0$, as long as we are in the region $0.1\|\theta^*\| \le \|\theta_n^t\| \le 0.9\|\theta^*\|$. Hence after $O(\max(1, \eta^{-2}))$ iterations, we reach to the desired initialization region.

Now we consider the case $\|\theta\| \le 0.1\|\theta^*\|$. In this case, by Lemma 8, we can show that

$$\|M_{\text{mlr}}(\theta)\| \ge \|\theta\|(1 + c_6 \min\{1, \|\theta\|^2, \|\theta^*\|^2\}),$$

for some universal constant $c_6 > 0$. After $O(\max\{\|\theta\|^{-2}, \|\theta^*\|^{-2}\})$ iterations, we enter $\|\theta\| \ge \|\theta^*\|/10$. Note that when we start with $\|\theta_n^0\| = \Omega(1)$, $\|\theta_n^t\|$ will stay above $\min\{\Omega(1), \|\theta^*\|/10\}$ throughout all iterations due to Lemma 8 and Lemma 8.

D Deferred Lemmas

In this appendix, we collect proofs for auxiliary lemmas which were postponed in the proof of main theorems: the contraction of population EM operators under both middle and low SNR regimes, uniform deviation of finite-sample EM operators, and the lower bounds on the norms of population EM operators.

D.1 Contraction of the Population EM Operator under Low SNR Regime

D.1.1 Proof of Lemma 6

We use notations and definitions stated in A.

Upper Bound: We first bound the first coordinate of the population operator from equation (10):

$$M_{mlr}(\theta)^\top v_1 = \mathbb{E}_{x_1,x_2,y}\left[\tanh(yx_1\|\theta\|)x_1 y\right],$$

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

$$x^2 - \frac{x^4}{3} \le x\tanh(x) \le x^2 - \frac{x^4}{3} + \frac{2x^6}{15}. \tag{28}$$

Now we unfold the equation above, we have

$$\begin{aligned}
M_{mlr}(\theta)^\top v_1 &= \frac{1}{\|\theta\|}\mathbb{E}_{x_1,x_2,y}\left[\tanh(yx_1\|\theta\|)yx_1\|\theta\|\right] \\
&\le \frac{1}{\|\theta\|}\mathbb{E}_{x_1,x_2,y}\left[(yx_1\|\theta\|)^2 - \frac{(yx_1\|\theta\|)^4}{3} + \frac{2(yx_1\|\theta\|)^6}{15}\right] \\
&\le \frac{1}{\|\theta\|}\mathbb{E}_{x_1,z}\Bigg[(x_1\|\theta\|(z + x_1b_1^* + x_2b_2^*))^2 - \frac{(x_1\|\theta\|(z + x_1b_1^* + x_2b_2^*))^4}{3} \\
&\qquad + \frac{2(x_1\|\theta\|(z + x_1b_1^* + x_2b_2^*))^6}{15}\Bigg],
\end{aligned}$$

where $z \sim \mathcal{N}(0, 1)$. Note here that, any (constantly) higher order terms of Gaussian distribution is constant. Hence instead of computing all coefficients explicitly for all monomials, we can simplify the argument as

$$\begin{aligned}
M_{mlr}(\theta)^\top v_1 &\le \frac{1}{\|\theta\|}\mathbb{E}_{x_1,z}\left[(x_1\|\theta\|z)^2 - \frac{(x_1\|\theta\|z)^4}{3} + \frac{2(x_1\|\theta\|z)^6}{15}\right] + c_1\|\theta\|\|\theta^*\|^2, \\
&= \|\theta\|(1 - 3\|\theta\|^2 + 30\|\theta\|^4) + c_1\|\theta\|\|\theta^*\|^2,
\end{aligned} \tag{29}$$

for some universal constant $c_1 > 0$. Since we assumed $\|\theta\| < 0.2$, we have $3\|\theta\|^2 - 30\|\theta\|^4 \ge \|\theta\|^2$. We conclude that

$$M_{mlr}(\theta)^\top v_1 \le \|\theta\|(1 - \|\theta\|^2 + c_1\|\theta^*\|^2).$$

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

$$M_{mlr}(\theta)^\top v_2 = \mathbb{E}_{x_1,x_2,y}\left[\tanh(yx_1\|\theta\|)yx_2\right],$$

where $y|(x_1, x_2) \sim \mathcal{N}(x_1 b_1^* + x_2 b_2^*, 1)$. In order to derive an upper bound for the above equation, we rely on the following equation which we defer the proof to the end of this section:

$$\mathbb{E}\left[\tanh(yx_1\|\theta\|)yx_2\right] = b_2^* \, \mathbb{E}\left[x_1^2 \tanh(x_1\|\theta\|(z + x_1 b_1^*)) - \|\theta\| b_1^* x_1^2 \tanh'(x_1\|\theta\|(z + x_1 b_1^*))\right], \tag{30}$$

where $z \sim \mathcal{N}(0, 1 + b_2^{*2})$ with subsuming x_2 from the equation. From (30), we can check that

$$\begin{aligned}
\mathbb{E}\left[\tanh(yx_1\|\theta\|)yx_2\right] &\le b_2^* \, \mathbb{E}\left[x_1^2 \tanh(x_1\|\theta\|(z + x_1 b_1^*))\right] \\
&= \frac{b_2^*}{2} \mathbb{E}\left[x_1^2 \tanh(x_1\|\theta\|(z + x_1 b_1^*)) + x_1^2 \tanh(x_1\|\theta\|(-z + x_1 b_1^*))\right] \\
&\le b_2^* \mathbb{E}\left[x_1^2 \tanh(x_1^2\|\theta\| b_1^*)\right], \\
&\le \|\theta\| b_1^* b_2^* \mathbb{E}\left[x_1^4\right] \le \frac{1}{2}\|\theta\|\|\theta^*\|^2,
\end{aligned}$$

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

From the above results, we have shown that

$$\|M_{mlr}(\theta)\| \le |M_{\mathrm{mlr}}(\theta)^\top v_1| + |M_{\mathrm{mlr}}(\theta)^\top v_2| \le \|\theta\| \left(1 - \|\theta\|^2 + c\|\theta^*\|^2\right), \tag{31}$$

for some universal constant $c > 0$.

Lower Bound: To prove the lower bound of the population EM operator, we again expand the equation using Taylor series (28):

$$\|M_{\mathrm{mlr}}(\theta)\| \ge |M_{\mathrm{mlr}}(\theta)^\top v_1| \ge \|\theta\|(1 - 3\|\theta\|^2) - c_2\|\theta\|\|\theta^*\|^2. \tag{32}$$

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

Proof of equation (30): For the left hand side, we apply the Stein's lemma with respect to x_2. It gives that

$$\begin{aligned}
\mathbb{E}[\tanh(\|\theta\|x_1 y)yx_2] &= \mathbb{E}\left[\frac{d}{dx_2}\tanh(\|\theta\|x_1 y)y\right] \\
&= \mathbb{E}\left[\frac{d}{dx_2}\tanh(\|\theta\|x_1(\bar{z} + x_1 b_1^* + x_2 b_2^*))(\bar{z} + x_1 b_1^* + x_2 b_2^*)\right] \\
&= \mathbb{E}[b_2^* \tanh(\|\theta\|x_1(\bar{z} + x_1 b_1^* + x_2 b_2^*)) \\
&\quad + (\|\theta\|x_1 b_2^*)(\bar{z} + x_1 b_1^* + x_2 b_2^*)\tanh'(\|\theta\|x_1(\bar{z} + x_1 b_1^* + x_2 b_2^*)] \\
&= b_2^* \, \mathbb{E}[\tanh(\|\theta\|x_1(z + x_1 b_1^*)) + \|\theta\|x_1(z + x_1 b_1^*)\tanh'(\|\theta\|x_1(z + x_1 b_1^*)))]
\end{aligned}$$

where $\bar{z} \sim \mathcal{N}(0, 1)$ and $z \sim \mathcal{N}(0, 1 + b_2^{*2})$. For the right hand side, we apply the Stein's lemma with respect to x_1. First, we check the first term in the right hand side that

$$\begin{aligned}
&\mathbb{E}[x_1^2 \tanh(\|\theta\|x_1(z + x_1 b_1^*))] \\
&\quad = \mathbb{E}\left[\frac{d}{dx_1}(x_1 \tanh(\|\theta\|x_1(z + x_1 b_1^*)))\right] \\
&\quad = \mathbb{E}\left[\tanh(\|\theta\|x_1(z + x_1 b_1^*)) + x_1 \frac{d}{dx_1}\tanh(\|\theta\|x_1(z + x_1 b_1^*)\right] \\
&\quad = \mathbb{E}\left[\tanh(\|\theta\|x_1(z + x_1 b_1^*)) + \|\theta\|x_1(z + 2x_1 b_1^*)\tanh'(\|\theta\|x_1(z + x_1 b_1^*)\right].
\end{aligned}$$

Plugging this into (30) and subtracting the remaining term gives the result that matches to the left hand side.

D.2 Contraction of the Population EM Operator under Middle SNR Regime

In this appendix, we provide the proofs for contraction of the population EM operator under middle SNR regime.

D.2.1 Proof of Corollary 1

In Lemma 3, note that $\kappa \le 1 - \frac{1}{2}\min\{\|\theta\|^2, \frac{\|\theta^*\|^2}{\|\theta^*\|^2+1}\}$ and $(\|\theta^*\|\sin\alpha) < \|\theta - \theta^*\|$ where $\sin\alpha < 1/10$. Therefore, whenever $\|\theta^*\| \ge 1$, with the initialization condition $\|\theta\| \ge 0.9\|\theta^*\|$

$$\|M_{\mathrm{mlr}}(\theta) - \theta^*\| \le (1 - 1/4)\,\|\theta - \theta^*\| + \kappa 16(\sin^2\alpha)\|\theta - \theta^*\| \le 0.9\|\theta - \theta^*\|,$$

which completes the proof.

D.2.2 Proof of Corollary 2

From Lemma 3, note that $\frac{\eta^2}{1+\eta^2} \le \eta^2 = \|\theta^*\|^2$. Using $\kappa \le 1 - \frac{1}{2}\min\{\|\theta\|^2, \frac{\|\theta^*\|^2}{\|\theta^*\|^2+1}\}$, $(\|\theta^*\|\sin\alpha) < \|\theta - \theta^*\|$ and $\sin\alpha < 1/10$. With the initialization condition $\|\theta\| \ge 0.9\|\theta^*\|$, we have

$$\|M_{mlr}(\theta) - \theta^*\| \le \left(1 - \frac{1}{4}\|\theta^*\|^2\right)\|\theta - \theta^*\| + \frac{1}{8}\|\theta^*\|^2\|\theta - \theta^*\| \le \left(1 - \frac{1}{8}\|\theta^*\|^2\right)\|\theta - \theta^*\|.$$

D.3 Uniform deviation of finite-sample EM operator: Proof of Lemma 2

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

$$\begin{aligned}\|M_{n,\mathrm{mlr}}(\theta) - M_{\mathrm{mlr}}(\theta)\| \le |\!|\!|\hat{\Sigma}_n^{-1}|\!|\!|_{\mathrm{op}}\Big\|\frac{1}{n}\sum_{i=1}^n Y_i X_i \tanh(Y_i X_i^\top\theta) - M_{\mathrm{mlr}}(\theta)\Big\| \\ + |\!|\!|\hat{\Sigma}_n^{-1} - I|\!|\!|_{\mathrm{op}}\|M_{\mathrm{mlr}}(\theta)\|.\end{aligned}$$

The first term can be bounded by $c_1 r\sqrt{d\log^2(n/\delta)/n}$ with some absolute constant $c_1 > 0$ using the results of (8) and Lemma 9 in Appendix E.

For the second term, we first know from Lemma 9 that $|\!|\!|\hat{\Sigma}_n^{-1} - I|\!|\!|_{\mathrm{op}} = |\!|\!|\hat{\Sigma}_n^{-1}|\!|\!|_{\mathrm{op}}|\!|\!|\hat{\Sigma}_n - I|\!|\!|_{\mathrm{op}} \le c_2\sqrt{d/n})$ for some universal constant $c_2 > 0$. If we can show that $\|M_{\mathrm{mlr}}(\theta)\| \le O(r)$, then we are done. To see this, first we check that

$$\|M_{\mathrm{mlr}}(\theta)\| = \|\mathbb{E}[YX\tanh(YX^\top\theta)]\| \le \|\theta\||\!|\!|\mathbb{E}[Y^2XX^\top]|\!|\!|_{\mathrm{op}}.$$

It is easy to check that $\mathbb{E}[Y^2XX^\top] = I + 2\theta^*\theta^{*\top}$, hence $|\!|\!|\mathbb{E}[Y^2XX^\top]|\!|\!|_{\mathrm{op}} = 1 + 2\|\theta^*\|^2 \le 1 + 2C^2 = O(1)$. Therefore, $\|M_{\mathrm{mlr}}(\theta)\| \le c_3\|\theta\| \le c_3 r$ with a constant $c_3 = (1 + 2C^2)$. This completes the proof of Lemma 2. □

D.4 Lower Bound on the Norm: Proof of Lemma 8

This Lemma is in fact a more refined statement of Lemma 23 in Kwon et al. (2019) where they give a lower bound on the norms for the same purpose. We give a more refined result here.

Let $\alpha = \angle(\theta, \theta^*)$. We use the notations defined in Appendix A. We recall here that $b_1^* = \theta^*\cos\alpha$, $b_2^* = \theta^*\sin\alpha$. We consider three cases as in Kwon et al. (2019).

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

$$\begin{aligned}\|M_{\mathrm{mlr}}(\theta)\| \ge |M_{\mathrm{mlr}}(\theta)^\top v_1| &= \mathbb{E}_{x_1,x_2,y}[\tanh(yx_1\|\theta\|)yx_1] \\ &= \mathbb{E}_{x_1,x_2,z}[\tanh((x_1b_1^* + x_2b_2^* + z)x_1\|\theta\|)(x_1b_1^* + x_2b_2^* + z)x_1],\end{aligned}$$

where $x_1, x_2, z \sim \mathcal{N}(0,1)$. From the argument in Kwon et al. (2019), the above quantity is larger than the following $b_1^* = 0$ case (see Lemma 23 in Kwon et al. (2019) for details):

$$\mathbb{E}_{x_1,x_2,z}[\tanh((x_2b_2^* + z)x_1\|\theta\|)(x_2b_2^* + z)x_1] = \mathbb{E}_{x_1,\bar{z}}[\tanh(\bar{z}x_1\|\theta\|)\bar{z}x_1],$$

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

$$\begin{aligned}\mathbb{E}_{x_1,\bar{z}}[\tanh(\bar{z}x_1\|\theta\|)\bar{z}x_1] &\geq \sigma_2 \mathbb{E}_{x_1,z}[\tanh(\sigma_2 z x_1\|\theta\|)zx_1] \\ &\geq \sigma_2 \mathbb{E}_{x_1,z}[\tanh(zx_1\|\theta\|)zx_1].\end{aligned}$$

If $\|\theta\| > 0.5$, then through the numerical integration we can check that $\mathbb{E}_{x_1,z}[\tanh(0.5zx_1)zx_1] > 1/\pi$. Hence, we immediately have that

$$|M_{\mathrm{mlr}}(\theta)^\top v_1| \geq \frac{1}{\pi}\sigma_2 \geq \frac{\sin\alpha}{\pi}\|\theta^*\| \geq \frac{1}{5}\|\theta^*\|,$$

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

$$\|M_{\mathrm{mlr}}(\theta)\| \geq \|\theta\|(1 + 1 \cdot \min(1, \|\theta\|^2)).$$

If $\|\theta\| < 0.5$, then we get a lower bound using Taylor expansion:

$$\begin{aligned}\mathbb{E}_{x_1,\bar{z}}[\tanh(\bar{z}x_1\|\theta\|)\bar{z}x_1] &\geq \sigma_2\left(\mathbb{E}_{x_1,z}[\|\theta\|(zx_1)^2] - \frac{1}{3}\mathbb{E}_{x_1,z}[\|\theta\|^3(zx_1)^4]\right) \\ &= \sigma_2\|\theta\|(1 - 3\|\theta\|^2) = \|\theta\|\sqrt{1 + 0.96\eta^2}(1 - 3\|\theta\|^2),\end{aligned}$$

where $\|\theta^*\| = \eta$. Here, we consider three cases when $\eta \geq 5$, $5 \geq \eta \geq 1$, $1 \geq \eta$. When $\eta \geq 5$, then we immediately have $|M_{\mathrm{mlr}}(\theta)^\top v_1| \geq 1.25\|\theta\|$. In case $5 \geq \eta \geq 1$, we first note that since $\|\theta\| \leq \|\theta^*\|/10$, we check the value of

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

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

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

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

$$\|M_{\mathrm{mlr}}(\theta)\| \geq \|\theta\|(1 + 0.25 \cdot \min(1, \|\theta^*\|^2)) \geq \|\theta\|(1 + 0.25 \cdot \min(1, \|\theta\|^2)).$$

Now note that $M_{\mathrm{mlr}}(\theta)^\top v_1$ increases in $\|\theta\|$, hence for all $\|\theta\| \geq \|\theta^*\|/10$, it holds that

$$\|M_{\mathrm{mlr}}(\theta)\| \geq \frac{\|\theta^*\|}{10}(1 + 0.25 \cdot \min(1, \|\theta^*\|^2)).$$

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

In the first case when $\sigma_2^2\|\theta\| \leq b_1^*$, it can be shown that (see equation (50) in Kwon et al. (2019) for details)

$$b_1^* - M_{\mathrm{mlr}}(\theta)^\top v_1 \leq \kappa^3(b_1^* - \sigma_2^2\|\theta\|),$$

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

$$\begin{aligned}M_{\mathrm{mlr}}(\theta)^\top v_1 &\geq \|\theta^*\|(1 - \kappa^3)\cos\alpha + \kappa^3(1 + \eta^2\sin^2\alpha)\|\theta\| \\ &\geq \|\theta\|2(1 - \kappa^3) + \kappa^3(1 + \eta^2\sin^2\alpha)\|\theta\| \\ &\geq \|\theta\| + (1 - \kappa^3)\|\theta\|.\end{aligned}$$

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

$$M_{\mathrm{mlr}}(\theta)^\top v_1 \geq \|\theta^*\|/5 \geq \frac{\|\theta^*\|}{10}(1 + 1 \cdot \min(1, \|\theta^*\|^2)) \geq \|\theta\|(1 + 1 \cdot \min(1, \|\theta\|^2)).$$

Combining two cases, we have that

$$\|M_{\text{mlr}}(\theta)\| \geq \|\theta\|(1 + c_1 \cdot \min(1, \|\theta\|^2)).$$

Now similarly to *Case (i)*, since $M_{\text{mlr}}(\theta)^\top v_1$ is increasing in $\|\theta\|$, when $\|\theta\| \geq \|\theta^*\|/10$, we have

$$\|M_{\text{mlr}}(\theta)\| \geq \frac{\|\theta^*\|}{10}(1 + c_2 \cdot \min(1, \|\theta^*\|^2)),$$

where $c_2 = c_1/100$.

Collecting all results in two cases, we have Lemma 8.

E Concentration of Measures in Finite-Sample EM

In all lemmas that follow, we assume that $n \geq Cd$ for sufficiently large constant $C > 0$, such that the tail probability of the sum of n independent sub-exponential random variables are in sub-Gaussian decaying rate.

Lemma 9. *Suppose $X \sim \mathcal{N}(0, I)$ and $Y|X \sim \frac{1}{2}\mathcal{N}(X^\top\theta^*, 1) + \frac{1}{2}\mathcal{N}(-X^\top\theta^*, 1)$. Then, with probability at least $1 - \delta$,*

$$\frac{1}{n}\sum_{i=1}^n Y_i^2 - 1 = O\left((\|\theta^*\| + 1)^2\sqrt{\frac{\ln(1/\delta)}{n}}\right), \tag{33}$$

$$\Big\|\frac{1}{n}\sum_{i=1}^n X_i X_i^\top - I\Big\|_{op} = O\left(\sqrt{\frac{d\ln(1/\delta)}{n}}\right). \tag{34}$$

The above lemma is standard concentration lemmas for standard Gaussian distributions.

Lemma 10. *Let X, Y be the random variables as in Lemma 9. With probability at least $1 - \delta$, we have*

$$\Big\|\frac{1}{n}\sum_{i=1}^n Y_i^2 X_i X_i^\top - I\Big\|_{op} = O\left((\|\theta^*\| + 1)^2\sqrt{\frac{d\ln^2(n/\delta)}{n}}\right), \tag{35}$$

Proof. Let ν_i be an independent Rademacher variable and $Z_i = \mathcal{N}(0, 1)$. We can write $Y_i = \nu_i X_i^\top\theta^* + Z_i$. We use the truncation argument for the of concentration of higher order moments. First define the good event $\mathcal{E} := \{\forall i \in [n], |Z_i| \leq \tau, |X_i^\top\theta^*| \leq \tau_2|\}$. We will decide the order of τ later such that $P(\mathcal{E}) \geq 1 - \delta$. Let $\widetilde{Y} \sim Y|\mathcal{E}, \widetilde{X} \sim X|\mathcal{E}$ and $(\widetilde{Y}_i, \widetilde{X}_i)$ be independent samples of $(\widetilde{Y}, \widetilde{X})$. It is easy to check that $\widetilde{Y}\widetilde{X}$ is a sub-Gaussian vector with Orlicz norm $O(\tau + \tau_2)$ (Vershynin, 2010). To see this,

$$\left\|\widetilde{Y}\widetilde{X}\right\|_{\psi_2} = \sup_{u\in\mathbb{S}^{d-1}}\sup_{p\geq 1} p^{-1/2}\mathbb{E}\left[|Y(X^\top u)|^p|\mathcal{E}\right]^{1/p} \tag{36}$$

$$\leq (\tau + \tau_2)\sup_{u\in\mathbb{S}^{d-1}}\sup_{p\geq 1} p^{-1/2}\mathbb{E}\left[|X^\top u|^p 1_{\mathcal{E}}\right]^{1/p}/P(\mathcal{E})^{1/p} \tag{37}$$

$$\leq (\tau + \tau_2)K, \tag{38}$$

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

Now we decompose the probability as the following:

$$\mathbb{P}\left(\Big\|\frac{1}{n}\sum_{i=1}^n Y_i^2 X_i X_i^\top - I\Big\|_{\text{op}} \geq t\right) \leq \mathbb{P}\left(\Big\|\frac{1}{n}\sum_{i=1}^n Y_i^2 X_i X_i^\top - I\Big\|_{\text{op}} \geq t|\mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)$$

$$\leq \underbrace{\mathbb{P}\left(\Big\|\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2\widetilde{X}_i\widetilde{X}_i^\top - \mathbb{E}[\widetilde{Y}^2\widetilde{X}\widetilde{X}^\top]\Big\|_{\text{op}} \geq t/2\right)}_{(a)}$$

$$+ \underbrace{\mathbb{P}\left(\|\mathbb{E}[\widetilde{Y}^2\widetilde{X}\widetilde{X}^\top] - I\|_{\mathrm{op}} \geq t/2\right)}_{(b)} + \underbrace{\mathbb{P}(\mathcal{E}^c)}_{(c)}.$$

We can use a measure of concentration for random matrices for (a) given that $n \geq Cd$ for sufficiently large $C > 0$ (Vershynin, 2010), and bound by $\exp\left(-\frac{nt^2}{C(\tau+\tau_2)^4} + C'd\right)$ for some constants $C, C' > 0$. The bound for (c) is given by $n\exp(-\tau^2)$, hence we set

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

Finally, for (b), we first note that

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

Rearranging the terms,

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

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

Lemma 11. *Let X, Y be the random variables as in Lemma 9. Suppose $\|\theta^*\| \leq C$ for some universal constant $C > 0$. Then for any given $r > 0$, with probability at least $1-\delta$, we have*

$$\sup_{\theta:\|\theta\|\leq r}\left\|\frac{1}{n}\sum_{i=1}^n Y_iX_i\tanh\left(Y_iX_i^\top\theta\right) - M_{mlr}(\theta)\right\| \leq cr\sqrt{\frac{d\ln^2(n/\delta)}{n}}, \tag{39}$$

for some universal constant $c > 0$.

Proof. We start with the standard discretization argument for bounding the concentration of measures in l_2 norm. Let $Z(\theta) := \frac{1}{n}\sum_{i=1}^n Y_iX_i\tanh\left(Y_iX_i^\top\theta\right) - M_{\mathrm{mlr}}(\theta)$. The standard symmetrization argument gives that (van der Vaart and Wellner, 1996; Wainwright, 2019).

$$\mathbb{P}\left(\sup_{\|\theta\|\leq r}\|Z(\theta)\| \geq t\right) \leq 2\mathbb{P}\left(\sup_{\|\theta\|\leq r}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_iY_iX_i\tanh\left(Y_iX_i^\top\theta\right)\right\| \geq t/2\right), \tag{40}$$

where ε_i are independent Rademacher random variables. We define a good event $\mathcal{E} := \{\forall i \in [n], |Y_i| \leq \tau, |X_i^\top\theta^*| \leq C\tau\}$ as before, where $\tau = \Theta\left(\sqrt{\log(n/\delta)}\right)$. Then the probability defined in (40) can be decomposed as

$$\mathbb{P}\left(\sup_{\|\theta\|\leq r}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_iY_iX_i\tanh\left(Y_iX_i^\top\theta\right)\right\| \geq t/2\middle|\mathcal{E}\right) + P(\mathcal{E}^c).$$

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

$$\mathbb{E}\left[\exp\left(\sup_{\|\theta\|\leq r}\frac{\lambda}{n}\left\|\sum_{i=1}^n \varepsilon_iY_iX_i\tanh\left(Y_iX_i^\top\theta\right)\right\|\right)\middle|\mathcal{E}\right],$$

where we used Chernoff-Bound with some $\lambda > 0$ for the last inequality. We first go some steps before we can apply the Ledoux-Talagrand contraction arguments (Ledoux and Talagrand, 1991), with $f_i(\theta) := \tanh\left(|Y_i|X_i^\top\theta\right)$. First, we use discretization argument for removing l_2 norm inside the expectation.

$$\mathbb{E}\left[\exp\left(\sup_{\|\theta\|\leq r}\frac{\lambda}{n}\left\|\sum_{i=1}^n \varepsilon_iY_iX_i\tanh\left(Y_iX_i^\top\theta\right)\right\|\right)\middle|\mathcal{E}\right]$$

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

where M is $1/2$-covering number of the unit sphere and $\{u_1,...,u_M\}$ is the corresponding covering set. Now for each u_j, we can apply the Ledoux-Talagrand contraction lemma since $|f_i(\theta_1)-f_i(\theta_2)|\leq |Y_i||X_i^\top\theta_1 - X_i^\top\theta_2|$ for $\theta\in\mathbb{B}(0,r)$:

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\sup_{\|\theta\|\leq r}\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i Y_iX_i^\top u_j\tanh\left(Y_iX_i^\top\theta\right)\right)\middle|\mathcal{E}\right]\\
&\quad= \mathbb{E}\left[\exp\left(\sup_{\|\theta\|\leq r}\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i |Y_i|X_i^\top u_j\tanh\left(|Y_i|X_i^\top\theta\right)\right)\middle|\mathcal{E}\right]\\
&\quad\leq \mathbb{E}\left[\exp\left(\sup_{\|\theta\|\leq r}\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i Y_i^2(X_i^\top\theta)(X_i^\top u_j)\right)\middle|\mathcal{E}\right]\\
&\quad\leq \mathbb{E}\left[\exp\left(\sup_{\|\theta\|\leq r}\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i Y_i^2(X_i^\top v)(X_i^\top u_j)\right)\middle|\mathcal{E}\right],
\end{aligned}
\tag{41}
$$

where we define $v:=\theta/\|\theta\|$.

We have already seen in (36) that $Y_i(X_i^\top u_j)|\mathcal{E}$ is sub-Gaussian with Orcliz norm $O(\tau(1+\|\theta^*\|))=O(\tau)$. Since the multiplication of two sub-Gaussian variables is sub-exponential, it implies that $Y_i^2(X_i^\top v)(X_i^\top u_1)|\mathcal{E}$ is sub-exponential with Orcliz norm $O(\tau^2)$ (Vershynin, 2010). Now we need the lemma for the exponential moment of sub-exponential random variables from Vershynin (2010).

Lemma 12 (Lemma 5.15 in Vershynin (2010))**.** *Let X be a centered sub-exponential random variable. Then, for t such that $t\leq c/\|X\|_{\psi_1}$, one has*

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

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

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

$$
\mathbb{E}\left[\exp\left(4\lambda r\frac{1}{n}\sum_{i=1}^n \varepsilon_i Y_i^2(X_i^\top v)(X_i^\top u_1)\right)\middle|\mathcal{E}\right]\leq \exp(C\lambda^2r^2\tau^4/n),\qquad \forall|\lambda r/n|\leq c/\tau^2,
$$

which yields

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

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

$$
\mathbb{P}\left(\sup_{\theta\in\mathbb{B}(\theta^*,r)}\|Z(\theta)\|\geq t\right)\leq \exp\left(C_0\lambda^2r^2\tau^4/n + C_1d - \lambda t/2\right)+\mathbb{P}(\mathcal{E}^c).
$$

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

F Supplementary Results

In this appendix, we collect an additional result clarifying the initialization in Theorem 1 and the proof for super-linear convergence of population EM operator in very high SNR regime.

F.1 Initialization with Spectral Methods

Lemma 13. *Let $M = \frac{1}{n}\sum_{i=1}^n Y_i^2 X_i X_i^\top - I$ where X, Y are as given in Lemma 9. Let the largest eigenvalue and corresponding eigenvector of M be (λ_1, v_1). Then, there exists universal constants $c_0, c_1 > 0$ such that*

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

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

$$\sin\angle(v_1, \theta^*) \le c_0\left(1 + \frac{1}{\|\theta^*\|^2}\right)\sqrt{\frac{d\log^2(n/\delta)}{n}} \le \frac{1}{10}.$$

Proof. The lemma is a direct consequence of Lemma 10 and matrix perturbation theory (Wainwright, 2019). Note that $\mathbb{E}[Y_i^2 X_i X_i^\top] = I + 2\theta^*\theta^{*\top}$ (e.g., see Lemma 1 in Yi et al. (2016)). □

The above lemma states that when $\|\theta^*\|$ is not too small, we can always start from the well-initialized point where it is well aligned with ground truth θ^*. In low SNR regime where $\|\theta^*\|^2 \lesssim (d/n)^{1/2}$, we cannot guarantee such a well-alignment with θ^* since the eigenvector is perturbed too much. However, the largest eigenvalue can still serve as an indicator that $\|\theta^*\|$ is small. Hence in all cases, we can initialize the estimator with $\theta_n^0 = \max\{0.2, \sqrt{\lambda_1}\}v_1$ to satisfy the initialization condition that we required in Theorem 1.

F.2 Super-Linear Convergence of Population EM Operator in Very High SNR Regime

In this appendix, we prove Lemma 1 on the super-linear convergence behavior of population EM operator in very high SNR regime.

Proof. We start from the following equation:

$$\begin{aligned}\|M_{\mathrm{mlr}}(\theta) - \theta^*\| &= \mathbb{E}[XY(\tanh(YX^\top\theta) - \tanh(YX^\top\theta^*))] \\ &= \mathbb{E}[XY\Delta_{(X,Y)}(\theta)],\end{aligned}$$

where $\Delta_{(X,Y)}(\theta) := \tanh(YX^\top\theta) - \tanh(YX^\top\theta^*)$. We define good events as follows:

$$\begin{aligned}\mathcal{E}_1 &= \{2|X^\top(\theta^* - \theta)| \le |X^\top\theta^*|\}, \\ \mathcal{E}_2 &= \{|X^\top\theta^*| \ge 2\tau\}, \\ \mathcal{E}_3 &= \{|Z| \le \tau\},\end{aligned} \tag{42}$$

where we set $\tau = \Theta\left(\sqrt{\log\|\theta^*\|}\right)$.

Let the good event $\mathcal{E}_{good} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. From Lemma 4, under the good event, we have $\Delta_{(X,Y)}(\theta) \le \exp(-\tau^2)$. To simplify the notation, let $\Delta(\theta) = \Delta_{(X,Y)}(\theta)$ and $W = \nu XX^\top\theta^*\Delta(\theta)$. Then we can decompose the estimation error as the following:

$$\begin{aligned}\|M_{\mathrm{mlr}}(\theta) - \theta^*\| &= \|\mathbb{E}[XZ\Delta(\theta)] + \mathbb{E}[W\Delta(\theta)]\| \\ &\le \sup_{u\in\mathbb{S}^{d-1}} |\mathbb{E}[(X^\top u)Z\Delta(\theta)]| + |\mathbb{E}[(W^\top u)\Delta(\theta)]| \\ &\le \sup_{u\in\mathbb{S}^{d-1}} \sqrt{\mathbb{E}\left[(X^\top u)^2|\Delta(\theta)|\right]}\sqrt{\mathbb{E}\left[Z^2|\Delta(\theta)|\right]}\end{aligned}$$

$$+\sqrt{\mathbb{E}\left[(X^\top u)^2|\Delta(\theta)|\right]}\sqrt{\mathbb{E}\left[(X^\top\theta^*)^2|\Delta(\theta)|\right]}.$$

We use again the event-wise decomposition strategy. For population EM, note that we set $\tau = \Theta(\sqrt{\log\|\theta^*\|})$ unlike in finite-sample EM case in Appendix B.1. We need to prove the following lemma:

Lemma 14. *For any $u \in \mathbb{S}^{d-1}$, we have*

$$\mathbb{E}\left[(X^\top u)^2|\Delta(\theta)|\right] \le 4\exp(-\tau^2/2) + 2(\tau + 2\|\theta-\theta^*\|)/\|\theta^*\|. \tag{43}$$

Furthermore, we have

$$\mathbb{E}\left[(X^\top\theta^*)^2|\Delta(\theta)|\right] \le 4\|\theta^*\|^2\exp(-\tau^2/2) + 8\tau^3/\|\theta^*\| + 4\|\theta-\theta^*\|^3/\|\theta^*\|. \tag{44}$$

On the other hand, we have

$$\mathbb{E}\left[Z^2|\Delta(\theta)|\right] \le 4\exp(-\tau^2/4) + 2(\tau + \|\theta-\theta^*\|)/\|\theta^*\|. \tag{45}$$

Equipped with the above lemma, whenever $\|\theta-\theta^*\| \ge C\tau$ with $\tau = c_2\sqrt{\log\|\theta^*\|}$ for sufficiently large constants $C, c_2 > 0$, we have

$$\begin{aligned}
\mathbb{E}[(X^\top u)^2|\Delta(\theta)|] &\le 5\|\theta-\theta^*\|/\|\theta^*\|,\\
\mathbb{E}[(X^\top\theta^*)^2|\Delta(\theta)|] &\le 5\|\theta-\theta^*\|^3/\|\theta^*\|,\\
\mathbb{E}[Z^2|\Delta(\theta)|] &\le 5\|\theta-\theta^*\|/\|\theta^*\|,
\end{aligned}$$

which yields $\|M_{\mathrm{mlr}}(\theta)-\theta^*\| \le 6\|\theta-\theta^*\|^2/\|\theta^*\|$, given that $\|\theta^*\|$ is sufficiently large and $\|\theta-\theta^*\| \le \|\theta^*\|/10$. □

Proof of Lemma 14 For equation (43), we can check that

$$\begin{aligned}
\mathbb{E}[(X^\top u)^2|\Delta(\theta)|] &\le \mathbb{E}[(X^\top u)^2|\Delta(\theta)||\mathcal{E}_{good}]P(\mathcal{E}_{good}) + \mathbb{E}[(X^\top u)^2|\Delta(\theta)||\mathcal{E}_1^c]P(\mathcal{E}_1^c)\\
&\quad + \mathbb{E}[(X^\top u)^2|\Delta(\theta)||\mathcal{E}_2^c]P(\mathcal{E}_2^c) + \mathbb{E}[(X^\top u)^2|\Delta(\theta)||\mathcal{E}_3^c]P(\mathcal{E}_3^c)\\
&\le \exp(-\tau^2)\mathbb{E}[(X^\top u)^2 1_{\mathcal{E}_{good}}] + \mathbb{E}[(X^\top u)^2|\mathcal{E}_1^c]P(\mathcal{E}_1^c) +\\
&\quad + \mathbb{E}[(X^\top u)^2|\mathcal{E}_2^c]P(\mathcal{E}_2^c) + \mathbb{E}[(X^\top u)^2|\mathcal{E}_3^c]P(\mathcal{E}_3^c).
\end{aligned}$$

We now recall Lemma 1 in Yi et al. (2014), which is given by:

Lemma 15 (Lemma 1 in Yi et al. (2014))**.** *Given vectors $u, v \in \mathbb{R}^d$ and a Gaussian random vector $X \sim \mathcal{N}(0, I)$, the matrix $\Sigma = \mathbb{E}[XX^\top|(X^\top u)^2 > (X^\top v)^2]$ has singular values*

$$\left(1+\frac{\sin\alpha}{\alpha}, 1-\frac{\sin\alpha}{\alpha}, 1, 1, ..., 1\right), \qquad \text{where} \quad \alpha = \cos^{-1}\left(\frac{(u-v)^\top(u+v)}{\|u-v\|\|u+v\|}\right).$$

Furthermore, if $\|v\| \le \|u\|$, then we have

$$P((X^\top u)^2 > (X^\top v)^2) \le \frac{\|v\|}{\|u\|}.$$

Based on the results of Lemma 15, we obtain

$$\|\mathbb{E}[XX^\top|\mathcal{E}_1^c]\|_{\mathrm{op}} \le 2, \quad P(\mathcal{E}_1^c) \le 2\|\theta-\theta^*\|/\|\theta^*\|.$$

From standard property of Gaussian distribution, (see also Lemma 9 in Balakrishnan et al. (2017)), we also have

$$\|\mathbb{E}[XX^\top|\mathcal{E}_2^c]\|_{\mathrm{op}} \le 1, \quad P(\mathcal{E}_2^c) \le 2\tau/\|\theta^*\|.$$

Finally, from standard Gaussian tail bound, $P(\mathcal{E}_3^c) \le 2\exp(-\tau^2/2)$. Plugging these relations, we get equation (43).

Similarly, we can check that

$$\begin{aligned}\mathbb{E}[(X^\top u)^2|\Delta(\theta)|] &\le \exp(-\tau^2)\mathbb{E}[(X^\top\theta^*)^2 1_{\mathcal{E}_{good}}] + \mathbb{E}[(X^\top\theta^*)^2|\mathcal{E}_1^c]P(\mathcal{E}_1^c) + \\ &\quad + \mathbb{E}[(X^\top\theta^*)^2|\mathcal{E}_2^c]P(\mathcal{E}_2^c) + \mathbb{E}[(X^\top\theta^*)^2|\mathcal{E}_3^c]P(\mathcal{E}_3^c) \\ &\le \exp(-\tau^2)\mathbb{E}[(X^\top\theta^*)^2] + \mathbb{E}[(X^\top(\theta^*-\theta))^2|\mathcal{E}_1^c]P(\mathcal{E}_1^c) + \\ &\quad + 4\mathbb{E}[\tau^2|\mathcal{E}_2^c]P(\mathcal{E}_2^c) + \mathbb{E}[(X^\top\theta^*)^2|\mathcal{E}_3^c]P(\mathcal{E}_3^c) \\ &\le \exp(-\tau^2)\|\theta^*\|^2 + 4\|\theta^*-\theta\|^3/\|\theta^*\| + 8\tau^3/\|\theta^*\| + 2\|\theta^*\|^2\exp(-\tau^2/2),\end{aligned}$$

which gives equation (44).

Finally, for equation (45),

$$\begin{aligned}\mathbb{E}[Z^2|\Delta(\theta)|] &\le \exp(-\tau^2)\mathbb{E}[Z^2 1_{\mathcal{E}_{good}}] + \mathbb{E}[Z^2|\mathcal{E}_1^c]P(\mathcal{E}_1^c) + \mathbb{E}[Z^2|\mathcal{E}_2^c]P(\mathcal{E}_2^c) + \mathbb{E}[Z^2 1_{\mathcal{E}_3^c}] \\ &\le \exp(-\tau^2) + \mathbb{E}[Z^2]P(\mathcal{E}_1^c) + \mathbb{E}[Z^2]P(\mathcal{E}_2^c) + \sqrt{\mathbb{E}[Z^2]}\sqrt{P(\mathcal{E}_3^c)} \\ &\le 4\exp(-\tau^2/4) + 2\tau/\|\theta^*\| + 2\|\theta-\theta^*\|/\|\theta^*\|,\end{aligned}$$

where we used the independence between Z and $\mathcal{E}_1, \mathcal{E}_2$. This concludes the proof of Lemma 14. □

F.3 Proof of Theorem 3

F.3.1 Convergence in the Population Level

Given the EM updates of location and variance in equation (7), the population version of the EM operation is given as follows:

$$M_{\text{ind}}(\theta) := \mathbb{E}_{(X,Y)}\left[XY\tanh\left(\frac{YX^\top\theta}{1+\|\theta^*\|^2-\|\theta\|^2}\right)\right], \tag{46}$$

We recall some notations we defined in the beginning of the section. $\{v_1, ..., v_d\}$ is the standard basis in the transformed coordinate such that $v_1 = \theta/\|\theta\|$, and $\mathbf{span}(v_1, v_2) = \mathbf{span}(\theta, \theta^*)$. Let x_1, x_2 be $X^\top v_1, X^\top v_2$ respectively. Furthermore, denote $b_1^* = \theta^{*\top} v_1 = \|\theta^*\|\cos\angle(\theta,\theta^*)$, and $b_2^* = \theta^{*\top} v_2 = \|\theta^*\|\sin\angle(\theta,\theta^*)$. We will denote $\Delta = \|\theta^*\|^2 - \|\theta\|^2$.

We can rewrite the form of population operator in equation (46) as follows:

$$\begin{aligned}M_{\text{ind}}(\theta) &= \mathbb{E}_{(X,Y)}\left[XY\tanh\left(\frac{YX^\top\|\theta\|}{1+\Delta}\right)\right] \\ &= \mathbb{E}_{(x_1,x_2,y)}\left[yx_1\tanh\left(\frac{yx_1\|\theta\|}{1+\Delta}\right)\right]v_1 + \mathbb{E}_{(x_1,x_2,y)}\left[yx_2\tanh\left(\frac{yx_1\|\theta\|}{1+\Delta}\right)\right]v_2.\end{aligned}$$

In fact, this expression is equivalent to (10) by replacing $\|\theta\| \leftarrow \frac{\|\theta\|}{1+\Delta}$. Therefore we can use the equation (29) with replacing $\|\theta\|$ such that,

$$\begin{aligned}M_{\text{ind}}(\theta)^\top v_1 &\le \frac{\|\theta\|}{1+\Delta}\left(1 - \frac{3\|\theta\|^2}{(1+\Delta)^2} + \frac{30\|\theta\|^4}{(1+\Delta)^4}\right) + c_1\frac{\|\theta\|}{1+\Delta}\|\theta^*\|^2, \\ M_{\text{ind}}(\theta)^\top v_2 &\le c_2\frac{\|\theta\|}{1+\Delta}\|\theta^*\|^2,\end{aligned}$$

for some absolute constants $c_1, c_2 > 0$. We will show that $\frac{3}{(1+\Delta)^2} - \frac{30\|\theta\|^2}{(1+\Delta)^4} \ge 1.25$ whenever $\|\theta\| < 0.2$. Then we can conclude that $|M_{\text{ind}}(\theta)^\top v_1| \le \|\theta\|(1 - 0.25\|\theta\|^2 + O(\|\theta^*\|^2))$.

Now it is easy to check that

$$\frac{3}{(1+\Delta)^2} - \frac{30\|\theta\|^2}{(1+\Delta)^4} = \frac{3(1+\Delta)^2 - 30\|\theta\|^2}{(1+\Delta)^4} = \frac{3 - 36\|\theta\|^2 + 6\|\theta^*\|^2 + 3\Delta^2}{(1+\Delta)^4}.$$

If $\|\theta\| < 0.2$, then $|\Delta| < 0.04$, $3 - 36\|\theta\|^2 \geq 1.5$ and $(1+\Delta)^4 \leq 1.16$, giving the desired bound for the first coordinate. Note that the second coordinate is already less than $O(\|\theta\|\|\theta^*\|)$.

We can also check that this is the best speed at which EM can converge. Observe that

$$\begin{aligned}\|M_{\text{ind}}(\theta)\| \geq |M_{\text{ind}}(\theta)^\top v_1| &\geq \frac{\|\theta\|}{1+\Delta}\left(1 - \frac{3\|\theta\|^2}{(1+\Delta)^2}\right) - c_3\frac{\|\theta\|}{1+\Delta}\|\theta^*\|^2 \\ &\geq \|\theta\|\left(1 - 4\|\theta\|^2 - c_4\|\theta^*\|^2\right),\end{aligned}$$

for some absolute constants $c_3, c_4 > 0$ where we simplify the coefficients using $\|\theta\| < 0.2$. Together with the upper bound we can conclude that

$$\|\theta\|(1 - 4\|\theta\|^2 - c_l\|\theta^*\|^2) \leq \|M_{\text{ind}}(\theta)\| \leq \|\theta\|(1 - 0.25\|\theta\|^2 + c_u\|\theta^*\|^2), \tag{47}$$

for some absolute constants $c_l, c_u > 0$, completing the proof.

F.3.2 Uniform Deviations of Finite-Sample EM Operators

Note that we assume $n \gtrsim d\ln^2(n/\delta)/\epsilon^2$ for sufficiently small $\epsilon > 0$. To simplify the notation, we use $\hat{\Sigma}_n = \frac{1}{n}\sum_i X_i X_i^\top$ and $\bar{\sigma}_n^2 = \frac{1}{n}\sum_i Y_i^2 - \frac{1}{n}\sum_i (X_i^\top\theta)^2$. We also let the $\widetilde{M}_{ind}(\theta) := (\sum_{i=1}^n X_i X_i^\top)^{-1}\sum_{i=1}^n Y_i X_i \tanh\left(\frac{Y_i X_i^\top\theta}{1+\Delta}\right)$. Then we can see that

$$\begin{aligned}\|M_{n,\text{ind}}(\theta) - M_{\text{ind}}(\theta)\| &\leq \left\|M_{n,\text{ind}}(\theta) - \widetilde{M}_{ind}(\theta)\right\| + \left\|\widetilde{M}_{ind}(\theta) - M_{\text{ind}}(\theta)\right\| \\ &\leq \|\hat{\Sigma}_n^{-1}\| \underbrace{\left\|\frac{1}{n}\sum_i X_i Y_i\left(\tanh\left(\frac{Y_i X_i^\top\theta}{\sigma_n^2}\right) - \tanh\left(\frac{Y_i X_i^\top\theta}{1+\Delta}\right)\right)\right\|}_{(a)} \\ &\quad + \|\hat{\Sigma}_n^{-1}\| \underbrace{\left\|\frac{1}{n}\sum_i X_i Y_i \tanh\left(\frac{Y_i X_i^\top\theta}{1+\Delta}\right) - \mathbb{E}\left[XY\tanh\left(\frac{YX^\top\theta}{1+\Delta}\right)\right]\right\|}_{(b)} \\ &\quad + \underbrace{\|\hat{\Sigma}_n^{-1} - I\|\left\|\mathbb{E}\left[XY\tanh\left(\frac{YX^\top\theta}{1+\Delta}\right)\right]\right\|}_{(c)}.\end{aligned}$$

For bounding (a), we first note that by the concentration lemmas, we have $\bar{\sigma}_n^2 \approx 1 + \Delta + O(\epsilon)$. It is also easy to verify that $|\tanh(a) - \tanh(b)| \leq |a - b|$ for any $a, b \in \mathbb{R}$. Now for any unit vector $u \in \mathbb{S}^d$,

$$\begin{aligned}&\frac{1}{n}\sum_i (X_i^\top u) Y_i\left(\tanh\left(\frac{Y_i X_i^\top\theta}{\bar{\sigma}_n^2}\right) - \tanh\left(\frac{Y_i X_i^\top\theta}{1+\Delta}\right)\right) \\ &\qquad \leq \frac{1}{n}\sqrt{\sum_i (X_i^\top u)^2 Y_i^2}\sqrt{\sum_i\left(\frac{Y_i X_i^\top\theta}{\bar{\sigma}_n^2} - \frac{Y_i X_i^\top\theta}{(1+\Delta)}\right)^2} \\ &\qquad \leq \frac{1}{n}\sqrt{\|\sum_i Y_i^2 X_i X_i^\top\|_{\text{op}}}\sqrt{\sum_i \epsilon^2 Y_i^2 \frac{(X_i^\top\theta)^2}{(1+\Delta)^2}} \\ &\qquad \leq \frac{\epsilon\|\theta\|}{1+\Delta}\|\frac{1}{n}\sum_i Y_i^2 X_i X_i^\top\|_{\text{op}} \leq 2\epsilon\|\theta\|,\end{aligned}$$

Finally, we can use Lemma 10 to get $(a) \leq O(\epsilon\|\theta\|)$.

For the left two terms, (b) is bounded with applying the Lemma 11 by plugging $\theta \leftarrow \theta/(1+\Delta)$. (c) is bounded by the concentration of $\hat{\Sigma}_n$ in Lemma 9 and the fact $\|M_{\text{ind}}(\theta)\| \leq \|\theta\|$ from (47). The rest of the steps follow the same argument as in the case for known variances (see Appendix B.3). This conclude the Theorem 3.