Math. Stat. Learn. 4 (2021), 143–220
DOI 10.4171/MSL/29

Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations

Yihong Wu and Harrison H. Zhou

Abstract. We analyze the classical EM algorithm for parameter estimation in the symmetric two-component Gaussian mixtures in d dimensions. We show that, even in the absence of any separation between components, provided that the sample size satisfies $n = \Omega(d \log^4 d)$, the randomly initialized EM algorithm converges to an estimate in at most $O(\sqrt{n})$ iterations with high probability, which is at most $O((d/n)^{1/4} \log n)$ in Euclidean distance from the true parameter and within logarithmic factors of the minimax rate of $(d/n)^{1/4}$. Both the nonparametric statistical rate and the sublinear convergence rate are direct consequences of the zero Fisher information in the worst case. Refined pointwise guarantees beyond worst-case analysis and convergence to the MLE are also shown under mild conditions.

This improves the previous result of Balakrishnan, Wainwright, and Yu (2017), which requires strong conditions on both the separation of the components and the quality of the initialization, and that of Daskalakis, Tzamos, and Zampetakis (2017), which requires sample splitting and restarting the EM iteration.

1. Introduction

The Expectation-Maximization (EM) algorithm [8] is a powerful heuristic aiming at approximating the maximal likelihood estimator (MLE) in the presence of latent variables. The general setting can be described as follows: Let (X, Y) be random variables distributed according to some parametrized joint distribution with density $p_{\theta_*}(x, y)$. Observing Y (but not the latent X), the goal is to estimate the true parameter θ_*. Let

$$p_\theta(y) = \int p_\theta(x, y)\, dx$$

denote the marginal density of Y. Given $Y = y$, the MLE for θ_* is

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta} \log p_\theta(y), \tag{1}$$

2020 Mathematics Subject Classification. Primary 62H15; Secondary 62C20.
Keywords. EM algorithm, Gaussian mixture, minimax rates, convergence rate, random initialization.

which is frequently expensive to compute due to the nonconvexity of the likelihood and the computational cost of the marginalization. To this end, the EM algorithm was proposed as an iterative algorithm to approximate the MLE. Given the current estimate θ_t, the next estimate θ_{t+1} is obtained by executing the following two steps:

- "E step": compute

$$Q(\theta|\theta_t) \triangleq \int p_{\theta_t}(x|y) \log p_\theta(x,y)\, dx. \tag{2}$$

- "M step": update

$$\theta_{t+1} = \arg\max_\theta Q(\theta|\theta_t). \tag{3}$$

The algorithm then proceeds by iterating these two steps and generates a sequence of estimators $\{\theta_t : t \geq 0\}$. The interpretation of this methodology is that (3) is equivalent to maximizing the following lower bound of the log-likelihood:

$$\int p_{\theta_t}(x|y) \log \frac{p_\theta(x,y)}{p_{\theta_t}(x|y)}\, dx = \log p_\theta(y) - D\big(p_{\theta_t}(\cdot|y) \| p_\theta(\cdot|y)\big),$$

where $D(\cdot\|\cdot)$ denote the Kullback–Leibler (KL) divergence. Consequently,

$$\log p_\theta(y) - \log p_{\theta_t}(y) \geq Q(\theta|\theta_t) - Q(\theta_t|\theta_t)$$

for any θ, and hence the likelihood along the EM trajectory $\{\theta_t\}$ is nondecreasing.

1.1. Gaussian mixture model

We consider the symmetric two-component Gaussian mixture (2-GM) model in d dimensions:

$$P_\theta = \frac{1}{2}N(-\theta, I_d) + \frac{1}{2}N(\theta, I_d), \tag{4}$$

which corresponds to two equally weighted clusters centered at $\pm\theta$, respectively. Recall that $\cosh(x) = \frac{e^x + e^{-x}}{2}$, $\sinh(x) = \frac{e^x - e^{-x}}{2}$, and $\tanh(x) = \frac{\sinh(x)}{\cosh(x)}$. The density function of P_θ is

$$p_\theta(y) \triangleq \frac{1}{2}\big[\varphi(y-\theta) + \varphi(y+\theta)\big] = \exp(-\|y\|^2/2)\varphi(\theta)\cosh\langle y, \theta\rangle, \tag{5}$$

where φ denotes the standard normal density in $\mathbb{R}^d$, $\|\cdot\|$ denotes the Euclidean norm.

Let $\theta_* \in \mathbb{R}^d$ denote the ground truth. Given i.i.d. samples $Y = (Y_1, \ldots, Y_n) \overset{\text{i.i.d.}}{\sim} P_{\theta_*}$, the goal is to estimate θ_* up to a global sign flip, under the following loss function:

$$\ell(\widehat{\theta}, \theta) \triangleq \min\big\{\|\widehat{\theta} - \theta\|, \|\widehat{\theta} + \theta\|\big\}.$$

Here the latent variables $(X_1, \ldots, X_n)$ correspond to the labels of each sample, which are i.i.d. and equally likely to be ±1 (Rademacher). Then, we have

$$Y_i = X_i \theta_* + Z_i, \tag{6}$$

where $Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_d)$ and are independent of X_i's. Since

$$p_\theta(x, y) \propto e^{-\frac{1}{2}\sum_{i=1}^n \|y_i - x_i\theta\|^2} \propto e^{-\frac{1}{2}\sum_{i=1}^n \|\theta\|^2 - \langle x_i y_i, \theta\rangle},$$

the M-step in (3) simplifies to

$$\begin{aligned}
\theta_{t+1} &= \arg\min_\theta \sum_{i=1}^n \sum_{x_i \in \{\pm\}} \|y_i - x_i\theta\|^2 p_{\theta_t}(x_i|y) \\
&= \arg\min_\theta \left\{ n\|\theta\|^2 - 2\left\langle \theta, \sum_{i=1}^n y_i \mathbb{E}_{\theta_t}[X_i|Y_i = y_i] \right\rangle \right\} \\
&= \frac{1}{n}\sum_{i=1}^n y_i \mathbb{E}_{\theta_t}[X_i|Y_i = y_i],
\end{aligned}$$

where the conditional mean is given by

$$\mathbb{E}_\theta[X|Y = y] = \tanh\langle\theta, y\rangle. \tag{7}$$

Thus, specialized to the symmetric 2-GM model, the EM algorithm takes the following form:

$$\theta_{t+1} = f_n(\theta_t), \tag{8}$$

where

$$f_n(\theta) \triangleq \mathbb{E}_n[Y \tanh\langle\theta, Y\rangle] \triangleq \frac{1}{n}\sum_{i=1}^n Y_i \tanh\langle\theta, Y_i\rangle. \tag{9}$$

In the case of infinite sample size ($n \to \infty$), (9) reduces to the following

$$f(\theta) \triangleq \mathbb{E}[Y \tanh\langle\theta, Y\rangle], \quad Y \sim P_{\theta_*}. \tag{10}$$

We refer to (9) and (10) as the *sample version* and the *population version* of the EM map, respectively.

In the special case of symmetric Gaussian mixture,[1] EM algorithm can also be interpreted as maximizing the likelihood by means of gradient ascent with *constant step size*. Indeed, denote the average n-sample log likelihood by

$$\ell_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \log p_\theta(Y_i) = \mathbb{E}_n[\log p_\theta(Y)] \tag{11}$$

[1] In fact, this holds for any Gaussian mixture distribution, where the center of each component has the same Euclidean norm.

and its population version by

$$\ell(\theta) \triangleq \mathbb{E}[\log p_\theta(Y)], \quad Y \sim P_{\theta_*}. \tag{12}$$

Since $\nabla \ell_n(\theta) = \mathbb{E}_n[\nabla_\theta \log p_\theta(Y)] = -\theta + \mathbb{E}_n[Y \tanh\langle \theta, Y\rangle]$, the EM iteration (8) can be written as in the following gradient ascent form (with step size equal to one)

$$\theta_{t+1} = \theta_t + \nabla \ell_n(\theta_t). \tag{13}$$

Recently there has been a sequence of work on the performance of the EM algorithm [1, 6, 19, 42], in particular, on the global convergence of the population version. For finite sample size, either strong conditions on the initializations and the separation need to be assumed, or certain variants of the algorithm (such as sample splitting or restart) need to be executed. Despite these progress, the performance guarantee of the classical EM algorithm remains not fully understood, especially with random initializations, which are widely adopted in practice. The main focus of this paper is to provide a better understanding of the statistical and computational guarantees for the randomly initialized EM algorithm in high dimensions, thereby assessing the optimality of the EM estimate and the number of iterations needed to reach the statistical optimum. To this end, we consider the symmetric 2-GM model, which has been well-studied in the literature as a prototypical example for both parameter estimation and clustering [1, 6, 21, 24, 26, 42].

1.2. Main results

We focus on the regime of bounded $\|\theta_*\|$. This is the most interesting case for parameter estimation, wherein consistent clustering is impossible but accurate estimation of θ_* is nevertheless possible. In fact, for the purpose of parameter estimation, it is not necessary to impose any separation between the two clusters, since the parameter θ_* is perfectly identifiable even when $\theta_* = 0$ is allowed, in which case the data are simply generated from a single standard Gaussian component.

Formally, throughout the paper we assume that

$$\|\theta_*\| \leq r \tag{14}$$

for some constant r.

Theorem 1. *There exist constants C, C_0 depending only on r, such that the following holds. Assume that $n \geq C d \log^3 d$. Initialize the EM iteration (8) with*

$$\theta_0 = C_0 \Big(\frac{d}{n} \log n\Big)^{1/4} \eta_0,$$

where η_0 is drawn uniformly at random from the unit sphere S^{d-1}. For any $\|\theta_*\| \le r$, with probability $1 - o_n(1)$,

$$\ell(\theta_t, \theta_*) \le C\left(\frac{d}{n}\right)^{1/4} \log n \tag{15}$$

for all $t \ge C\sqrt{n}$.

Theorem 1 provides a statistical and computational guarantee for the EM algorithm for all θ_*, with the worst case occurring for θ_* close to zero. In fact, if $\|\theta_*\| = O((d/n)^{1/4})$, the 2-GM model is statistically indistinguishable from the standard normal model. The following result is a refined version of Theorem 1 under the modest assumption that θ_* is slightly bounded away from zero, which also shows the convergence to the MLE:

Theorem 2. *In the setting of Theorem 1, assume in addition that*

$$\|\theta_*\| \ge C\left(\frac{d}{n}\right)^{1/4} \log n.$$

Then, with probability at least $1 - o_n(1)$,

$$\ell(\theta_t, \theta_*) \le \frac{C}{\|\theta_*\|}\sqrt{\frac{d \log n}{n}} \tag{16}$$

holds for all $t \ge C\log(n)/\|\theta_*\|^2$ and, furthermore, $\lim_{t\to\infty} \theta_t$ exists and coincides with $\widehat{\theta}_{\mathrm{MLE}}$, the unique (up to a global sign change) global maximizer of the likelihood (11) and $\ell(\theta_t, \widehat{\theta}_{\mathrm{MLE}}) = o(1/n)$ for all $t \ge C\log(n)/\|\theta_*\|^2$.

The statistical optimality of the EM estimate can be seen by comparing Theorems 1 and 2 with the following minimax results (which are consequences of Theorem 10 in Appendix B): For any $r \gtrsim 1$ and $n \gtrsim d$, we have

$$\inf_{\widehat{\theta}} \sup_{\|\theta_*\| \le r} \mathbb{E}_{\theta^*}[\ell(\widehat{\theta}, \theta_*)] \asymp \left(\frac{d}{n}\right)^{1/4}, \tag{17}$$

where the infimum is taken over all estimators $\widehat{\theta}$ measurable with respect to $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} P_{\theta_*}$. Furthermore, for any fixed $\|\theta_*\| = s \lesssim 1$ and $n \gtrsim d$, we have

$$\inf_{\widehat{\theta}} \sup_{\|\theta_*\| = s} \mathbb{E}_{\theta^*}[\ell(\widehat{\theta}, \theta_*)] \asymp \min\left\{s, \frac{1}{s}\sqrt{\frac{d}{n}}\right\}. \tag{18}$$

Comparing (17) with (15), we conclude that the performance of the EM algorithm is within logarithmic factors[2] of the minimax rate, which can be attained in at most

[2] In the one-dimensional case, it is possible to show that the EM algorithm attains the minimax rate (17) without logarithmic factors; see Corollary 1 in Section 2.

$O(\sqrt{n})$ iterations in the worst case. In addition, (18) shows that the transition from the worst-case rate $(d/n)^{1/4}$ to the parametric rate $\frac{1}{\|\theta_*\|}\sqrt{d/n}$ occurs when $\|\theta_*\|$ exceeds $(d/n)^{1/4}$, in which case the more refined guarantee (16) demonstrates the near-optimality of the EM algorithm and its adaptivity to $\|\theta_*\|$.

We pause to clarify that the main objective of this paper is not to exhibit nearly minimax optimal methods, as other procedures (e.g., spectral method; cf. Appendix B) are known to achieve the minimax rate (17) without the extraneous logarithmic factors, but rather to show the popular EM algorithm with a single random initialization achieves near optimality and, furthermore, approaches the MLE (see Theorem 9). Compared to spectral methods, the statistical advantages of the EM algorithm are inherited from the MLE, including the asymptotic efficiency, which holds for example when the dimension is fixed and the center θ_* is bounded away from zero (cf. [35, Theorem 5.39], for example).

We conclude this subsection with a remark interpreting the results of the preceding theorems:

Remark 1 (Statistical and computational consequences of flat likelihood). In Theorem 1, the statistical estimation rate $O((d/n)^{1/4})$ which is slower than the typical parametric rate. Furthermore, the convergence rate is in fact $O(1/\sqrt{t})$ which is much slower than the typical linear convergence rate that is exponential in t. Both guarantees are tight in the worst case which occurs when $\|\theta_*\| = O((d/n)^{1/4})$, and both phenomena are due to the zero curvature of log likelihood function. To explain this, let us consider the simple setting of one dimension and $\theta_* = 0$.

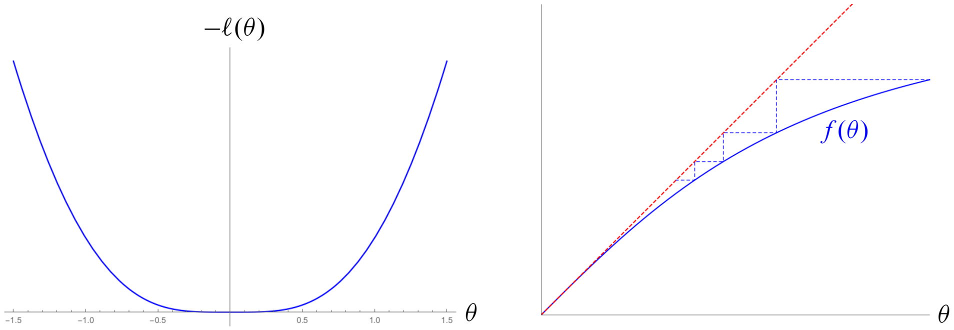
Vanishing Fisher information and nonparametric rate. When $\theta_* = 0$, a simple Taylor expansion shows that the population likelihood (12) satisfies

$$\ell_n(\theta) = \ell_n(0) - \frac{1}{4}\theta^4 + O(\theta^6) \quad \text{when } \theta \to 0,$$

corresponding to the flat maxima at $\theta = 0$ as shown in Figure 1a. In particular, the Fisher information is zero, resulting in an estimation rate slower than the typical rate $\sqrt{d/n}$ for parametric models. Furthermore, for $\theta_* \neq 0$, the Fisher information behaves as $\Theta(\theta_*^2)$ (cf. Remark 2). Therefore (16) shows that the EM algorithm achieves the local minimax rate within logarithmic factors.

Noncontraction and sublinear convergence rate. In typical analysis of iterative methods, linear convergence rate is a direct consequence of contractive mapping theorem. This however fails for the case of $\theta_* = 0$. Indeed, using (13) we obtain that the population EM map $f(\theta)$ satisfies

$$f(\theta) = \theta - \theta^3 + O(\theta^5) \quad \text{with } f'(0) = 1.$$



(a) Flat minimum of the negative log likelihood.

(b) Nonlinear contraction of $f(\theta)$ and the resulting sublinear rate of convergence.

Figure 1. Population version of the negative log likelihood and the EM map for $\theta_* = 0$.

Thus, the EM iteration roughly behaves as $\theta_{t+1} \approx \theta_t - \theta_t^3$. Despite this nonstrict contraction, the iteration nevertheless converges monotonically to the unique fixed point at zero (see Figure 1b); however, the resulting convergence rate is $O(1/\sqrt{t})$ (cf. Lemma 22 in Appendix A). This gives theoretical quantification of the slow convergence rate of EM algorithm for poorly separated Gaussian mixtures, which has been widely observed in practice [20, 28].

1.3. Related work

Since the original paper [8], the EM algorithm has been widely used in Gaussian mixture models [28, 43]. As can be seen from its gradient ascent interpretation (13), a limiting point of the EM iteration is only guaranteed to be a critical point of the likelihood function rather than the global MLE. Various techniques for choosing the initialization has been proposed (cf. the survey [20] and the references therein); however, in practice random initializations are often preferred due to its simplicity over more costly approaches such as spectral methods [2]. Furthermore, it is well-known in practice [20, 28] that the convergence of the EM iteration can be very slow when the components are not well separated, agreeing with the theoretical findings in Theorem 1 and Theorem 2.

Recently there is a renewed interest on the EM algorithm in high dimensions from both statistical and optimization perspectives. General conditions (such as strong concavity and smoothness) are given in [1] to guarantee the local convergence of the EM algorithm as well as its statistical performance. Particularized to the 2-GM model (4), the result [1, Corollary 2] shows that if $\|\theta_*\|$ exceeds some large constant and the initialization satisfies $\|\theta_0 - \theta_*\| \le \frac{1}{4}\|\theta_*\|$, then with probability $1-\delta$ the EM iteration converges exponentially fast to a neighborhood at θ_* of radius $\sqrt{Cd/n \log(1/\delta)}$ for some constant C depending on $\|\theta_*\|$. There are two major distinctions between [1] and the current paper: First, the requirement on the initialization in [1] is very strong, which implies that θ_0 has a nontrivial angle with θ_* and clearly cannot be afforded by random initializations. Second, to bound the deviation between the sample EM trajectory and its population counterpart, [1] proved that

$$\sup_{\|\theta\|\leq C} \|f_n(\theta) - f(\theta)\| = \widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$$

with high probability, where $\widetilde{O}(\cdot)$ hides logarithmic factors. Such a concentration inequality in terms of *absolute deviation* is too weak to yield the sharp rates in Theorem 1 and 2 even in one dimension. Instead, in order to obtain the optimal statistical and computational guarantees, it is crucial to bound the *relative deviation* and show that with high probability,

$$\sup_{\|\theta\|\leq C} \frac{\|f_n(\theta) - f(\theta)\|}{\|\theta\|} = \widetilde{O}\left(\sqrt{\frac{d}{n}}\right) \tag{19}$$

i.e., $f_n - f$ is $\widetilde{O}(\sqrt{d/n})$-Lipschitz at zero, the reason being that when the iterates are close to zero, the finite-sample deviation is proportionally small as well. In addition, in Section 6 we show that the EM iterations converge to the MLE under mild conditions.

The global convergence of the population EM iterates has been analyzed in [6,42]. The following deterministic result was shown: Provided that the initial value θ_0 is not orthogonal to θ_*, the population version of the EM iteration, that is, the sequence (8) with f_n replaced by f, converges to the global maximizer of the population log likelihood ℓ in (12), namely, θ_* (resp., $-\theta_*$) if $\langle\theta_0, \theta_*\rangle > 0$ (resp., < 0). If $\langle\theta_0, \theta_*\rangle = 0$, then the population EM iteration converges to 0, the unique saddle point of ℓ. For the sample EM, [42, Theorem 7] showed that when the dimension and θ_* are fixed, the difference of the sample and population EM iteration vanishes in the double limit of $t \to \infty$ followed by $n \to \infty$; neither finite-sample nor finite-iteration guarantees are provided. As for high dimensions, a variant of the EM algorithm using sampling splitting is analyzed in [6] consisting of two steps: First, run EM with a random and sufficiently small initialization for $\Theta(\log(d)/\|\theta_*\|^2)$ iterations. Next, renormalize the resulting estimate so that its norm is a large constant, and continue to run EM for another $\Theta(1/\|\theta_*\|^2 \log(1/\varepsilon))$ iterations. The final output achieves a loss of ε with high probability provided that each iteration operates on a fresh batch of $\widetilde{\Theta}(d/\varepsilon^2\|\theta_*\|^4)$ samples. The use of sampling splitting conveniently ensures independence among iterations and circumvents the major difficulty of analyzing the entire trajectory; however, for the desired accuracy of $\varepsilon = O(1/\|\theta_*\|\sqrt{d/n})$, the total number of samples is $\widetilde{\Theta}(n/\|\theta_*\|^2)$, which far exceeds n when $\|\theta_*\|$ is small.

Based on the population results in [42], the paper [25] shows that if $\|\theta_*\|$ is at least a constant, the landscape of the log likelihood ℓ_n is close to that of the population version (in terms of the critical points and the Hessian). Specifically, [25, Theorem 8] showed the following: There exist constants C, C' depending on $\|\theta_*\|$ and δ, such that if $n \geq Cd \log d$, then with probability $1 - \delta$, ℓ_n has two local maxima in the ball $B(0, C')$, which are within Euclidean distance $C\sqrt{d \log(n)/n}$ of $\pm\theta_*$. As a corollary of the empirical landscape analysis, with appropriately chosen parameters and initialized from any point in $B(0, C')$, standard trust-region method (cf. [5, Algorithm 6.1.1], for example) is guaranteed to converge to a local maximizer of ℓ_n. It should be noted that trust-region method is a second-order method using the Hessian information, which is more expensive than first-order methods such as gradient descent including the EM algorithm (8). Furthermore, the number of iterations needed to reach the statistical optimum is unclear.

On the technical side, the main difficulty of analyzing a sample-driven iterative scheme, such as (8), is the dependency between the iterates $\{\theta_t\}$ and the data, since each iteration takes one pass over the same set of samples. Of course, one can conduct a uniform analysis by taking a supremum over the realization of θ_t; however, since the supremum is over a d-dimensional space, the resulting bound is too crude to characterize the growth of the "signal" $\langle\theta_*, \theta_t\rangle$, which is very close to zero initially (that is, $O_P(1/\sqrt{d})$, due to random initialization). It is for this reason that the analysis is significantly more challenging than those using sample splitting such as [1,6], which sidesteps the difficulty of dependency. Furthermore, such *trajectory analysis*, which tracks the signal growth from random initializations, does not follow from landscape analysis.

In this vein, the most related to the current paper is the recent seminal work [4] on analyzing gradient descent for nonconvex phase retrieval with random initializations, where the goal is to recover a d-dimensional signal x_* from noiseless quadratic measurements $\langle a_i, x_*\rangle^2$ with i.i.d. Gaussian a_i. To overcome the aforementioned difficulties due to dependency, the main idea of [4] is two-fold: In addition to the commonly used "leave-one-sample-out" method that analyzes the auxiliary iteration when one measurement is replaced by an independent copy, [4] introduced a "leave-one-coordinate-out" auxiliary iteration where a single coordinate of each measurement vector is replenished with a random sign. This is possible thanks to the rotational symmetry of the Gaussian measurement vectors, which allows one to assume, without loss of generality, that the ground truth is proportional to a coordinate vector. By comparing the auxiliary dynamics to the original one, one can effectively decouple the data and the iterates. The idea of leave-one-coordinate-out turns out to be crucial in our analysis of randomly initialized EM, where we introduce an auxiliary sequence with a randomized label but otherwise identical to the original sequence; on the other hand, we are able to conduct the analysis without resorting to the leave-one-sample-out method. Compared to [4] which relies on the strong convexity of the population objective function and the resulting contraction of the iteration, for the EM algorithm since we do not assume θ_* is bounded away from zero, none of these applies which creates additional challenges for the analysis.

Finally, we note that the very recent and independent work [10,11] obtained a tight analysis of the performance of EM algorithm when the true model is a single Gaussian and the postulated model is an over-specified Gaussian mixture. In particular, guarantees similar to Theorem 1 are shown for the special case of $\theta_* = 0$, and both balanced and unbalanced mixture model are considered as well as the more general location-scale mixtures. More recently, for two-component mixture in high dimensions with known (possibly unequal) weights and nonzero centers, the recent work [38] characterizes the statistical optimality and provides computational guarantee for the corresponding EM algorithm, in which case the EM algorithm enjoys improved statistical accuracy and faster convergence, thanks to the nonvanishing Fisher information in the unbalanced case.

1.4. Notation

Throughout the paper, $c, C, C_0, C_1, \ldots, C', C''$ denote constants whose values vary from place to place and only depend on an upper bound on $\|\theta_*\|$, and the notations $\lesssim$, $\gtrsim$, $\asymp$ are within these constant factors. Since we assume that $\|\theta_*\| \leq r$ for some absolute constant r, these constant factors are absolute as well.

Let $\mathcal{L}(X)$ denote the distribution (law) of a random variable X. The generic notation $\mathbb{E}_n[\cdot]$ denotes the empirical average over n i.i.d. samples, namely,

$$\mathbb{E}_n[f(X)] \triangleq \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

where X_i's are i.i.d. copies of X. We say a random variable X is s-subgaussian (resp., s-subexponential) if

$$\|X\|_{\psi_2} \triangleq \inf\{t > 0 : \mathbb{E}e^{X^2/t^2} \leq 2\} \leq \sqrt{s}$$

(resp., $\|X\|_{\psi_1} \triangleq \inf\{t > 0 : \mathbb{E}e^{|X|/t} \leq 2\} \leq s$).

Let $\|x\|$ denotes the Euclidean norm of a vector x. Let $B(x, R)$ denote the ball of radius R centered at x and $B(0, R)$ is abbreviated as $B(R)$. For any matrix M, $\|M\|_{\mathrm{op}}$ and $\|M\|_{\mathrm{F}}$ denote its operator (spectral) norm and Frobenius norm, respectively.

Standard asymptotic notation is adopted in the paper: For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write

$$a_n = O(b_n) \quad \text{if } a_n \leq C b_n$$

for an absolute constant C and for all n;

$$\begin{aligned} a_n &= \Omega(b_n) && \text{if } b_n = O(a_n); \\ a_n &= \Theta(b_n) && \text{if } a_n = O(b_n) \text{ and } a_n = \Omega(b_n); \\ a_n &= o(b_n) \text{ or } b_n = \omega(a_n) && \text{if } a_n/b_n \to 0 \text{ as } n \to \infty. \end{aligned}$$

In addition, we denote

$$a_n = \widetilde{O}(b_n) \quad \text{if } a_n = O\big(b_n(\log n)^{O(1)}\big),$$

and $\widetilde{\Theta}(\cdot)$ is similarly defined.

1.5. Organization

The rest of the paper is organized as follows. Section 2 gives the statistical and computational guarantees for EM algorithm in one dimension, showing the achievability of the optimal average risk up to constant factors. Section 3 states and proves the relative concentration result (19) for the sample EM map. Section 4 presents the analysis of the EM algorithm in d dimensions and give near-optimal statistical and computational guarantees assuming a modest condition on the initialization. In Section 5 we show that starting from a single random initialization, such a condition is fulfilled in at most $O(\log(n)/\|\theta_*\|^2)$ iterations with high probability. Section 6 proves the convergence of the EM iteration to the MLE. Discussions and open problems are presented in Section 7. Proofs for Sections 2–6 are given in Sections 8–12, respectively.

In particular, the main result Theorem 2 previously announced in Section 1.2 follows from Theorem 7 in conjunction with Theorem 8 (on random initialization) and Theorem 9 (on convergence to MLE), while Theorem 1 follows from combining Theorems 2 and 6.

Complementing the performance guarantee on the EM algorithm, Theorem 10 in Appendix B determines the minimax rates for the 2-GM model in any dimension, which may be of independent interest. Auxiliary results are given in Appendix A.

2. EM iteration in one dimension

In this section we present the analysis for the one-dimension case. This turns out to be significantly simpler than the d-dimensional case and we are able to obtain a tighter result; nevertheless, several proof ingredients, both statistical and computational, will re-appear in the analysis for d dimensions later in Section 4. To bound the relative deviation between the sample and population EM trajectories, we use the concentration inequality for empirical distributions under the Wasserstein distance. Although perhaps not crucial, this method simplifies the analysis and yields the optimal rate of the average risk without unnecessary log factors in one dimension.

2.1. Concentration via Wasserstein distance

Recall the 1-Wasserstein distance between probability distributions μ and ν [37]:

$$W_1(\mu, \nu) = \inf \mathbb{E}|X - Y|,$$

where the infimum is over all couplings of μ and ν, i.e., joint law $\mathcal{L}(X, Y)$ such that $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$.

To relate the Wasserstein distance to the EM map, we start with the following simple observation:

Lemma 1. *For any $x, y \in \mathbb{R}$,*

$$\sup_{\theta \in \mathbb{R}} \frac{|x \tanh(x\theta) - y \tanh(y\theta)|}{|\theta|} = |x^2 - y^2|.$$

Proof. Without loss of generality, assume that $x \geq y \geq 0$. Then, by symmetry,

$$\begin{aligned} \sup_{\theta \in \mathbb{R}} \frac{|x \tanh(x\theta) - y \tanh(y\theta)|}{|\theta|} &= \sup_{\theta \geq 0} \frac{|x \tanh(x\theta) - y \tanh(y\theta)|}{\theta} \\ &= \sup_{\theta \geq 0} \frac{x \tanh(x\theta) - y \tanh(y\theta)}{\theta}. \end{aligned} \tag{20}$$

A straightforward calculation gives

$$\frac{\partial}{\partial \theta} \frac{\partial}{\partial x} \left(\frac{x \tanh(x\theta)}{\theta} \right) = \frac{1}{\theta^2 \cosh^2(\theta x)} \left(\theta x - \frac{1}{2} \sinh(2\theta x) - 2(\theta x)^2 \tanh(\theta x) \right) \leq 0,$$

where the inequality follows from $\sinh(t) \geq t$ and $\tanh(t) \geq 0$ for $t \geq 0$. Therefore,

$$\theta \mapsto \frac{\partial}{\partial x} \left(\frac{x \tanh(x\theta)}{\theta} \right)$$

is decreasing on $\mathbb{R}_+$, which implies that the supremum on the right-hand side of (20) is attained at $\theta = 0$. ∎

By coupling, an immediate corollary to Lemma 1 is the following:

Lemma 2. *For any random variables X and Y,*

$$\sup_{\theta \in \mathbb{R}} \frac{|\mathbb{E}[Y \tanh(\theta Y)] - \mathbb{E}[X \tanh(\theta X)]|}{|\theta|} \leq W_1(\mathcal{L}(X^2), \mathcal{L}(Y^2)).$$

As mentioned earlier in Section 1.3, it is crucial to establish the relative deviation in the sense of (19) for the sample EM trajectory. Let $\Delta_n = f_n - f$, where f_n and f are the sample and population EM map defined in (9) and (10). As a consequence of Lemma 2, we have, for all $\theta \in \mathbb{R}$,

$$|\Delta_n(\theta)| \le |\theta| W_1(\nu, \nu_n), \tag{21}$$

where $\nu = \mathcal{L}(Y^2)$ and ν_n is the empirical distribution of the squared samples $Y_1^2, \ldots, Y_n^2$. In other words, Δ_n is $W_1(\nu, \nu_n)$-Lipschitz at zero. To bound the Lipschitz constant, since $\mathbb{E}[\exp(Y^2)] \le C(r)$, applying the concentration inequality in [12, Theorems 1 and 2] (with $d = p = 1$, $\alpha = 2/3$, $\varepsilon = 1/3$ and $\gamma = 1$), we have

$$\mathbb{E}[W_1(\nu, \nu_n)] \le \frac{c_0}{\sqrt{n}} \tag{22}$$

and

$$\begin{aligned}\mathbb{P}[W_1(\nu, \nu_n) \ge x] \le c_1\bigl[&\exp\bigl(-c_2 n x^2\bigr)\mathbf{1}_{\{x \le 1\}} \\ &+ \exp\bigl(-c_2 (nx)^{1/3}\bigr)\mathbf{1}_{\{x \le 1\}} + \exp\bigl(-c_2 (nx)^{2/3}\bigr)\bigr], \quad x > 0 \end{aligned} \tag{23}$$

where c_0, c_1, c_2 depend only on r. Therefore, for any $1 \lesssim a \lesssim n^{1/10}$,

$$\mathbb{P}\left[W_1(\nu, \nu_n) \ge \frac{a}{\sqrt{n}}\right] \le \exp\bigl(-\Omega(a^2)\bigr).$$

2.2. Finite-sample analysis

The population EM map defined in (10) satisfies the following properties:

Lemma 3. *For any $\theta_* \ge 0$,*

1. *$\theta \mapsto f(\theta)$ is an increasing odd and bounded function on $\mathbb{R}$, with*

$$-(1 + \theta_*) \le -\mathbb{E}|Y| = f(-\infty) \le f(\theta) \le f(\infty) = \mathbb{E}|Y| \le 1 + \theta_*.$$

2. *$\theta \mapsto f(\theta)$ is concave on $\mathbb{R}_+$ and convex on $\mathbb{R}_-$.*
3. *$f(0) = 0$, $f'(0) = 1 + {\theta_*}^2$, $f''(0) = 0$, and $f'(\theta_*) \le \exp(-{\theta_*}^2/2)$.*
4. *Define*

$$q(\theta) \triangleq \frac{f(\theta)}{\theta}. \tag{24}$$

Then, q is decreasing on $\mathbb{R}_+$. Furthermore, for $\theta \ge 0$,

$$q'(\theta) = -\mathbb{E}\left[\frac{Y \sinh(2\theta Y) - 2\theta Y^2}{2\theta^2 \cosh^2(\theta Y)}\right] \le -\frac{2\theta}{3}\mathbb{E}\left[\frac{Y^4}{\cosh^2(\theta Y)}\right]. \tag{25}$$

The sample-based EM iterates are given by (8), that is,

$$\theta_{t+1} = f_n(\theta_t).$$

Here the samples $Y_1, \ldots, Y_n$ are i.i.d. drawn from $P_{\theta_*} = \frac{1}{2}N(-\theta_*, 1) + \frac{1}{2}N(\theta_*, 1)$. By the global assumption (30), we have $0 \leq \theta_* \leq r$. Without loss of generality, we assume that $\theta_0 > 0$ for otherwise we can apply the same analysis to the sequence $\{-\theta_t\}$. By (21), $\Delta_n = f_n - f$ is γ_n-Lipschitz at zero, where

$$\gamma_n \triangleq W_1(\nu, \nu_n)$$

is a random variable. Define the high-probability event

$$E = \{\gamma_n \leq c_\gamma\}, \tag{26}$$

where c_γ is a small constant depending only on r that satisfies $c_\gamma < \frac{1}{4}$. By (23), we have $\mathbb{P}[E] \geq 1 - \exp(-\Omega(n^{1/3}))$.

Define the following auxiliary iterations:

$$\begin{cases} \theta_{t+1}^+ = f(\theta_t^+) + \gamma_n \theta_t^+, \\ \theta_{t+1}^- = f(\theta_t^-) - \gamma_n \theta_t^-, \end{cases} \qquad \theta_0^+ = \theta_0^- = \theta_0. \tag{27}$$

By Lemma 3, q is decreasing and maps $\mathbb{R}_+$ onto $(0, 1 + \theta_*^2]$. Define

$$\theta^+ \triangleq q^{-1}(1 - \gamma_n), \tag{28}$$

$$\theta^- \triangleq \begin{cases} q^{-1}(1 + \gamma_n) & |\theta_*| \geq \sqrt{\gamma_n}, \\ 0 & |\theta_*| < \sqrt{\gamma_n}. \end{cases} \tag{29}$$

We will show that on the high-probability event (26), the EM iterates $\{\theta_t\}$ is sandwiched between the two auxiliary iterates $\{\theta_t^+\}$ and $\{\theta_t^-\}$ (see Figure 2). This is made precise by the following theorem, which gives the estimation error bound and finite-iteration guarantees for the EM algorithm in one dimension:

Theorem 3 (Statistical and computational guarantees for one-dimensional EM). *Assume that*

$$0 \leq \theta_* \leq r \tag{30}$$

for some constant r. Assume that

$$0 < \theta_0 \leq r_0.$$

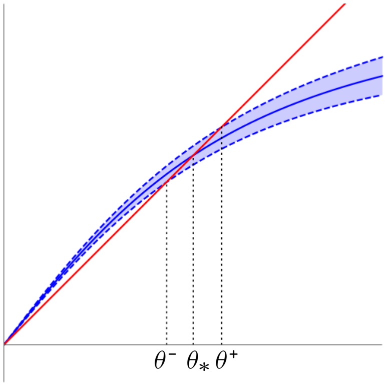


Figure 2. Perturbed EM trajectory and fixed points. The blue solid curve is the population EM map $f(\theta)$ and the two dashed curves correspond to its perturbation $f(\theta) \pm \gamma_n \theta$ in (27).

Then, there exist constants $\tau_1, \dots, \tau_4$ depending on r only, and a constant $n_0 = n_0(r, r_0)$, such that for all $n \geq n_0$, in the event (26)*, the following hold:*

1. *For all $t \geq 0$,*

$$0 \leq \theta_t^- \leq \theta_t \leq \theta_t^+ \leq \tau_1. \tag{31}$$

2. *The inequality*

$$\ell(\theta_t, \theta_*) \leq \tau_2 \min\left\{\frac{\gamma_n}{\theta_*}, \sqrt{\gamma_n}\right\}, \tag{32}$$

holds for all $t \geq T = T(\theta_0, \theta_*, \gamma_n)$, where

$$T = \begin{cases} \tau_3/\gamma_n, & \theta_* \leq \tau_4\sqrt{\gamma_n}, \\ \tau_3/\theta_*^2 \log(1/\theta_0\gamma_n), & \theta_* \geq \tau_4\sqrt{\gamma_n}, \end{cases} \tag{33}$$

and $\gamma_n = W_1(\nu, \nu_n)$.

A corollary of Theorem 3 is the following guarantee on the average risk:

Corollary 1. *There exist constants c_1, c_2 depending only on r, such that*

$$\mathbb{E}[\ell(\theta_t, \theta_*)] \leq c_1 \min\left\{\frac{1}{\theta_*\sqrt{n}}, \frac{1}{n^{1/4}}\right\}, \tag{34}$$

holds for all

$$t \geq c_2 \min\left\{\sqrt{n}, \frac{1}{\theta_*^2}\right\} \log\frac{n}{\theta_0}. \tag{35}$$

Remark 2. The rate in (34) is optimal in the following sense: the second term $n^{-1/4}$ matches the minimax lower bound in Appendix B, while the first term corresponds to the local minimax rate since the Fisher information behaves as $\Theta(\theta_*^2)$ for small θ_*.[3] Indeed, we will show in Section 6 that the EM iteration converges to the MLE which is asymptotically efficient.

In the special case of $\theta_* = 0$, results similar to Theorem 3 have been shown in [11, Theorem 3]. Furthermore, [11, Theorem 4] provided a matching lower bound showing that any limiting point of the EM iteration is $\Omega(n^{-1/4})$ with constant probability.

Computationally, suppose we initialize with $\theta_0 = 1$. Then, regardless of the value of θ_*, we have the worst-case computational guarantee: With high probability, the EM algorithm achieves the optimal rate (34) in at most $O(\sqrt{n}\log n)$ iterations. The number of needed iterations can be pre-determined on the basis of n and θ_0, without knowing θ_*.

3. Concentration of the EM trajectory: relative error bound

Recall that $\Delta_n = f_n - f$ denotes the difference between the sample and the population EM maps. In one dimension, we have shown that the random function $\Delta_n : \mathbb{R} \to \mathbb{R}$ is $O_P(1/\sqrt{n})$-Lipschitz at zero by means of the Wasserstein distance between the empirical distribution and the population. The goal of this section is to extend this result to d dimensions, by showing with high probability that $\Delta_n : \mathbb{R}^d \to \mathbb{R}^d$ is $O(\sqrt{d\log(n)/n})$-Lipschitz at zero with respect to the Euclidean distance on a ball of radius $R = \Theta(\sqrt{d})$.[4] Since with high probability the EM map f_n takes values within this radius, this result allows us to control the fluctuation of the EM trajectory with respect to its population counterpart proportionally to the distance to the origin. This *relative error bound* given next is crucial for obtaining the optimal statistical and computational guarantees.

Theorem 4. *Assume that* $\|\theta_*\| \le r$ *and*

$$n \ge Cd\log d$$

for some universal constant C. *There exist universal constants* c_0, C_0, *such that with probability at least* $1 - \exp(-c_0 d\log n)$, *the following hold:*

[3] Indeed, by Taylor expansion and the dominated convergence theorem, we have

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\Big[\Big(\frac{\partial \log p_\theta(Y)}{\partial\theta}\Big)^2\Big] = \mathbb{E}_\theta\big[\big(Y\tanh(\theta Y) - \theta\big)^2\big] \\ &= \theta^2\big(\mathbb{E}_\theta\big[(Y^2-1)^2\big] + o(1)\big) = (2+o(1))\theta^2 \quad \text{as } \theta \to 0. \end{aligned}$$

[4] It is also possible to show that Δ_n is $O(\sqrt{d\log^3(n)/n})$-Lipschitz at zero on the entire space $\mathbb{R}^d$.

1. *For all* $\theta \in \mathbb{R}^d$, $f_n(\theta) \in B(R)$, *where* $R = 10(\sqrt{d} + r)$.
2. *The function* Δ_n *is* L*-Lipschitz at zero on* $B(R)$, *where*

$$L = C_0(1 + r)\sqrt{d/n \log n}.$$

The proof is given in Section 9. We note that it is straightforward to extend the argument in one dimension (cf. (21)–(22)) to bound the Lipschitz constant of Δ_n by the Wasserstein (in fact, W_2) distance between the empirical distribution and the population. Nevertheless, it is well-known that the Wasserstein distance suffers from the curse of dimensionality; for example, the W_1 distance behaves as $O_P(n^{-1/d})$ (cf. [12, 33], for example). This effect is due to the high complexity of Lipschitz functions in d dimensions. In contrast, the EM map (9) depends on the d-dimensional randomness only through its *linear projections*, and the fact that the sliced Wasserstein distance (i.e., maximal W_1-distance between one-dimensional projections) behaves as $\widetilde{O}_P(\sqrt{d/n})$ suggests that it is possible to obtain a similar guarantee for the EM algorithm.

4. Analysis in d dimensions

In this section we analyze the EM algorithm in high dimensions. By using properties of the population EM iteration in Section 4.1 and the relative deviation bound in Section 3, in Section 4.2 we prove optimal statistical and computational guarantees for the sample EM iteration, assuming a modest condition on the initialization which is much weaker than those in [1]. Although not necessarily satisfied by random initialization, later in Section 5 we show that randomly initialized EM iteration will eventually fulfill such a condition with high probability.

4.1. Properties of the population EM map

Consider the population version of the EM iterates, driven by the population EM map (10):

$$\boldsymbol{\theta}_{t+1} = f(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_0 = \theta_0.$$

We use bold face to delineate it from the finite-sample iteration (8). Let $\eta_* = \theta_*/\|\theta_*\|$. Let

$$\theta_0 = \alpha_0\eta_* + \beta_0\xi_0,$$

where $\xi_0 \perp \theta_*$ and $\|\xi_0\| = 1$, so that $\mathrm{span}(\theta_0, \theta_*) = \mathrm{span}(\eta_*, \xi_0)$. The next lemma shows that the population EM iterates cannot escape the two-dimensional subspace spanned by θ_* and θ_0:

Lemma 4. *For each $t \geq 1$,*

$$\boldsymbol{\theta}_t \in \operatorname{span}(\theta_*, \theta_0). \tag{36}$$

Furthermore, let

$$\boldsymbol{\theta}_t = \boldsymbol{\alpha}_t \eta_* + \boldsymbol{\beta}_t \boldsymbol{\xi}_t,$$

where $\boldsymbol{\xi}_t \perp \eta_*$ and $\|\boldsymbol{\xi}_t\| = 1$. Then, $\{(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\}$ satisfies the recursion

$$\boldsymbol{\alpha}_{t+1} = F(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t), \tag{37}$$

$$\boldsymbol{\beta}_{t+1} = G(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t), \tag{38}$$

where

$$F(\alpha, \beta) \triangleq \mathbb{E}\big[V \tanh(\alpha V + \beta W)\big], \tag{39}$$

$$G(\alpha, \beta) \triangleq \mathbb{E}\big[W \tanh(\alpha V + \beta W)\big], \tag{40}$$

with $W \sim N(0, 1)$ and $V \sim \frac{1}{2}N(-\|\theta_*\|, 1) + \frac{1}{2}N(\|\theta_*\|, 1)$ being independent.

Proof. It suffices to show (36), which was proved in [42]. To give some intuitions, we provide a simple argument below by induction on t. Clearly, (36) holds for $t = 0$. Next, fix any $u \in \operatorname{span}(\theta_*, \theta_0)^\perp$. By the induction hypothesis, $u \perp \boldsymbol{\theta}_t$. Therefore,

$$\langle u, \boldsymbol{\theta}_{t+1}\rangle = \mathbb{E}\big[\langle u, Y\rangle \tanh\big(\langle Y, \boldsymbol{\theta}_t\rangle\big)\big] = \mathbb{E}\big[\langle u, Z\rangle \tanh\big(\langle \theta_*, \boldsymbol{\theta}_t\rangle X + \langle Z, \boldsymbol{\theta}_t\rangle\big)\big] = 0$$

since $\langle u, Z\rangle$, $\langle \boldsymbol{\theta}_t, Z\rangle$ and X are mutually independent. This proves that (36) holds for $t + 1$. ■

Next, we analyze the convergence of $(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$. Without loss of generality (otherwise we can negate θ_* and ξ), we assume that

$$\boldsymbol{\alpha}_0 \geq 0, \quad \boldsymbol{\beta}_0 \geq 0.$$

Therefore, $\boldsymbol{\theta}_t \to \theta_*$ is equivalent to $\boldsymbol{\alpha}_t \to \|\theta_*\|$ and $\boldsymbol{\beta}_t \to 0$. The convergence is easily justified by the following lemma:

Lemma 5 (Properties of F and G). *For any α and $\beta \geq 0$, we have*

1. *$\alpha \mapsto F(\alpha, \beta)$ is increasing, odd, concave (resp., convex) on $\mathbb{R}_+$ (resp., $\mathbb{R}_-$), with $F(0, \beta) = 0$, $F(\pm\|\theta_*\|, 0) = \pm\|\theta_*\|$.*
2. *$F(\alpha, \beta) \geq 0$ for any $\alpha \geq 0$.*
3. *$\beta \mapsto G(\alpha, \beta)$ is increasing and concave, with $G(\alpha, 0) = 0$.*
4. *$\alpha \mapsto G(\alpha, \beta)$ is even, decreasing on $\mathbb{R}_+$; $\beta \mapsto F(\alpha, \beta)$ is decreasing for $\alpha \geq 0$ and increasing for $\alpha \leq 0$.*

5. *Boundedness:*

$$|F(\alpha, \beta)| \leq \|\theta_*\| + \sqrt{2/\pi}, \quad 0 \leq G(\alpha, \beta) \leq \sqrt{2/\pi}.$$

6. *The following holds:*

$$G(\alpha, \beta) \leq G(0, \beta) = \mathbb{E}[W \tanh(\beta W)].$$

7. *The following hold:*

$$f(\alpha) \geq F(\alpha, \beta) \geq f(\alpha) - \big(1 + \|\theta_*\|^2\big)\alpha\beta^2, \quad \alpha \geq 0, \tag{41}$$

$$f(\alpha) \leq F(\alpha, \beta) \leq f(\alpha) - \big(1 + \|\theta_*\|^2\big)\alpha\beta^2, \quad \alpha \leq 0, \tag{42}$$

where

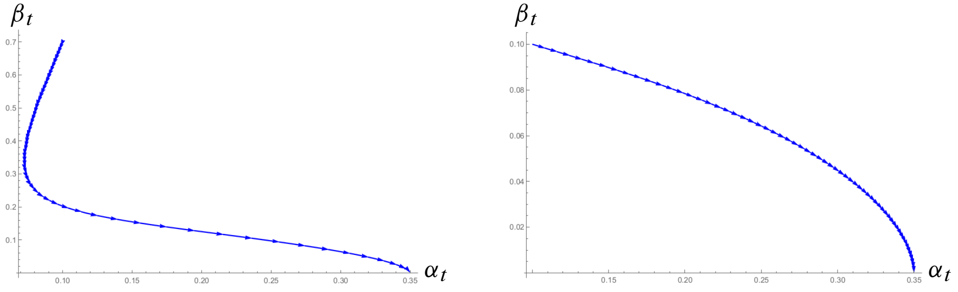
$$f(\alpha) \triangleq F(\alpha, 0) = \mathbb{E}[V \tanh(\alpha V)] \tag{43}$$

coincides with the one-dimensional EM map defined in (10) *with* θ_* *replaced by* $\|\theta_*\|$.

8. *The following holds:*

$$G(\alpha, \beta) \leq \beta\left(1 - \frac{\alpha^2 + \beta^2}{2 + 4(\alpha^2 + \beta^2)}\right). \tag{44}$$

From Lemma 5 it is clear that in the population case, the only fixed points are the desired $(\pm\|\theta_*\|, 0)$ and undesired $(0, 0)$. As long as the initial value is not orthogonal to the ground truth (i.e., $\boldsymbol{\alpha}_0 \neq 0$), $\boldsymbol{\theta}_t$ converges to $\pm\theta_*$; this has been previously shown in [6,42]. In fact, the orthogonal component $\boldsymbol{\beta}_t$ converges to 0 monotonically regardless of the signal component $\boldsymbol{\alpha}_t$. Furthermore, if we start out with $\boldsymbol{\alpha}_0 > 0$, then $\boldsymbol{\alpha}_t > 0$ remains true for all t, and when $\boldsymbol{\beta}_t$ gets sufficiently close to 0, $\boldsymbol{\alpha}_t$ converges to $\|\theta_*\|$ following the one-dimensional EM dynamics (cf. (43)). However, a major distinction between the one-dimensional and d-dimensional case is that $\boldsymbol{\alpha}_t$ need *not* converge monotonically even in the infinite-sample setting. In fact, if the initial value has little overlap with the ground truth (as is the case for random initialization in high dimensions), $\boldsymbol{\beta}_t$ is large initially which causes $\boldsymbol{\alpha}_t$ to decrease and $\boldsymbol{\theta}_t$ to move closer to the undesired fixed point at zero (see Figure 3a). Therefore, in the finite-sample setting, we need to assume conditions on the initialization (namely lower bound on $|\alpha_t|$) in order to avoid being trapped near zero – we will return to this point in the finite-sample analysis in the next subsection. This is in stark contrast to the one-dimensional case: even with finite samples, for any nonzero initialization, the EM iteration eventually converges to a neighborhood of the ground truth with optimal accuracy (cf. Theorem 3).



(a) Nonmonotone convergence of $\boldsymbol{\alpha}_t$ ($\boldsymbol{\alpha}_0 = 0.1, \boldsymbol{\beta}_0 = 0.7$).

(b) Monotone convergence of $\boldsymbol{\alpha}_t$ ($\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0 = 0.1$).

Figure 3. Convergence of $(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$ in the population dynamics in d dimensions with $\|\theta_*\| = 0.35$ for 60 iterations.

4.2. Finite-sample analysis

We now analyze the n-sample EM iteration (8), that is,

$$\theta_{t+1} = f_n(\theta_t).$$

Write

$$\theta_t = \alpha_t \eta_* + \beta_t \xi_t,$$

where $\xi_t \perp \eta_* = \frac{\theta_*}{\|\theta_*\|}$, $\|\xi_t\| = 1$ and $\beta_t \geq 0$. Thus, $\|\theta_t\| = \sqrt{\alpha_t^2 + \beta_t^2}$.

Recall that $\Delta_n = f_n - f$ denotes the difference between the sample and population EM maps. In view of Theorem 4, with probability at least $1 - \exp(-c_0 d \log n)$, the following event holds:

$$\begin{aligned} \sup_{\theta \in \mathbb{R}^d} \|f_n(\theta)\| &\leq R, \\ \|\Delta_n(\theta)\| &\leq \omega \|\theta\|, \quad \forall \theta \in B(R), \end{aligned} \tag{45}$$

where $R = 10(r + \sqrt{d})$, and

$$\omega \triangleq \sqrt{C_\omega \frac{d}{n} \log n} \tag{46}$$

and C_ω is a constant that only depends on r. We assume that n is sufficiently large so that ω is at most an absolute constant.

Recall from Lemma 4 that $f(\theta) \in \operatorname{span}(\eta_*, \theta)$ for any $\theta \in \mathbb{R}^d$. Furthermore,

$$f(\theta_t) = F(\alpha_t, \beta_t)\eta_* + G(\alpha_t, \beta_t)\xi_t,$$

where F and G are defined in (39)–(40). Therefore,

$$\alpha_{t+1} = \langle \theta_{t+1}, \eta_* \rangle = F(\alpha_t, \beta_t) + \langle \Delta_n(\theta_t), \eta_* \rangle.$$

In view of (45), we have

$$|\langle \Delta_n(\theta_t), \eta_* \rangle| \le \|\Delta_n(\theta_t)\| \le \omega\big(|\alpha_t| + \beta_t\big).$$

Hence,

$$\alpha_{t+1} \le F(\alpha_t, \beta_t) + \omega\big(|\alpha_t| + \beta_t\big) \tag{47}$$

$$\alpha_{t+1} \ge F(\alpha_t, \beta_t) - \omega\big(|\alpha_t| + \beta_t\big) \tag{48}$$

On the other hand, we have

$$(I - \eta_*\eta_*^\top)\theta_{t+1} = G(\alpha_t, \beta_t)\xi_t + (I - \eta_*\eta_*^\top)\Delta_n(\theta_t).$$

Taking norms on both sides, we have

$$\beta_{t+1} \le G(\alpha_t, \beta_t) + \omega\big(|\alpha_t| + \beta_t\big). \tag{49}$$

The equations (47)–(48) and (49) should be viewed as the finite-sample perturbation of the population dynamics (37) and (38), respectively.

We will show that the orthogonal component β_t unconditionally converges to $O(\sqrt{\omega}) = O((d \log(n)/n)^{1/4})$; however, for finite sample size we cannot expect β_t to converge to zero. To analyze α_t, let us assume that β_t have converged to this limiting value (in fact, by initializing near zero, we can ensure $\beta_t = O(\sqrt{\omega})$ for all t). Following the sandwich analysis in one dimension, we can define the auxiliary iterations similarly to (27) to give

$$\begin{cases} \alpha_{t+1}^+ = F(\alpha_t^+, \beta_t) + \omega\alpha_t^+ + \omega^{3/2}, \\ \alpha_{t+1}^- = F(\alpha_t^-, \beta_t) - \omega\alpha_t^- - \omega^{3/2}, \end{cases} \qquad \alpha_0^+ = \alpha_0^- = \alpha_0, \tag{50}$$

and show that the upper bound sequence $\{\alpha_t^+\}$ converges to α^+ which is within the optimal rate of the desired $\|\theta_*\|$. However, due to the additional intercept, the lower bound sequence $\{\alpha_t^-\}$ have two possible fixed points (see Figure 4): the "good" fixed point α^- that is within the optimal rate of $\|\theta_*\|$, and the "bad" fixed point α° that is close to zero (in fact, $\alpha^\circ = O(\sqrt{\omega})$).

Consequently, if the iteration starts from the left of the bad fixed points, i.e., $\alpha_0 < \alpha^\circ$, which is what happens when the initialization is nearly orthogonal to θ_*, the lower bound sequence α_t^- may be stuck at near zero and fail to converge to the desired neighborhood of $\|\theta_*\|$. Thus to rule this out it requires more refined argument than the above sandwich analysis, which is carried out in the next section. For this section we focus on proving the performance guarantee assuming a mild assumption on the initialization. Specifically, we establish the following claims:

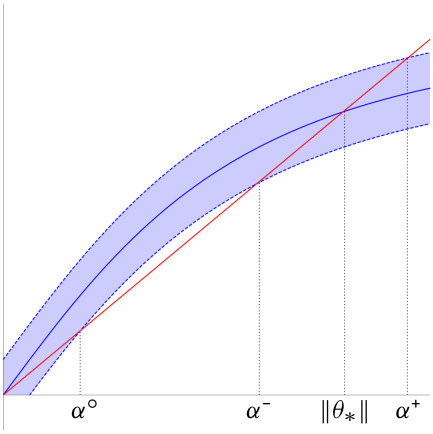


Figure 4. Perturbed EM trajectory for α_t and fixed points.

Orthogonal direction. We show that regardless of the initialization, $\{\beta_t\}$ unconditionally converges to the near-optimal rate $O(\sqrt{\omega})$. In particular, if we start from near zero (and we will), we can ensure that the entire sequence $\{\beta_t\}$ is $O(\sqrt{\omega})$ for *all* t.

Signal direction. We show that:

- For small θ_*, i.e., $\|\theta_*\| = O(\sqrt{\omega})$, $\{|\alpha_t|\}$ unconditionally converges to $O(\sqrt{\omega})$, and hence so does $\|\theta_t - \theta_*\|$.
- For large θ_*, i.e., $\|\theta_*\| = \Omega(\sqrt{\omega})$, provided that the initialization satisfies

$$|\langle \eta_0, \eta_* \rangle| \gtrsim \frac{1}{\|\theta_*\|^2} \sqrt{\frac{d}{n} \log n},$$

the signal part $\{\alpha_t\}$ converges to

$$\|\theta_*\| + O\left(\frac{1}{\|\theta_*\|} \sqrt{\frac{d}{n} \log n}\right).$$

The condition on the initialization improves that of [1], which requires that

$$|\langle \eta_0, \eta_* \rangle| \geq \Omega(1) \quad \text{and} \quad \|\theta_*\| = \Omega(1).$$

Note that if η_0 is drawn uniformly from the unit sphere, we have

$$|\langle \eta_0, \eta_* \rangle| = \Theta_P\left(\frac{1}{\sqrt{d}}\right).$$

Thus, in the special case of $\|\theta_*\|$ being a constant, the above condition is fulfilled when $n = \widetilde{\Omega}(d^2)$. Nevertheless, in Section 5 we will prove the refined result that as long as $n = \widetilde{\Omega}(d)$, starting from a single random initialization, the EM iterates will eventually satisfy the above condition with high probability.

In the rest of the paper, we always assume that the initialization lies in a bounded ball. To simplify the presentation, assume that

$$\|\theta_0\| \le 1. \tag{51}$$

The following theorems are the main result of this section. We note that results similar to Theorems 5–6 have been shown in [11, Theorem 3] in the special case of $\theta_* = 0$. The following result shows the unconditional convergence of β_t to zero within the minimax rate regardless of the ground truth or the initialization. An improved error bound is given later in Theorem 7 for large $\|\theta_*\|$.

Theorem 5 (Unconditional convergence of β_t). *There exist constants $\kappa_0, \kappa_1, \kappa_2$ depending only on r, such that in the event* (45)*, the following hold:*

1. *For all $t \ge 0$,*

$$\beta_{t+1} \le \beta_t(1+\omega) + \omega|\alpha_t| \tag{52}$$

and

$$\beta_{t+1} \le \beta_t(1+\omega) - \frac{\beta_t^3}{2+8\Gamma^2} + \min\left\{\frac{\omega^2(2+8\Gamma^2)}{2\beta_t}, \omega\Gamma\right\}, \tag{53}$$

where $\Gamma = 2 + 2r$.

2. *Consequently, regardless of θ_0,*

$$\limsup_{t\to\infty} \beta_t \le \kappa_1 \left(\frac{d}{n}\log n\right)^{1/4}. \tag{54}$$

3. *Furthermore, if $\omega \le \kappa_0$ and*

$$\|\theta_0\| \le \kappa_2 \left(\frac{d}{n}\log n\right)^{1/4}, \tag{55}$$

then for all $t \ge 0$,

$$\beta_t \le \kappa_2 \left(\frac{d}{n}\log n\right)^{1/4}. \tag{56}$$

Theorem 6 (Small $\|\theta_*\|$: unconditional convergence of α_t). *There exist absolute constants $K, L \ge 1$, such that in the event* (45)*, the following holds: Let s_0 be such that $K\sqrt{\omega} \le s_0 \le 1$. Assume that $\|\theta_*\| \le s_0$.*

1. *Regardless of θ_0,*

$$\limsup_{t\to\infty} |\alpha_t| \le 2s_0, \tag{57}$$

and hence

$$\limsup_{t\to\infty} \ell(\theta_t, \theta_*) \le 3s_0 + \kappa_1 \left(\frac{d}{n}\log n\right)^{1/4}. \tag{58}$$

2. *Furthermore, if the initializer θ_0 satisfies (55), then*

$$|\alpha_t| \leq Ls_0, \tag{59}$$

$$\ell(\theta_t, \theta_*) \leq 2Ls_0 \tag{60}$$

hold for all $t \geq 0$.

Theorem 7 (Large $\|\theta_*\|$: conditional convergence of α_t). *There exist constants $\lambda_0, \ldots, \lambda_4$ depending only on r, such that in the event (45), the following holds: Assume that $\|\theta_*\| \geq \lambda_0\sqrt{\omega}$. Let $\eta_0 \in S^{d-1}$ satisfy*

$$|\langle \eta_0, \eta_* \rangle| \geq \frac{\lambda_2}{\|\theta_*\|^2}\sqrt{\frac{d}{n}\log n}. \tag{61}$$

Set

$$\theta_0 = c\left(\frac{d}{n}\log n\right)^{1/4}\eta_0, \tag{62}$$

where $c \leq \kappa_2$ and κ_2 is from Theorem 5. Then, for all $t \geq \lambda_4 \log(n)/\|\theta_*\|^2$,

$$\big|\alpha_t - \|\theta_*\|\big| \leq \lambda_1 \frac{1}{\|\theta_*\|}\sqrt{\frac{d\log n}{n}} \tag{63}$$

and

$$\ell(\theta_t, \theta_*) \leq \lambda_3 \frac{1}{\|\theta_*\|}\sqrt{\frac{d\log n}{n}}. \tag{64}$$

Remark 3. We can take $s_0 = \lambda_0\sqrt{\omega}$ in Theorem 6, so that Theorems 6 and 7 gives the near-optimal rate of $O((d/n\log n)^{1/4})$ for the case of small and large $\|\theta_*\|$, respectively. Later in the refined analysis in Section 5 we will take s_0 slightly larger than $\sqrt{\omega}$; cf. (65).

Theorems 5–7 are proved in Section 10.1. Here we give a sketch of the proof of Theorem 7. The analysis consists of three phases:

Phase I: $\boldsymbol{\alpha_t \lesssim \sqrt{\omega}}$. By using the condition (61) on the initialization, we show that in this phase α_t increases geometrically according to

$$\alpha_{t+1} \geq \big(1 + \Omega\big(\|\theta_*\|^2\big)\big)\alpha_t.$$

Phase II: $\boldsymbol{\alpha_t \gtrsim \sqrt{\omega}}$. Now that α_t has escaped the undesired fixed point near zero (cf. Figure 4), one can apply the "sandwich bound" (50) to show that α_t follows a perturbed one-dimensional EM evolution

$$\alpha_{t+1} = f(\alpha_t) + O(\omega\alpha_t),$$

where f is defined (43) and coincides with the one-dimensional EM map (10) with θ_* replaced by $\|\theta_*\|$.

Phase III: $\alpha_t \asymp \|\theta_*\|$. Recall that Theorem 5 ensures that β_t converges to the worst-case rate $O(\sqrt{\omega})$. Now that α_t has reached a constant fraction of the desired limit $\|\theta_*\|$, we can obtain improved estimate $\beta_t \lesssim \omega/\|\theta_*\|$, leading to the optimal $\|\theta_*\|$-dependent bound (64).

5. Refined analysis for random initialization: the initial phase

In this section we analyze the EM iterates starting from a single random initialization. Since Theorems 5 and 6 have covered the case of small $\|\theta_*\|$, we only consider the case where $\|\theta_*\| \gg (d/n)^{1/4}$. We provide a refined analysis of Phase I in the proof of Theorem 7: if the initial direction is uniformly chosen at random, then with high probability, the iterates will satisfy $\alpha_t = \Omega(\sqrt{\omega})$ for sufficiently large constant C in at most $O(1/\|\theta_*\|^2 \log n)$ iterations and hence the analysis in the subsequent Phase II and III applies. This was previously shown in Theorem 7 under the stronger assumption (61) which need not be fulfilled by random initializations.

Recall that $\eta_* = \frac{1}{\|\theta_*\|}\theta_*$ denotes the true direction and

$$\alpha_t = \langle \theta_t, \eta_* \rangle, \quad \beta_t = \|(I - \eta_* \eta_*^\top)\theta_t\|.$$

Without loss of generality, we assume the following:

1. Thanks to the rotational invariance of the Gaussian distribution, we can assume that the true center is aligned with a coordinate vector, i.e., $\theta_* = \|\theta_*\| e_1$, so that
$$\alpha_t = \theta_{t,1}, \quad \beta_t = \|\theta_{t,\perp}\| = \|(\theta_{t,2}, \ldots, \theta_{t,d})\|.$$
2. The initialization satisfies $\alpha_0 > 0$. Otherwise, we can apply the same analysis to $\{-\theta_t\}$, which has the same law as $\{\theta_t\}$.

Furthermore, we assume that the ground truth satisfies

$$r \geq \|\theta_*\| \geq \left(\frac{C_\star d}{n}\right)^{1/4} \log n \tag{65}$$

for some absolute constant $C_\star$. Otherwise, applying Theorem 6 (with s_0 being the right-hand side of (65)) shows that regardless of the initialization, we achieve the near optimal rate for all $t \geq 0$:

$$\|\theta_t - \theta_*\| = O\left(\left(\frac{d}{n}\right)^{1/4} \log n\right). \tag{66}$$

Define

$$T_1 \triangleq \min\{t \in \mathbb{N} : \alpha_t > C_* \sqrt{\omega}\},$$

where C_* is some constant depending only on r (cf. (99)) and $\omega = \sqrt{C_\omega \, d/n \log n}$ as defined in (46). The main result of this section is the following:

Theorem 8. *Assume that θ_* satisfies (65). There exist constants C_0, C_1, C_2 depending only on r, such that the following holds: Let*

$$\theta_0 = C_0\Big(\frac{d}{n}\log n\Big)^{1/4}\eta_0, \tag{67}$$

where η_0 is drawn uniformly at random from the unit sphere S^{d-1}. Assume that

$$n \geq C_1 d \log^4 d. \tag{68}$$

Then, with probability at least $1 - \frac{C_2 \log\log n}{\sqrt{\log n}}$,

$$T_1 \leq T_\star \triangleq \frac{C_T(\log d + \log\log n)}{\|\theta_*\|^2}, \tag{69}$$

where C_T is some universal constant.

Theorem 8 shows that after $t \geq T_1$, the iteration enters Phase II and the statistical guarantee in Theorem 7 applies to all subsequent iterations; in particular, the optimal estimation error is achieved in another $O(\log(n)/\|\theta_*\|^2) = O(\sqrt{n/d})$ iterations, proving Theorem 2 previously announced in Section 1.2. Finally, since the case of $\|\theta_*\| = O((d/n)^{1/4}\log n)$ is covered by (66), the worst-case result in Theorem 1 follows.

5.1. Proof of Theorem 8

In this subsection we provide the main argument for proving Theorem 8, with key lemmas proved in Section 11.1. Suppose, for the sake of contradiction, that $\alpha_t \leq \sqrt{\omega}$ for all $t \leq T_\star$. Then, in view of (56), we conclude that for all $t \leq T_\star$,

$$\|\theta_t\| \leq 2C_1\Big(\frac{d}{n}\log n\Big)^{1/4} \tag{70}$$

for some constant C_1. In particular, θ_t belongs to the unit ball in view of the assumption (68).

We now introduce an *auxiliary sequence* of iterates $\{\widetilde{\theta}_t\}$, which is main apparatus for analyzing the initial growth of the signal. Since the law of $Y_{i,1}$ is symmetric, without loss of generality, we view the ith sample as $Y_i = (b_i Y_{i,1}, Y_{i,2}, \ldots, Y_{i,d})$, where b_i's are independent Rademacher variables, and the sample-based EM iterates is

$$\theta_{t+1} = f_n(\theta_t),$$

where

$$f_n(\theta) = \mathbb{E}_n\big[Y \tanh\langle\theta, Y\rangle\big] = \frac{1}{n}\sum_{i=1}^{n} Y_i \tanh\langle\theta, Y_i\rangle.$$

In comparison, the auxiliary iteration is based on the modified samples $(\widetilde{Y}_1, \ldots, \widetilde{Y}_n)$, where $\widetilde{Y}_i = (\widetilde{b}_i Y_{i,1}, Y_{i,2}, \ldots, Y_{i,d})$, the $\widetilde{b}_i$'s are independent Rademacher variables, and $\{\widetilde{b}_i, b_i, Y_i\}$ are mutually independent. Define the auxiliary iterates

$$\widetilde{\theta}_{t+1} = \widetilde{f}_n(\widetilde{\theta}_t),$$

where

$$\widetilde{f}_n(\theta) \triangleq \mathbb{E}_n\big[\widetilde{Y} \tanh\langle\theta, \widetilde{Y}\rangle\big] = \frac{1}{n}\sum_{i=1}^{n} \widetilde{Y}_i \tanh\langle\theta, \widetilde{Y}_i\rangle.$$

Both the main and the auxiliary sequence starts from the same random initialization:

$$\widetilde{\theta}_0 = \theta_0,$$

as specified by (67). The angle of a random initialization satisfies the following:

Lemma 6 (Random initialization). *There exists an absolute constant* C_0*, such that for any* $a \in (0, 1)$*,*

$$\mathbb{P}\left[|\langle\eta_0, e_1\rangle| < \frac{a}{\sqrt{d}}\right] \le C_0 a\sqrt{\log\frac{1}{a}}.$$

Proof. Note that $\langle\eta_0, e_1\rangle$ is equal in distribution to $Z_1/\|Z\|$, where $Z = (Z_1, \ldots, Z_d)$ is standard normal. Therefore,

$$\mathbb{P}\left[|\langle\eta_0, e_1\rangle| < \frac{a}{\sqrt{d}}\right] \le \mathbb{P}\big[\|Z\| \ge \sqrt{Cd}\big] + \mathbb{P}\big[|Z_1| < \sqrt{C}a\big].$$

Take $C = 2 + 3\log(1/a)$. By Lemma 20,

$$\mathbb{P}\big[\|Z\| \ge \sqrt{Cd}\big] \le a^d \le a \quad \text{and} \quad \mathbb{P}\big[|Z_1| < \sqrt{C}a\big] \le \sqrt{2C/\pi}a. \qquad \blacksquare$$

In the following, we conduct the analysis in the event:

$$\alpha_0 \ge \frac{1}{\sqrt{d\log n}}\|\theta_0\|, \tag{71}$$

which holds with probability at least $1 - O(\log(\log(n))/\sqrt{\log n})$, in view of Lemma 6.

The key argument is to show that the signal component α_t grows exponentially according to

$$\alpha_{t+1} \ge \alpha_t\big(1 + \|\theta_*\|^2 - o\big(\|\theta_*\|^2\big)\big). \tag{72}$$

More precisely, we prove a quantitative version of (72) (cf. (76) below).

Lemma 7. *With probability at least* $1 - O(n^{-1/2}\log n)$*, for all* $t = 0, 1, \ldots, T_\star$*, we have*

$$\|\theta_t - \widetilde{\theta}_t\| \leq \alpha_t \sqrt{\frac{Kd\log^2 n}{n}}\, t, \tag{73}$$

$$\frac{\beta_t}{\alpha_t} \leq \sqrt{d\log n} + \omega t, \tag{74}$$

and

$$\alpha_t \geq \frac{1}{\sqrt{Kd\log n}}\|\theta_t\|, \tag{75}$$

$$\alpha_{t+1} \geq \alpha_t\left(1 + \|\theta_*\|^2 - \sqrt{\frac{Kd\log^2 n}{n}}\right), \tag{76}$$

where K *is a constant depending only on* r*.*

The proof of Lemma 7 is by induction on t, relying on the following results that relate the actual iterations to the auxiliary ones.

Lemma 8. *For each* $t \geq 0$*, with probability at least* $1 - O(n^{-1})$*, we have*

$$\alpha_{t+1} \geq \alpha_t\left(1 + \|\theta_*\|^2 - \sqrt{\frac{C\log n}{n}} - C\|\theta_t\|^2\right) - \sqrt{\frac{C\log n}{n}}\|\theta_t\| - \sqrt{\frac{Cd\log n}{n}}\|\theta_t - \widetilde{\theta}_t\|, \tag{77}$$

where C *is some constant depending only on* r*.*

Lemma 9. *For each* $t \geq 0$*, with probability at least* $1 - O(n^{-1})$*, we have*

$$\|\widetilde{\theta}_{t+1} - \theta_{t+1}\| \leq \left(1 + \|\theta_*\|^2 + \sqrt{\frac{Cd\log n}{n}}\right)\|\widetilde{\theta}_t - \theta_t\| + \sqrt{\frac{Cd\log n}{n}}\alpha_t + \sqrt{\frac{C\log n}{n}}\|\theta_t\|, \tag{78}$$

where C *is some constant depending only on* r*.*

Now we complete the proof of Theorem 8. Suppose, for the sake of contradiction, $T_1 > T_\star$, so that $\alpha_t \leq \sqrt{\omega}$ for all $t \leq T_\star$. Since (76) holds for all $t \leq T_\star$, in view of the assumption (65), we have that

$$\alpha_{t+1} \geq \alpha_t\big(1 + c_0\|\theta_*\|^2\big)$$

holds for some constant c_0. Since

$$\alpha_0 \geq \|\theta_0\|\frac{1}{\sqrt{d\log n}} = \frac{C_0}{\sqrt{C_\omega}}\frac{\sqrt{\omega}}{\sqrt{d\log n}},$$

when

$$t \geq T_\star = \frac{C_T(\log d + \log\log n)}{\|\theta_*\|^2}$$

for sufficiently large constant C_T, we have $\alpha_t > \sqrt{\omega} = (C_\omega\, d/n \log n)^{1/4}$, which is the necessary contradiction.

6. Approaching the MLE

Despite being a heuristic of solving the maximum likelihood, in this section we show that the EM iteration converges to the MLE under minimal conditions. Define the MLE as any global maximizer of the likelihood function, i.e.,

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta\in\mathbb{R}^n} \ell_n(\theta), \tag{79}$$

where the log likelihood ℓ_n is given in (11). Note that from first principles it is unclear whether there exists a unique global maximizer (up to a global sign change). Furthermore, our previous analysis only shows that with high probability, the EM iterates are within the optimal rate of the true mean θ_* after a certain number of iterations. Indeed, for $\|\theta_*\| \geq (Cd/n)^{1/4}\log n$, Theorem 7 and Theorem 8 together imply that, with probability $1 - o(1)$,

$$\ell(\theta_t, \theta_*) \leq \left(\frac{Cd\,\log n}{n}\right)^{1/4}$$

for all $t \geq T \triangleq C\,\log(n)/\|\theta_*\|^2$, for some constant C. This, however, has no direct bearing on the convergence of the sequence θ_t, since it does not rule out the possibility that θ_t oscillates within the optimal rate of θ_*. Next we will address both questions by showing that the MLE is unique and coincides with the limit of the EM iteration.

Theorem 9. *Assume that $n \geq C_1 d\log^4 d$ and $(C_2\, d/n)^{1/4}\log n \leq \|\theta_*\| \leq r$, where C_1, C_2 are constants depending only on r. With probability at least $1 - 2n^{-1}$, for all $t \geq 1$,*

$$\|\theta_{T+t} - \widehat{\theta}_{\mathrm{MLE}}\| \leq e^{-ct\|\theta_*\|^2}\|\theta_T - \widehat{\theta}_{\mathrm{MLE}}\|, \tag{80}$$

for some absolute constant c, where $\widehat{\theta}_{\mathrm{MLE}}$ is the maximizer that satisfies

$$\|\theta_T - \widehat{\theta}_{\mathrm{MLE}}\| = \ell(\theta_T, \widehat{\theta}_{\mathrm{MLE}}).$$

In particular, $\lim_{t\to\infty}\theta_t$ exists and coincides with $\widehat{\theta}_{\mathrm{MLE}}$, the unique (up to a global sign change) global maximizer of (79).

Next we prove Theorem 9. Note that $\widehat{\theta}_{\rm MLE}$ is a critical point, i.e., $\nabla \ell_n(\widehat{\theta}_{\rm MLE})=0$. Recall from (13) that the EM iteration corresponds to gradient ascent of the log likelihood ℓ_n with step size one. Applying the Taylor expansion of $\nabla \ell_n$ at $\widehat{\theta}_{\rm MLE}$, from (13) we get

$$\begin{aligned} \theta_{t+1} - \widehat{\theta}_{\rm MLE} &= \theta_t - \widehat{\theta}_{\rm MLE} + \nabla \ell_n(\theta_t) \\ &= \big(I + \nabla^2 \ell_n(\xi_t)\big)(\theta_t - \widehat{\theta}_{\rm MLE}), \end{aligned} \tag{81}$$

where $\xi_t = \alpha\theta_t + (1-\alpha)\widehat{\theta}_{\rm MLE}$ for some $\alpha \in [0,1]$. The key lemma is then as follows:

Lemma 10. *Under the setting of Theorem 9, denote $\delta \triangleq c(d/n)^{1/4} \log n$ for some constant c depending only on r. With probability at least $1-2n^{-1}$, for all θ such that $\ell(\theta, \theta_*) \le \delta$, we have*

$$0 \preceq I + \nabla^2 \ell_n(\theta) \preceq e^{-c\|\theta_*\|^2} I.$$

We now apply Lemma 10 to show the convergence of θ_t to $\widehat{\theta}_{\rm MLE}$. To do so, we need some crude guarantee on the MLE. We provide such an analysis in Appendix C. In particular, by (165) therein, with probability at least $1-\exp(-cd\log^2 n)$,

$$\ell(\widehat{\theta}_{\rm MLE}, \theta_*) \le \left(C \frac{d \log n}{n}\right)^{1/4}$$

for some constants c, C. Since $\|\theta_*\| > 2\delta$ for all sufficiently large n, in the event that

$$\ell(\widehat{\theta}_{\rm MLE}, \theta_*) \le \delta \quad \text{and} \quad \ell(\theta_T, \theta_*) \le \delta,$$

θ_T and $\widehat{\theta}_{\rm MLE}$ must both belong to exactly one of the two balls $B(\theta_*, \delta)$ and $B(-\theta_*, \delta)$. Without loss of generality, we can assume the former. Taking norms on both sides of (81) and applying Lemma 10, we have

$$\|\theta_{T+1} - \widehat{\theta}_{\rm MLE}\| \le e^{-c\|\theta_*\|^2} \|\theta_T - \widehat{\theta}_{\rm MLE}\|,$$

and hence (80) follows, which, in particular, implies the convergence of $\{\theta_t\}$ and the uniqueness of $\widehat{\theta}_{\rm MLE}$.

7. Discussions and open problems

We conclude this paper by discussing some technical aspects of the results and related or open problems:

Small initialization. In this paper, we showed that the EM algorithm achieves the near-optimal rate and converges to the MLE when the direction of the initialization θ_0 is uniform on the sphere and θ_0 is sufficiently close to zero, specifically, $\|\theta_0\| = \Theta((d/n \log n)^{1/4})$ (cf. Theorem 8). Computationally speaking, using a small initialization does not compromise the needed number of iterations as the signal grows rapidly according to (76) in the initial Phase I. Technically speaking, the main reason for using a small initialization in the proof is to ensure the orthogonal component β_t stays within the near-optimal rate throughout the entire trajectory, as shown in Theorem 5. An added bonus is that the signal component α_t converges monotonically; as demonstrated in Figure 3, this can fail for large initialization. We conjecture that the same result applies to $\|\theta_0\| = \Theta(1)$. Proving such a result entails a refined analysis of the initial phase since α_t initially decays due to β_t being as large as a constant (see Figure 3a).

Extensions. In this paper we considered the simple symmetric 2-GM model. It is of great interest to understand the performance or limitations of EM algorithms in more general Gaussian mixture models, e.g., multiple components, unknown covariance matrix, asymmetric and unknown weights, and, more generally, location-scale mixtures. The optimal and adaptive rates of location mixtures in one dimension were obtained in [16] and shown to be achieved by the generalized method of moments [40]. It remains open whether the corresponding EM algorithm achieves competitive performance. One immediate hurdle is the existence of bad fixed points, which can exist for population EM for 3-GM even in one dimension [19].

Beyond Gaussian mixture models, statistical problems with missing data, and other latent variable models such as mixture of regression and alignment problems in cryo-EM [31] are major avenues where EM algorithm are applied. Promising results have been obtained recently in [1, 22], although finite-sample finite-iteration guarantees and analysis for random initializations are still lacking.

The present paper concerns analyzing EM algorithm for the purpose of parameter estimation. For the related problem of *classification*, that is, recovering the labels of each sample with small error rate, we refer to the recent work on Lloyd's algorithm [24] and optimal rates [26]. It remains open to understand the performance of EM algorithm for clustering and whether it achieves the optimal rates.

8. Proofs in Section 2

8.1. Proofs of Theorem 3 and Corollary 1

Proof of Theorem 3. Step 1. We show that

$$\theta_t \leq \theta_t^+ \tag{82}$$

by induction on t. The base case of $t = 0$ is clearly true. Assume that (82) holds for t. Then,

$$\begin{aligned}\theta_{t+1} &= f(\theta_t) + \Delta_n(\theta_t)\\ &\le f(\theta_t) + \gamma_n \theta_t\\ &\le f(\theta_t^+) + \gamma_n \theta_t^+ = \theta_{t+1}^+,\end{aligned}$$

where we used the fact that $\theta \mapsto f(\theta) + \gamma_n \theta$ is increasing on $\mathbb{R}_+$.

Step 2. We show that $\theta_t^+ \le C_1$ for all t for some constant C_1. This simply follows from the fact that f is bounded. By Lemma 3 and the assumption $\theta_* \le r$,

$$\theta_{t+1}^+ = f(\theta_t^+) + \gamma_n \theta_t^+ \le 1 + r + \gamma_n \theta_t^+,$$

where $\gamma_n \le c_\gamma \le \frac{1}{2}$ in the event (26). Setting $C_1 = 2(1 + r)$ and letting $n \ge 4C_0^2$, the proof follows from induction on t.

Step 3. We show that

$$\theta_t \ge \theta_t^- \ge 0, \tag{83}$$

by induction on t. The base case of $t = 0$ is clearly true. Assume that (83) holds for t. Then,

$$\begin{aligned}\theta_{t+1} &\ge f(\theta_t) - \gamma_n \theta_t\\ &\ge f(\theta_t^-) - \gamma_n \theta_t^- = \theta_{t+1}^-,\end{aligned}$$

where we used the fact $C_1 \ge \theta_t \ge \theta_t^-$ as shown in the previous step and $\theta \mapsto f(\theta) - \gamma_n \theta$ is increasing on $[0, C_1]$. To see this, note that $f(\theta)$ is concave on $\mathbb{R}_+$. Therefore, $f'(\theta) \ge f'(C_1) \ge \gamma_n$, which holds in the event (26) provided that $c_\gamma \le f'(C_1)$. Finally, $\theta_{t+1}^- \ge 0$ follows again from monotonicity and $\theta_t^- \ge 0$. This completes the proof of (31).

Step 4. Next we prove the convergence of $\{\theta_t^+\}$ to θ^+. Recall $q(\theta) = f(\theta)/\theta$ from Lemma 3, which is a decreasing function on $\mathbb{R}_+$. By definition, we have

$$q(\theta^+) = 1 - \gamma_n. \tag{84}$$

Furthermore, we have, crucially,

$$f(\theta) + \gamma_n \theta \gtrless \theta \quad \text{if } \theta \lessgtr \theta^+.$$

Therefore,

$$|\theta_{t+1}^+ - \theta^+| < |\theta_t^+ - \theta^+|,$$

and hence $\theta_t^+ \to \theta^+$ as $t \to \infty$. Similarly, if $\theta_*^2 \geq \gamma_n$, then we have $\theta_t^- \to \theta^-$; if $\theta_*^2 < \gamma_n$, then $\theta^- = 0$ by definition and we have $\liminf \theta_t^- \geq \theta^-$.

Step 5. We show (32). Recall that $q(\theta) = f(\theta)/\theta$ from Lemma 3. If $\theta_*^2 \geq \gamma_n$, by the definitions (28)–(29), we have

$$\begin{aligned} q(\theta^+) &= 1 - \gamma_n, \\ q(\theta^-) &= 1 + \gamma_n, \\ q(\theta_*) &= 1. \end{aligned}$$

If $\theta_*^2 \leq \gamma_n$, then $\theta^- = 0$ by definition. In both cases, since q is decreasing on $\mathbb{R}_+$ by Lemma 3, we have

$$\theta^- \leq \theta_* \leq \theta^+.$$

Furthermore, since $\theta_* \in [0, r]$, by (25), for all $\theta \in [0, C_1]$,

$$q'(\theta) \leq -\frac{2\theta}{3}\mathbb{E}\left[\frac{Y^4}{\cosh^2(\theta Y)}\right] \leq -C_4\theta, \tag{85}$$

where C_4 is a constant that depends on r (recall $C_1 = 2(r+1)$).

Let $\varepsilon^+ = \theta^+ - \theta_*$. Then,

$$\begin{aligned} -\gamma_n &= q(\theta_* + \varepsilon^+) - q(\theta_*) = \int_{\theta_*}^{\theta_* + \varepsilon^+} q'(\tau)\, d\tau \\ &\overset{(85)}{\leq} -\frac{C_4}{2}\left((\theta_* + \varepsilon^+)^2 - \theta_*^2\right) = -\frac{C_4}{2}(2\theta_*\varepsilon^+ + \varepsilon^{+2}). \end{aligned}$$

Hence,

$$0 \leq \varepsilon^+ \leq \min\left\{\frac{\gamma_n}{C_4\theta_*}, \sqrt{\frac{2\gamma_n}{C_4}}\right\} \leq C_3 \min\left\{\frac{\gamma_n}{\theta_*}, \sqrt{\gamma_n}\right\}. \tag{86}$$

Similarly, let $\varepsilon^- = \theta_* - \theta^-$. Then, $0 \leq \varepsilon^- \leq \theta_*$. Furthermore, if $\theta_*^2 \geq \gamma_n$, we have

$$\begin{aligned} \gamma_n &= q(\theta_* - \varepsilon^-) - q(\theta_*) = \int_{\theta_* - \varepsilon^-}^{\theta_*} -q'(\tau)\, d\tau \\ &\overset{(85)}{\geq} \frac{C_4}{2}\left(\theta_*^2 - (\theta_* - \varepsilon^-)^2\right) = \frac{C_4}{2}(2\theta_* - \varepsilon^-)\varepsilon^- \geq \frac{C_4}{2}\theta_*\varepsilon^-. \end{aligned}$$

Hence,

$$0 \leq \varepsilon^- \leq \min\left\{\theta_*, \frac{2\gamma_n}{C_4\theta_*}\right\} \leq C_5 \min\left\{\frac{\gamma_n}{\theta_*}, \sqrt{\gamma_n}\right\}. \tag{87}$$

If $\theta_*^2 < \gamma_n$, since $\varepsilon^- \leq \theta_*$, then (87) holds automatically. Thus, combining (86) and (87) yields

$$\theta_* - \varepsilon \leq \theta^- \leq \liminf_{t\to\infty} \theta_t \leq \limsup_{t\to\infty} \theta_t \leq \theta^+ \leq \theta_* + \varepsilon, \tag{88}$$

where $\varepsilon \triangleq C_6 \min\{\gamma_n/\theta_*, \sqrt{\gamma_n}\}$.

Step 6. Finally, we provide a finite-iteration version of (88). In view of the sandwich inequality (31), it suffices to determine the convergence rate of $\{\theta_t^+\}$ and $\{\theta_t^-\}$. We consider two cases separately.

Case I: $\boldsymbol{\theta_*^2 \leq 2\gamma_n}$. Let $\varepsilon_t^+ = \theta_t^+ - \theta^+$. If $\varepsilon_t^+ \leq 0$, then we have

$$0 \leq \theta_t \leq \theta_t^+ \leq \theta^+ \leq \theta^* + \varepsilon \lesssim n^{-1/4},$$

which is already within the optimal rate of convergence. So it suffices to consider $\varepsilon_t^+ \geq 0$, i.e., θ_t^+ converging to θ^+ from above. Then,

$$\begin{aligned} \varepsilon_{t+1}^+ &= \theta_t^+\big(q(\theta_t^+) + \gamma_n\big) - \theta^+ \\ &\overset{(84)}{=} \varepsilon_t^+ + \theta_t^+\big[q(\theta_t^+) - q(\theta^+)\big] \\ &\overset{(85)}{\leq} \varepsilon_t^+ - C_6(\varepsilon_t^+ + \theta^+)\big(\theta^+\varepsilon_t^+ + (\varepsilon_t^+)^2\big) \\ &\leq \varepsilon_t^+ - C_6\big((\theta^+)^2\varepsilon_t^+ + (\varepsilon_t^+)^3\big) \\ &\leq \varepsilon_t^+ - C_6'(\varepsilon_t^+)^3, \end{aligned} \tag{89}$$

where $C_6' = \min\{C_6, 1/r_0^2\}$. Next we apply Lemma 22 with $h(x) = C_6'x^3$ to the sequence $\{\varepsilon_t^+\}$, which satisfies $h(x) < x$ for all $x \in (0, \varepsilon_0^+)$, since $\varepsilon_0^+ \leq \theta_0 \leq r_0$. We have

$$G(x) = \int_x^{r_0} \frac{1}{h(\tau)}\, d\tau = C_7\left(\frac{1}{x^2} - \frac{1}{r_0^2}\right),$$

and we conclude that

$$\varepsilon_t^+ \leq \frac{1}{\sqrt{t/C_7 + 1/r_0^2}} \leq \sqrt{\frac{C_7}{t}}.$$

Thus, for all $t \geq C_7/\gamma_n$, we have $\varepsilon_t^+ \leq \sqrt{\gamma_n}$ and hence $|\theta_t^+ - \theta_*| \lesssim \sqrt{\gamma_n}$.

Case II: $\boldsymbol{\theta_*^2 \geq 2\gamma_n}$. Let $\varepsilon_t^+ = \theta_t^+ - \theta^+$. First assume $\varepsilon_t^+ \geq 0$, in which case ε_t^+ converges to zero from above. Since $\theta_* \gtrsim \sqrt{\gamma_n}$, we have $\theta^- \asymp \theta^+ \asymp \theta_*$. Continuing from (89), we conclude that

$$\varepsilon_{t+1}^+ \leq (1 - C_8\theta_*^2)\varepsilon_t^+.$$

Therefore, for all sufficiently large n, as soon as $t \geq C_8' \log(n)/\theta_*^2$, we have

$$\theta_t - \theta_* \leq \varepsilon_t^+ \leq \frac{1}{\theta^*\sqrt{n}}.$$

Similarly, if $\varepsilon_t^+ \leq 0$, we have $\varepsilon_{t+1}^+ \geq \varepsilon_t^+(1 - C_8\theta_*^2)$, which converges to zero from below.

Next we analyze the convergence rate of $\{\theta_t^-\}$. Let $\varepsilon_t^- = \theta^- - \theta_t^-$. We only consider the case of $\varepsilon_t^- \geq 0$ as the other case is entirely analogous. Since $f(\theta) - \gamma_n\theta > \theta$ if and only if $\theta < \theta^-$, we have $\theta_t^- \to \theta^-$ from below and ε_t^- is a decreasing positive sequence. Let $c_0 = 1/(200\sqrt{3 + r^4})$. Consider two cases:

Case II.1: $\theta_t^- \geq c_0\theta_*$. Entirely analogous to (89), we have

$$\begin{aligned}\varepsilon_{t+1}^- &= \varepsilon_t^- - \theta_t^-\big[q(\theta_t^-) - q(\theta^-)\big]\\ &\leq \varepsilon_t^- - C_6(\theta^-)^2\varepsilon_t^-\\ &\leq \varepsilon_t^-(1 - C_9\theta_*^2).\end{aligned}$$

Since $\varepsilon_0^- = \theta^- - \theta_0 \leq \theta_* \leq r$, for all sufficiently large n, as soon as

$$t \geq C_9' \frac{\log(1/\gamma_n)}{\theta_*^2},$$

we have $\theta_t - \theta_* \geq -\varepsilon_t^- \geq -\gamma_n/\theta^*$.

Case II.2: $0 < \theta_t^- \leq c_0\theta_*$. Recall from Lemma 3 that $f(0) = f''(0) = 0$ and $f'(0) = 1 + \theta_*^2$. Furthermore, $f'''(\theta) = \mathbb{E}[Y^4 \tanh'''(\theta Y)]$. Since $|\tanh'''| \leq 2$, for all θ, we have

$$|f'''(\theta)| \leq 2\mathbb{E}[Y^4] \leq 16(3 + r^4). \tag{90}$$

Therefore, the Taylor expansion of f at zero yields

$$\theta_{t+1}^- = f(\theta_t^-) - \gamma_n\theta_t^- \geq \left(1 + \theta_*^2 - \gamma_n - \frac{16(3+r^4)}{6}c_0^2\theta_*^2\right)\theta_t^- \geq \left(1 + \frac{\theta_*^2}{4}\right)\theta_t^-,$$

where the last inequality is by the choice of c_0. Therefore, in at most $C_{11}/\theta_*^2 \log(\theta_*/\theta_0)$ iterations, we have $\theta_t^- \geq c_0\theta_*$ which enters the previous Case II.1.

In summary, for all $t \geq C_{12}/\theta_*^2 \log(\theta_*/\theta_0\gamma_n)$, we have $|\theta_t - \theta_*| \lesssim \gamma_n/\theta^*$. ■

Proof of Corollary 1. An inspection of the proof of Theorem 3 shows that the guarantees in (32) and (33) apply if $\gamma_n \equiv W_1(\nu_n, \nu)$ is replaced by any upper bound thereof, which we choose to be $\max\{\gamma_n, 1/\sqrt{n}\}$. Then, in the event E defined in (26), we have

$$\ell(\theta_t, \theta_*) \leq \tau_2 \min\left\{\frac{\max\{\gamma_n, 1/\sqrt{n}\}}{\theta_*}, \sqrt{\max\left\{\gamma_n, \frac{1}{\sqrt{n}}\right\}}\right\}$$

holds for all t satisfying (35). Taking expectation and using (22) and Jensen's inequality, we have

$$\mathbb{E}\big[\ell(\theta_t, \theta_*)\mathbf{1}_E\big] \leq \tau_2(1 + c_0)\min\left\{\frac{1}{\theta_*\sqrt{n}}, \frac{1}{n^{1/4}}\right\},$$

where the high-probability event E is in (26). Finally, by definition of the EM map, we have $|\theta_t| \leq \|f_n\|_\infty \leq \mathbb{E}_n|Y|$, and hence $|\ell(\theta_t, \theta_*)| \leq r + \mathbb{E}_n|Y|$. Therefore, by the Cauchy–Schwarz inequality, we have

$$\mathbb{E}\big[\ell(\theta_t, \theta_*)\mathbf{1}_{E^c}\big] \leq \sqrt{\mathbb{P}[E^c]}\sqrt{\mathbb{E}\big[\big(r + \mathbb{E}_n|Y|\big)^2\big]} \overset{(23)}{\leq} C\exp(-cn^{1/3})$$

for some constants c, C depending on r. Combining the previous two displays yields the desired (34). ■

8.2. Proof of Lemma 3

Proof. We consider each part in turn.

1. By definition,

$$f'(\theta) = \mathbb{E}\big[Y^2 \tanh'(\theta Y)\big] = \mathbb{E}\left[\frac{Y^2}{\cosh^2(\theta Y)}\right] \geq 0,$$
$$f''(\theta) = \mathbb{E}\big[Y^3 \tanh''(\theta Y)\big] = -2\mathbb{E}\left[\frac{Y^3\tanh(\theta Y)}{\cosh^2(\theta Y)}\right].$$

2. Clearly $f''(\theta)$ is negative (resp., positive) when θ is positive (resp., negative).
3. $f(0) = f''(0) = 0$ by definition, $f'(0) = \mathbb{E}[Y^2]$ and

$$\begin{aligned} f'(\theta_*) &= \mathbb{E}\left[\frac{Y^2}{\cosh^2(\theta_* Y)}\right] \\ &= \mathbb{E}\left[\frac{Z^2}{\cosh(\theta_* Z)}\right]\exp(-{\theta_*}^2/2), \qquad\qquad Z \sim N(0,1), \\ &\leq \mathbb{E}[Z^2]\exp(-{\theta_*}^2/2) = \exp(-{\theta_*}^2/2), \end{aligned}$$

where the second equality follows from a change of measure from Y to Z; compare with Lemma 26.

4. The monotonicity of q simply follows from the concavity of f on $\mathbb{R}_+$ and $f(0) = 0$. By the symmetry of the distribution of Y, we have

$$q'(\theta) = -\mathbb{E}\left[\frac{Y\sinh(2\theta Y) - 2\theta Y^2}{2\theta^2\cosh^2(\theta Y)}\bigg| Y \geq 0\right] \leq -\frac{2\theta}{3}\mathbb{E}\left[\frac{Y^4}{\cosh^2(\theta Y)}\bigg| Y \geq 0\right],$$

where we used the fact that $\sinh(x) \geq x + x^3/6$ for $x \geq 0$; (b) follows from $\cosh \geq 1$ and Jensen's inequality. ■

9. Proofs in Section 3

Proof of Theorem 4. First of all, by definition, we have

$$\|f_n(\theta)\| = \|\mathbb{E}_n\big[Y\tanh\langle\theta, Y\rangle\big]\| \leq \mathbb{E}_n\big[\|Y\|\big] \leq \sqrt{\mathbb{E}_n\big[\|Y\|^2\big]}.$$

Define the event

$$E_2 = \big\{\mathbb{E}_n\big[\|Y\|^2\big] \leq 2\|\theta_*\|^2 + 10d\big\}.$$

Since $\mathbb{E}_n[\|Y\|^2] \leq 2\|\theta_*\|^2 + 2\mathbb{E}_n[\|Z\|^2]$, where $n\mathbb{E}_n[\|Z\|^2] \sim \chi^2_{nd}$, and by the χ^2 tail bound (154) in Appendix A, we have

$$\mathbb{P}[E_2] \geq 1 - \exp(-nd).$$

Next, we show that with probability at least $1 - \exp(-c_0 d \log n)$,

$$\|\Delta_n(\theta)\| \leq C_0 \|\theta\|(1+r)\sqrt{\frac{d}{n}\log n}$$

for all $\theta \in B(R)$.

Let $Y, Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} P_{\theta_*}$. Let $\mathcal{C} \subset S^{d-1}$ be an ε-covering of S^{d-1} in Euclidean distance, where $\varepsilon \leq 1/2$ is to be specified later. It is well known (see [36]) that $\mathcal{C}$ can be chosen so that

$$|\mathcal{C}| \leq \Big(1 + \frac{2}{\varepsilon}\Big)^d \leq \Big(\frac{3}{\varepsilon}\Big)^d.$$

Furthermore, for any $y \in \mathbb{R}^d$, we have

$$\|y\| \leq \frac{1}{1-\varepsilon} \max_{u \in \mathcal{C}} \langle u, y\rangle,$$

and hence

$$\|\Delta_n(\theta)\| \leq 2 \max_{u \in \mathcal{C}} \mathbb{E}\big[\langle u, Y\rangle \tanh\langle \theta, Y\rangle\big] - \mathbb{E}_n\big[\langle u, Y\rangle \tanh\langle \theta, Y\rangle\big].$$

For each $\theta \in \mathbb{R}$, there exists $v \in \mathcal{C}$ such that $\|\|\theta\|v - \theta\| \leq \varepsilon\|\theta\|$. For any $u \in \mathcal{C}$, using Cauchy–Schwarz and the fact that tanh is 1-Lipschitz, we have

$$\begin{aligned}
&\big|\mathbb{E}\big[\langle u, Y\rangle \tanh\big(\langle \theta, Y\rangle\big)\big] - \mathbb{E}\big[\langle u, Y\rangle \tanh\big(\|\theta\|\langle v, Y\rangle\big)\big]\big| \\
&\quad \leq \mathbb{E}\big[|\langle u, Y\rangle||\langle \theta - \|\theta\|v, Y\rangle|\big] \leq \mathbb{E}\big[\|Y\|^2\big]\|u\|\|\theta - \|\theta\|v\| \leq \varepsilon\|\theta\|\mathbb{E}\big[\|Y\|^2\big].
\end{aligned}$$

Similarly,

$$\big|\mathbb{E}_n\big[\langle u, Y\rangle \tanh\big(\langle \theta, Y\rangle\big)\big] - \mathbb{E}_n\big[\langle u, Y\rangle \tanh\big(\|\theta\|\langle v, Y\rangle\big)\big]\big| \leq \varepsilon\mathbb{E}_n\big[\|Y\|^2\big]\|\theta\|.$$

Therefore,

$$\begin{aligned}
\|\Delta_n(\theta)\| \leq 2 \max_{u,v \in \mathcal{C}} &\big|\mathbb{E}\big[\langle u, Y\rangle \tanh\big(\|\theta\|\langle v, Y\rangle\big)\big] - \mathbb{E}_n\big[\langle u, Y\rangle \tanh\big(\|\theta\|\langle v, Y\rangle\big)\big]\big| \\
&+ \varepsilon\|\theta\|\big(\mathbb{E}\big[\|Y\|^2\big] + \mathbb{E}_n\big[\|Y\|^2\big]\big),
\end{aligned}$$

and hence

$$\begin{aligned}
&\sup_{0<\|\theta\|\leq R} \frac{\|\Delta_n(\theta)\|}{\|\theta\|} \\
&\quad \leq 2 \max_{u,v \in \mathcal{C}} \sup_{0<a\leq R} \underbrace{\frac{1}{a}\big|\mathbb{E}\big[\langle u, Y\rangle \tanh\big(a\langle v, Y\rangle\big)\big] - \mathbb{E}_n\big[\langle u, Y\rangle \tanh\big(a\langle v, Y\rangle\big)\big]\big|}_{\triangleq F(u,v,a)} \\
&\qquad + \varepsilon\big(\mathbb{E}\big[\|Y\|^2\big] + \mathbb{E}_n\big[\|Y\|^2\big]\big),
\end{aligned}$$

where $\mathbb{E}[\|Y\|^2] = d + \|\theta_*\|^2 \leq d + r^2$.

We consider two cases separately:

Case I: $0 < a \le \varepsilon$. Since $|\tanh'| \le 1$ and $|\tanh''| \le 1$ everywhere, we have

$$\left|\frac{1}{a}\mathbb{E}\big[\langle u, Y\rangle \tanh\big(a\langle v, Y\rangle\big)\big] - \mathbb{E}\big[\langle u, Y\rangle\langle v, Y\rangle\big]\right| \le \varepsilon\mathbb{E}\big[|\langle u, Y\rangle|\langle v, Y\rangle^2\big] \le \varepsilon\mathbb{E}\big[\|Y\|^3\big],$$

and similarly,

$$\left|\frac{1}{a}\mathbb{E}_n\big[\langle u, Y\rangle \tanh\big(a\langle v, Y\rangle\big)\big] - \mathbb{E}_n\big[\langle u, Y\rangle\langle v, Y\rangle\big]\right| \le \varepsilon\mathbb{E}_n\big[\|Y\|^3\big].$$

Therefore,

$$\sup_{0<a\le\varepsilon} F(u, v, a) \le \big|\mathbb{E}\big[\langle u, Y\rangle\langle v, Y\rangle\big] - \mathbb{E}_n\big[\langle u, Y\rangle\langle v, Y\rangle)\big]\big| \\ + \varepsilon\big(\mathbb{E}\big[\|Y\|^3\big] + \mathbb{E}_n\big[\|Y\|^3\big]\big).$$

For any $u, v \in \mathcal{C}$, note that

$$\langle u, Y\rangle\langle v, Y\rangle = \langle u, \theta_*\rangle\langle v, \theta_*\rangle + \langle XZ, \langle u, \theta_*\rangle v + \langle v, \theta_*\rangle u\rangle + \langle u, Z\rangle\langle v, Z\rangle.$$

Since $\|\theta_*\| \le r$ by assumption and $\|\langle u, Z\rangle\langle v, Z\rangle\|_{\psi_1} \le \|\langle u, Z\rangle\|_{\psi_2}\|\langle v, Z\rangle\|_{\psi_2} = 1$ (cf. [36, Lemma 2.7.7]), we conclude that $\langle u, Y\rangle\langle v, Y\rangle$ is $C_2(r+1)$-subexponential. By Bernstein's inequality (cf. [36, Theorem 2.8.1]), for any b such that $bd\log n \le n$,

$$\mathbb{P}\left[\big|\mathbb{E}\big[\langle u, Y\rangle\langle v, Y\rangle\big] - \mathbb{E}_n\big[\langle u, Y\rangle\langle v, Y\rangle\big]\big| \ge (1+r)\sqrt{\frac{bd\log n}{n}}\right] \\ \le \exp(-cbd\log n),$$

where c is some absolute constant. Furthermore,

$$\mathbb{E}\big[\|Y\|^3\big] \le C_4\big(r + \sqrt{d}\big)^3 \quad \text{and} \quad \mathbb{E}_n\big[\|Y\|^3\big] \le \max_{i\in[n]} \|Y_i\|^3.$$

Since $n \ge d\log d$, $\mathbb{P}[\|Y_i\| \ge \sqrt{n}\,] \le \exp(-cn)$. Therefore, by the union bound,

$$\mathbb{E}_n\big[\|Y\|^3\big] \le n^{3/2}$$

with probability at least $1 - \exp(-c'n)$.

Case II: $\varepsilon \le a \le R$. Let $\mathcal{R}$ be an ε^2-net for the interval $[\varepsilon, R]$, so that for any $a \in [\varepsilon, R]$, there exists $a' \in \mathcal{R}$ such that $|a - a'| \le \varepsilon^2$. Then,

$$\left|\frac{1}{a}\mathbb{E}\big[\langle u, Y\rangle \tanh\big(a\langle v, Y\rangle\big)\big] - \frac{1}{a'}\mathbb{E}\big[\langle u, Y\rangle \tanh\big(a'\langle v, Y\rangle\big)\big]\right| \\ \le 2\frac{|a - a'|}{a}\mathbb{E}\big[|\langle u, Y\rangle\langle v, Y\rangle|\big] \le 2\varepsilon\mathbb{E}\big[\|Y\|^2\big].$$

Therefore,

$$\sup_{\varepsilon \le a \le R} F(u, v, a) \le \max_{a \in \mathcal{R}} F(u, v, a) + 2\varepsilon\big(\mathbb{E}\big[\|Y\|^2\big] + \mathbb{E}_n\big[\|Y\|^2\big]\big).$$

For any $u, v \in \mathcal{C}$ and $a \in \mathcal{R}$,

$$\left|\frac{\langle u, Y\rangle \tanh\big(a\langle v, Y\rangle\big)}{a}\right| \le |\langle u, Y\rangle||\langle v, Y\rangle|.$$

Therefore, $|(\langle u, Y\rangle \tanh(a\langle v, Y\rangle))/a|$ is $C_2(1+r)$-subexponential. Again by Bernstein's inequality, we have

$$\mathbb{P}\left[|F(u, v, a)| \ge (1+r)\sqrt{\frac{bd \log n}{n}}\right] \le \exp(-cbd \log n).$$

Set $\varepsilon = n^{-4}$ so that $|\mathcal{C}| \le (3n^4)^d$ and $|\mathcal{R}| \le Rn^4$. Applying the union bound to both cases and choosing a sufficiently large constant b completes the proof. ■

10. Proofs in Section 4

10.1. Proofs of Theorems 5–7

Throughout this section denote for brevity $s \triangleq \|\theta_*\|$.

Proof of Theorem 5. We first show that the sequence $\{\alpha_t, \beta_t\}$ is bounded. By assumption, $\omega \le \frac{1}{2}$ and $\|\theta_0\| \le 1$ by (51). Using the bounded property of the F and G maps in Lemma 5 and induction on t, we have

$$|\alpha_t| \le \Gamma, \quad 0 \le \beta_t \le \Gamma, \tag{91}$$

where $\Gamma = 2(\|\theta_*\| + \sqrt{2/\pi}) \le 2r + 2$.

Combining (44) and (49), we have

$$\beta_{t+1} \le \beta_t\left(1 - \frac{\alpha_t^2 + \beta_t^2}{2 + 4(\alpha_t^2 + \beta_t^2)}\right) + \omega\big(|\alpha_t| + \beta_t\big)$$

from which (52) follows. To show (53), note that, in view of (91), we have

$$\begin{aligned}\beta_{t+1} &\le \beta_t\left(1 - \frac{\alpha_t^2 + \beta_t^2}{2 + 8\Gamma^2}\right) + \omega\big(|\alpha_t| + \beta_t\big) \\ &\le \beta_t(1+\omega) - \frac{\beta_t^3}{2 + 8\Gamma^2} + \sup_{0 \le \alpha \le \Gamma}\left(\omega\alpha - \frac{\alpha^2\beta_t}{2 + 8\Gamma^2}\right) \\ &\le \beta_t(1+\omega) - \frac{\beta_t^3}{2 + 8\Gamma^2} + \min\left\{\frac{\omega^2(2 + 8\Gamma^2)}{4\beta_t}, \omega\Gamma\right\}.\end{aligned} \tag{92}$$

Let $C_1 = 2 + 8\Gamma^2$. Let β be any limiting point of the sequence $\{\beta_t\}$. Taking limits on both sides we have

$$\frac{\beta^3}{C_1} \le \omega\beta + \frac{\omega^2 C_1}{4\beta} \le 2\left(\omega\beta \vee \frac{\omega^2 C_1}{4\beta}\right),$$

which implies that either $\beta \le \sqrt{2C_1\omega}$ or $\beta \le (\omega^2 C_1^2/2)^{1/4}$. So we conclude (54).

Finally, we prove (56). We show by induction that there exists some constant a depending only on r, such that $\beta_t \le a\sqrt{\omega}$ for all $t \ge 0$. The base case is the assumption (55). Next, fix some constant b to be specified and consider two cases:

Case I: $\boldsymbol{\beta_t \le b\omega}$. From (53), we get

$$\beta_{t+1} \le \beta_t(1+\omega) - \frac{\beta_t^3}{C_1} + \omega\Gamma \le \omega(b + \omega + \Gamma) \le a\sqrt{\omega},$$

provided that $\sqrt{\omega} \le a/(b + \omega + \Gamma)$.

Case II: $\boldsymbol{b\omega \le \beta_t \le a\sqrt{\omega}}$. Again from (53), we get $\beta_{t+1} \le h(\beta_t)$, where

$$h(\beta) \triangleq \beta(1+\omega) - \frac{\beta^3}{C_1} + \frac{\omega^2 C_1}{2\beta}.$$

Note that

$$\frac{d}{d\beta}h(\beta) = 1 + \omega - \frac{3\beta^2}{C_1} - \frac{\omega^2 C_1}{2\beta^2} \ge 1 - \frac{C_1}{3b^2} + \omega\left(1 - \frac{3a^2}{C_1}\right) \ge 0,$$

provided that $C_1/3b^2 \le \frac{1}{2}$ and $\omega(1 - (3a^2/C_1)) \ge -\frac{1}{2}$. Therefore,

$$\beta_{t+1} \le \sup_{b\omega \le \beta \le a\sqrt{\omega}} h(\beta) \le h(a\sqrt{\omega}) = a\sqrt{\omega} + \omega^{3/2}\left(a - \frac{a^3}{C_1} + \frac{C_1}{2a}\right) \le a\sqrt{\omega},$$

provided that $a^3/C_1 \ge 2a$ and $a^3/C_1 \ge C_1/a$. Finally, choosing $a = 2C_1$ and $b = C_1$, then the above conditions hold simultaneously as long as $\omega \le c_0 = c_0(r)$ for some small constant c_0. ■

Proof of Theorem 6. It suffices to show (57) which, together with (54), implies (58). Combining (47) with (41) and (48) with (42), we have

$$\alpha_{t+1} \le f(\alpha_t) + \Gamma|\alpha_t|\beta_t^2 + \omega\big(|\alpha_t| + \beta_t\big), \tag{93}$$

$$\alpha_{t+1} \ge f(\alpha_t) - \Gamma|\alpha_t|\beta_t^2 - \omega\big(|\alpha_t| + \beta_t\big) \tag{94}$$

with $\Gamma = 1 + s^2$. Since $\|\theta_*\| = s \le s_0 \le 1$, we have $\Gamma \le 2$. Furthermore, in this case the constant κ_2 in (56) is also absolute. Let α be any limiting point of $\{\alpha_t\}$. We show that $|\alpha| \leq 2s_0$. Assume for the sake of contradiction that $\alpha \geq 2s_0$. Sending $t \to \infty$ in (93) and in view of (56), we have

$$\alpha \leq f(\alpha) + C_3(\alpha\omega + \omega^{3/2}), \tag{95}$$

for some absolute constant C_3. Let $q(\alpha) = f(\alpha)/\alpha$ be defined in (24) with θ_* replaced by s. As shown in Lemma 3, q is a decreasing function on $\mathbb{R}_+$ with $q(s) = 1$. Dividing both sides of (95) by α leads to

$$1 \leq q(\alpha) + C_3\left(\omega + \frac{\omega^{3/2}}{\alpha}\right) \leq q(2s_0) + \frac{3C_3}{2}\omega,$$

where the last inequality holds because of the assumption $s_0 \geq \sqrt{\omega}$. Furthermore, for all $\alpha \in [0, 2]$, we have $q'(\alpha) \leq -C_4\alpha$ for some absolute constant C_4. Thus,

$$q(2s_0) - 1 = \int_s^{2s_0} q'(\alpha)d\alpha \leq -C_4(4s_0^2 - s^2) \leq -3C_4s^2.$$

Therefore, we reach the desired contradiction that

$$q(2s_0) + \frac{3C_3}{2}\omega \leq 1 - C_4s^2 + \frac{3C_3}{2}\omega < 1,$$

provided that $s^2 \geq (3C_3/2C_4)\omega$. The proof is completed by taking

$$K = \max\left\{1, \sqrt{\frac{3C_3}{2C_4}}\right\}.$$

For the other direction, if $\alpha < -2s_0$, then the above proof applies to (94) with α replaced by $-\alpha$ and in view of the fact that $f(-\alpha) = -f(\alpha)$. This completes the proof of (57).

Finally, we show the second part for small initialization satisfying (55). We prove (59) by induction on t. The base case of $t = 0$ follows from

$$\alpha_0 \leq \|\theta_0\| \leq \kappa_2\left(\frac{d \log n}{n}\right)^{1/4} \leq LK\sqrt{\omega} \leq Ls_0,$$

provided that $L \geq \kappa_2/KC_\omega^{1/4}$, where both κ_2 and C_ω in (46) are absolute constants since $\|\theta_*\| \leq 1$ by assumption. Next, using (93) and the argument that leads to (95), we have

$$\alpha_{t+1} \leq f(\alpha_t) + C_3(\alpha_t\omega + \omega^{3/2}).$$

By the monotonicity of f, it suffices to show that

$$f(Ls_0) + C_3(Ls_0\omega + \omega^{3/2}) \leq Ls_0.$$

To this end, recalling from (85) and the fact that $q(0) = f'(0) = 1 + s^2 \leq 1 + s_0^2$, we have $q(\alpha) \leq 1 + s_0^2 - C_4\alpha^2/2$, where C_4 is absolute since $\|\theta_*\| \leq 1$. Thus,

$$f(\alpha) = \alpha q(\alpha) \leq \alpha(1 + s_0^2) - C_4\alpha^3/2.$$

Therefore, using the assumption that $s_0 \geq K\sqrt{\omega}$, we have

$$\begin{aligned} f(Ls_0) + C_3(Ls_0\omega + \omega^{3/2}) &= Ls_0 + s_0\big(L - C_4L^3/2 + C_3(L/K^2 + 1/K^3)\big) \\ &\leq Ls_0, \end{aligned}$$

provided that L exceeds some large absolute constant. This completes the proof of (59), which implies (60) in view of (56) provided that $L \geq \kappa_2$. ■

Proof of Theorem 7. By assumption, $s \geq C_0\sqrt{\omega}$. Without loss of generality, we can assume that $\alpha_0 \geq 0$ (otherwise we the same argument applies with α_t replaced by $-\alpha_t$ and s by $-s$). By design, α_t is close to zero at $t = 0$. The argument entails proving that initially α_t increases geometrically approximately as $\alpha_{t+1} = (1 + \Omega(s^2))\alpha_t$, until α_t exceeds $\Omega(\sqrt{\omega})$. After this point, the sandwich bound (93)–(94) behave as the linear perturbation of the one-dimensional EM iteration in (27), and consequently the one-dimensional analysis in Theorem 3 applies, yielding both error bound and speed of convergence.

By the assumption (62),

$$\|\theta_0\| = c\Big(\frac{d}{n}\log n\Big)^{1/4} \leq c'\sqrt{\omega}$$

for some small constant c' proportional to c. Since $c \leq \kappa_2$, (56) in Theorem 5 ensures that $\beta_t \leq \kappa_2(d/n \log n)^{1/4}$ for *all* $t \geq 0$. Then, (93)–(94) imply the following

$$\alpha_{t+1} \leq f(\alpha_t) + C_5(\alpha_t\omega + \omega^{3/2}), \tag{96}$$

$$\alpha_{t+1} \geq f(\alpha_t) - C_5(\alpha_t\omega + \omega^{3/2}). \tag{97}$$

Let C_* be a constant to be specified. Consider the following phases:

Phase I: $\alpha_t \leq C_*\sqrt{\omega}$. We will show that throughout Phase I, for some sufficiently large constant C_4,

$$\alpha_t \geq \frac{C_4}{s^2}\omega^{3/2}. \tag{98}$$

In view of the choice of the initialization (62), the assumption (62) ensures that (98) holds for the base case of $t = 0$, where C_4 is proportional to λ_2/c and can be made sufficiently large. Assume (98) holds at time t. By Lemma 3 and using (90), the Taylor expansion of f at 0 gives $f(\alpha_t) \geq (1 + s^2)\alpha_t - C_6\alpha_t^3$. So (97) implies

$$\alpha_{t+1} \geq (1 + s^2)\alpha_t - C_6\alpha_t^3 - C_5(\alpha_t\omega + \omega^{3/2}) \geq (1 + s^2/4)\alpha_t$$

where, since $s \geq C_0\sqrt{\omega}$ and $(C_4/s^2)\omega^{3/2} \leq \alpha_t \leq C_*\sqrt{\omega}$ by assumption, the last inequality holds provided that

$$C_0 \geq \frac{C_*}{\sqrt{4C_6}}, \quad C_0 \geq \sqrt{4C_5}, \quad C_4 \geq 4C_5. \tag{99}$$

Therefore, (98) holds at time $t+1$. Furthermore, α_t grows exponentially and in $T_1 = O(1/s^2 \log(s/\omega)) = O(\log(n)/s^2)$ iterations enters the next phase.

Phase II: $\boldsymbol{\alpha_t \geq C_*\sqrt{\omega}}$. Then, (96)–(97) imply

$$\alpha_{t+1} \leq f(\alpha_t) + C_5'\omega\alpha_t, \tag{100}$$

$$\alpha_{t+1} \geq f(\alpha_t) - C_t'\omega\alpha_t, \tag{101}$$

where $C_5' = C_5(1 + (1/C_*))$. Comparing (100)–(101) with (27), by replacing γ_n with ω, θ_* with $s = \|\theta_*\|$, and the initial value θ_0 by $\alpha_{T_1} \geq C_*\sqrt{\omega}$, we see that Theorem 3 applies to the convergence of $\{\alpha_t : t \geq T_1\}$. In particular, (32) and (33) yield

$$|\alpha_t - s| \leq C_7 \min\left\{\frac{\omega}{s}, \sqrt{\omega}\right\}, \tag{102}$$

for all $t - T_1 \geq T_2 \triangleq C_8/s^2 \log(ns/\sqrt{\omega}) = O(1/s^2 \log n)$. This completes the proof of (63).

Phase III: improved estimate on $\boldsymbol{\beta_t}$. Since $s \geq C_0\sqrt{\omega}$ by assumption, from (102), we conclude that for all $t \geq T_1 + T_2$, we have $\alpha_t \in [s/2, 2s]$. Recall that the prior unconditional analysis in Theorem 5 treats α_t as zero (which is the worst case) and shows that $\beta_t = O(\sqrt{\omega})$. Now that $\alpha_t = \Theta(s)$, we will use the α-dependent bound (44) to upgrade the error bound to $\beta_t = O(\omega/s)$. Continuing from (92), for all $t \geq T_1 + T_2$, we have

$$\begin{aligned}
\beta_{t+1} &\leq \beta_t\left(1 - \frac{\alpha_t^2 + \beta_t^2}{2 + 8\Gamma^2}\right) + \omega\big(|\alpha_t| + \beta_t\big) \\
&\overset{\text{(a)}}{\leq} \beta_t\left(1 - \frac{s^2}{4(2 + 8\Gamma^2)}\right) + \omega(2s + \beta_t) \\
&\overset{\text{(b)}}{\leq} \beta_t(1 - C_9 s^2) + 2\omega s,
\end{aligned}$$

where (a) follows from $s/2 \leq \alpha_t \leq 2s$ and (b) follows from the assumption $s \geq C_0\sqrt{\omega}$ for sufficiently large C_0, where C_9 is a constant depending only on Γ (hence on r). Thus, $\beta_t \leq 4\omega/s$ for all $t - (T_1 + T_2) \geq T_3 \triangleq C_{10}/s^2 \log(s/\omega) = O(1/s^2 \log n)$. This completes the proof of (64). ■

10.2. Proof of Lemma 5

Proof. Let $s = \|\theta_*\|$. Let $W = \langle \xi, Z\rangle$ and $U = \langle \eta, Z\rangle$, which are independent standard normals. Then,

$$\langle \theta, Y\rangle = \alpha\|\theta_*\|X + \alpha U + \beta W = \alpha V + \beta W,$$

where $V \sim \frac{1}{2}N(-s,1) + \frac{1}{2}N(s,1)$ is independent of W.

1. The function $\alpha \mapsto \mathbb{E}[V \tanh(\alpha V + \beta W)]$ is because of the symmetry of the distribution of W. Furthermore,

$$\frac{\partial F}{\partial \alpha} = \mathbb{E}\left[\frac{V^2}{\cosh^2(\alpha V + \beta W)}\right] \geq 0,$$
$$\frac{\partial^2 F}{\partial \alpha^2} = \mathbb{E}\big[V^3 \tanh''(\alpha V + \beta W)\big] = \mathbb{E}\big[Z^3 \tanh''(\alpha Z + \beta Z)\cosh(sZ)\big]e^{-s^2/2},$$

where the last equality follows from a change of measure (Lemma 26) with $Z \sim N(0,1)$ independent of W. Consider $\alpha \geq 0$. By symmetry,

$$\mathbb{E}\big[Z^3 \cosh(sZ) \mid \alpha Z + \beta W = y\big]$$

is an odd function which is nonnegative if and only if $y \geq 0$.

Since $\tanh'' = -2\tanh \operatorname{sech}^2$, we have

$$\mathbb{E}\big[Z^3 \cosh(sZ) \mid \alpha Z + \beta W\big]\tanh''(\alpha Z + \beta W) \leq 0$$

almost surely. Therefore, $\alpha \mapsto F(\alpha, \beta)$ is concave on $\mathbb{R}_+$, and convex on $\mathbb{R}_-$ by symmetry.

2. This is simply because $F(\cdot, \beta)$ is an odd function and increasing on $\mathbb{R}_+$.

3. Entirely analogously,

$$\frac{\partial G}{\partial \beta} = \mathbb{E}\left[\frac{W^2}{\cosh^2(\alpha V + \beta W)}\right] \geq 0,$$
$$\frac{\partial^2 G}{\partial \beta^2} = -2\mathbb{E}\left[\frac{W^3 \tanh(\alpha V + \beta W)}{\cosh^2(\alpha V + \beta W)}\right] \leq 0.$$

4. For $\alpha \geq 0$,

$$\begin{aligned}\frac{\partial F}{\partial \beta} = \frac{\partial G}{\partial \alpha} &= \mathbb{E}\big[WV \tanh'(\alpha V + \beta W)\big] \\ &= \beta\mathbb{E}\big[V \tanh''(\alpha V + \beta W)\big] \qquad (103) \\ &= -2\beta\mathbb{E}\left[\frac{V \tanh(\alpha V + \beta W)}{\cosh^2(\alpha V + \beta W)}\right]\end{aligned}$$

$$= -2\beta\mathbb{E}\Bigg[\underbrace{\frac{\mathbb{E}\big[V \mid \alpha V + \beta W\big]\tanh(\alpha V + \beta W)}{\cosh^2(\alpha V + \beta W)}}_{\geq 0}\Bigg] \leq 0, \tag{104}$$

where (103) follows from Stein's lemma, and (104) follows from the fact that, in view of Lemma 23 and the symmetry of the distribution of V,

$$\widehat{V}(y) \triangleq \mathbb{E}\big[V \mid \alpha V + \beta W = y\big]$$

is an odd and increasing function such that $\widehat{V}(y) \gtrless 0$ when $y \gtrless 0$.

The case for $\alpha \leq 0$ follows from the fact that

$$G(-\alpha, \beta) = G(\alpha, \beta) \quad \text{and} \quad F(-\alpha, \beta) = -F(\alpha, \beta).$$

5. We have

$$|F(\alpha, \beta)| = \big|\mathbb{E}\big[V \tanh(\alpha V + \beta W)\big]\big| \leq \mathbb{E}\big[|V|\big] \leq \|\theta_*\| + \mathbb{E}|U|,$$

and similarly, $|G(\alpha, \beta)| \leq \mathbb{E}[|W|]$.

6. By the third property, $\alpha \mapsto G(\alpha, \beta)$ is maximized at $\alpha = 0$.

7. We only prove (41) for $\alpha \geq 0$; (42) follows from the fact that $F(-\alpha, \beta) = -F(\alpha, \beta)$. The left inequality follows from (104). To show the right inequality, note that since $\mathbb{E}[V \mid \alpha V + \beta W]\tanh(\alpha V + \beta W) \geq 0$ almost surely, using the fact that $\cosh(x) \geq 1$ and $\tanh(x) \lesseqgtr x$ for $x \gtrless 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{E}[V \mid \alpha V + \beta W]\tanh(\alpha V + \beta W)}{\cosh^2(\alpha V + \beta W)}\right] &\leq \mathbb{E}\big[\mathbb{E}[V \mid \alpha V + \beta W](\alpha V + \beta W)\big] \\
&= \mathbb{E}\big[V(\alpha V + \beta W)\big] \\
&= \alpha\mathbb{E}[V^2] = \alpha\big(1 + \|\theta_*\|^2\big).
\end{aligned}$$

Consequently,

$$\frac{\partial F}{\partial \beta} \geq -2\beta\alpha\big(1 + \|\theta_*\|^2\big).$$

Integrating over β yields the right inequality in (41).

8. By symmetry, without loss of generality we assume $\alpha \geq 0$. By Stein's identity,

$$G(\alpha, \beta) = \mathbb{E}\big[W \tanh(\alpha V + \beta W)\big] = \beta\mathbb{E}\big[\tanh'(\alpha V + \beta W)\big].$$

Recall that $V = sX + U$, where X is Rademacher and $U \sim N(0, 1)$. Let

$$T = \alpha(sX + U) + \beta W = \alpha s X + (\alpha U + \beta W).$$

Then,

$$\frac{G(\alpha, \beta)}{\beta} = \mathbb{E}\left[\frac{1}{\cosh^2(T)}\right].$$

Since $\mathbb{E}[X \mid T = t] = \tanh((\alpha s/(\alpha^2 + \beta^2))t)$, we have

$$\frac{\partial}{\partial s}\left(\frac{G(\alpha,\beta)}{\beta}\right) = \alpha\mathbb{E}\big[X\tanh''(T)\big] = -2\alpha\mathbb{E}\left[\frac{X\tanh(T)}{\cosh^2(T)}\right]$$
$$= -2\alpha\mathbb{E}\underbrace{\left[\frac{\tanh((\alpha s/\alpha^2+\beta^2)T)\tanh(\alpha T)}{\cosh^2(T)}\right]}_{\geq 0} \leq 0.$$

Therefore, $G(\alpha,\beta)/\beta$ is decreasing in s, and it suffices to consider $s = 0$. Next we show for any $\sigma \geq 0$ and $Z \sim N(0,1)$,

$$\mathbb{E}\left[\frac{1}{\cosh^2(\sigma Z)}\right] \leq 1 - \frac{\sigma^2}{2(1+2\sigma^2)}, \tag{105}$$

which applied to $\sigma^2 = \alpha^2 + \beta^2$ implies the desired result.

Using the inequality $\cosh(x) \geq 1 + x^2/2$, and hence $\cosh^2(x) \geq 1 + x^2$, we have[5]

$$\mathbb{E}\left[\frac{1}{\cosh^2(\sigma Z)}\right] \leq \mathbb{E}\left[\frac{1}{1+\sigma^2 Z^2}\right] = \frac{\bar{\Phi}(1/\sigma)}{\sigma\varphi(1/\sigma)}. \tag{106}$$

Using Lemma 24, we have

$$\mathbb{E}\left[\frac{1}{\cosh^2(\sigma Z)}\right] \leq 1 - \frac{2\sigma^2}{(\sqrt{1+2\sigma^2}+1)^2} \leq 1 - \frac{\sigma^2}{2(1+2\sigma^2)}.$$

This proves (105) and the desired (44). ■

11. Proofs in Section 5

11.1. Proof of Lemmas 7, 8 and 9

We start by defining a few typical events which will be used subsequently for several times.

[5] The last inequality in (106) is due to the following integral representation of Mill's ratio [15, 3.466.1]:

$$\mathbb{E}\Big[\frac{1}{t^2+Z^2}\Big] = \frac{\bar{\Phi}(t)}{t\varphi(t)}.$$

To see this, let $f(t) = \mathbb{E}[t/(t^2+Z^2)]$. By Stein's identity, one can verify that f satisfies the differential equation $f'(t) = tf(t) - 1$. Thus, $g(t) = f(t)\varphi(t)$ satisfies $g'(t) = -\varphi(t)$, which implies that $g(t) = \bar{\Phi}(t)$ since $g(\infty) = 0$.

Lemma 11. *Define*

$$H_2 = \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_{i,1}^2 \geq 1 + \|\theta_*\|^2 - \sqrt{\frac{\kappa \log n}{n}} \right\}, \tag{107}$$

$$H_4 = \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_{i,1}^4 \leq \kappa \right\}, \tag{108}$$

$$H_3 = \left\{ \sum_{i=1}^{n} \|Y_i\|^3 \leq \kappa d^{3/2} \right\}, \tag{109}$$

$$H_\infty = \left\{ \max_{i \in [n]} |Y_{i,1}| \leq \sqrt{\kappa \log n} \right\}. \tag{110}$$

Then, there exists some $\kappa = \kappa(\|\theta_*\|)$ such that $\mathbb{P}[H_i] \geq 1 - n^{-1}$ for $i = 2, 3, 4, \infty$.

Next we provide the supporting lemmas:

Lemma 12 (Smoothness of the sample-EM map). *Let f_n be defined in (9). Then, f_n is $\|\Sigma_n\|_{\mathrm{op}}$-Lipschitz continuous on $\mathbb{R}^d$, where $\Sigma_n \triangleq \mathbb{E}_n[YY^\top]$ is the sample covariance matrix. In particular, with probability at least $1 - e^{-C'd \log n}$,*

$$\|\Sigma_n\|_{\mathrm{op}} \leq 1 + \|\theta_*\|^2 + \sqrt{\frac{Cd}{n}},$$

where the constants C, C' depend only on r.

Lemma 13. *Assume that $n \geq d$. Let $Y_\perp = [Y_{1,\perp}, \dots, Y_{1,\perp}]$. Then,*

$$\mathbb{P}\left[\|Y_\perp\|_{\mathrm{op}} \geq 4\sqrt{n} \right] \leq e^{-n}. \tag{111}$$

Furthermore, there exists some constant C depending only on r, such that with probability at least $1 - 4n^{-1}$,

$$\frac{1}{n} \sum_{i=1}^{n} Y_{i,1}^2 |\langle Y_{i,\perp}, \theta \rangle|^2 \leq C \|\theta\|^2,$$

for all $\theta \in \mathbb{R}^{d-1}$.

Lemma 14. *Let $b = (b_1, \dots, b_n)$ consist of independent Rademacher random variables and let $x = (x_1, \dots, x_n)$ be independent of b. Then, for any $a, s > 0$,*

$$\mathbb{P}\left[\frac{1}{n} \left| \sum_{i=1}^{n} x_i b_i \right| \geq \sqrt{\frac{as}{n}} \right] \leq 2 \exp(-s/8) + \mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} x_i^2 \geq a \right]. \tag{112}$$

Furthermore, given a finite collection $\{x^\theta : \theta \in \Theta\}$ independent of b,

$$\mathbb{P}\left[\sup_{\theta \in \Theta} \frac{1}{n} \left| \sum_{i=1}^{n} x_i^\theta b_i \right| \geq \sqrt{\frac{as}{n}} \right] \leq 2 \exp(-s/8) |\Theta| + \mathbb{P}\left[\frac{1}{n} \sup_{\theta \in \Theta} \sum_{i=1}^{n} (x_i^\theta)^2 \geq a \right]. \tag{113}$$

Lemma 15. *Assume that $n \geq Cd$ for some absolute constant C. Let $q\colon \mathbb{R} \to \mathbb{R}$ be a function with bounded first two derivatives, such that*

$$\max\left\{\|q'\|_\infty, \|q''\|_\infty\right\} \leq L_0, \tag{114}$$

for some constant L_0. Define a (random) function $D\colon \mathbb{R}^d \to \mathbb{R}$ by

$$D(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} Y_{i,1} b_i q\big(\langle \theta, Y_i\rangle\big),$$

where $\{b_i\}$ are independent Rademacher variables and independent of $\{Y_i\}$. Let $R > 0$. Then, there exists a constant L_1 depending only on L_0, r and R, such that with probability at least $1 - 10n^{-1}$, D is $\sqrt{L_1 d \log(n)/n}$-Lipschitz on the ball

$$B(R) = \{\theta \in \mathbb{R}^d : \|\theta\| \leq R\}.$$

Lemma 16. *For $\theta = (\theta_1, \theta_\perp) \in \mathbb{R}^d$, define*

$$M(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} b_i Y_{i,\perp} Y_{i,1} Q\big(\theta_1 Y_{i,1}, \langle \theta_\perp, Y_{i,\perp}\rangle\big),$$

where $Q\colon \mathbb{R}^2 \to \mathbb{R}$ satisfies $\max\{\|Q\|_\infty, \|\partial_x Q\|_\infty, \|\partial_y Q\|_\infty\} \leq L_0$ for some constants L_0. Let $R > 0$. Then, there exist constants L_1 depending only on L_0, r and R, such that with probability at least $1 - 10n^{-1}$,

$$\sup_{\|\theta\| \leq R} \|M(\theta)\| \leq \sqrt{\frac{L_1 d \log n}{n}}.$$

We now prove the main lemmas:

Proof of Lemma 7. By the definition in (69), we have $T_\star = O(\sqrt{n}\log n)$. By the union bound, with probability at least $1 - O(T_\star n^{-1}) = 1 - O(n^{-1/2}\log n)$, (77) and (78) hold for all $t \leq T_\star$. On this event, we proceed by induction on t.

For the base case of $t = 0$, (73) is trivially true, and (74)–(75) hold by virtue of the random initialization in the event (71).

Next, assume that (73) and (74) hold at time t. Recall from (46) that

$$\omega \asymp \sqrt{\frac{d \log n}{n}}.$$

In particular, thanks to the assumption (65) and (68), we have

$$T_\star \sqrt{\frac{d \log^2 n}{n}} \lesssim \sqrt{\frac{d \log^2 n (\log d + \log\log n)^2}{n\|\theta_*\|^4}} \lesssim 1. \tag{115}$$

Thus, (73) implies that

$$\|\theta_t - \widetilde{\theta}_t\| \le C_3 \alpha_t. \tag{116}$$

By (115), (74) implies that

$$\alpha_t \ge \frac{1}{\sqrt{d \log n} + C_2} \beta_t,$$

which further implies the desired (75), since $\|\theta_t\|^2 = \alpha_t^2 + \beta_t^2$.

To show that (74) holds at time $t+1$, by (77) in Lemma 8, we have

$$\begin{aligned}\alpha_{t+1} &\ge \alpha_t \left(1 + \|\theta_*\|^2 - \sqrt{\frac{C \log n}{n}} - C\|\theta_t\|^2 \right) \\ &\quad - \sqrt{\frac{C \log n}{n}} \|\theta_t\| - \sqrt{\frac{Cd \log n}{n}} \|\theta_t - \widetilde{\theta}_t\| \\ &\ge \alpha_t \left(1 + \|\theta_*\|^2 - C_4 \sqrt{\frac{d \log^2 n}{n}} \right),\end{aligned} \tag{117}$$

where the last step follows from (70), (75), and (116). This proves (76). Combined with (52), we have

$$\begin{aligned}\frac{\beta_{t+1}}{\alpha_{t+1}} &\le \frac{\beta_t}{\alpha_t} \frac{1+\omega}{1 + \|\theta_*\|^2 - C_4\sqrt{d \log^3(n)/n}} + \frac{\omega}{1 + \|\theta_*\|^2 - C_4\sqrt{d \log^3(n)/n}} \\ &\le \frac{\beta_t}{\alpha_t} + \omega,\end{aligned}$$

where the last step follows from the assumption (65) with the constant $C_\star$ chosen to be sufficiently large. Thus, the ratio β_t/α_t grows at most linearly and satisfies

$$\frac{\beta_t}{\alpha_t} \le \frac{\beta_0}{\alpha_0} + \omega t \le \sqrt{d \log n} + \omega t,$$

in the event (71). This is the desired (74).

It remains to show (73) holds at time $t+1$. To this end, write abstractly

$$\|\theta_t - \widetilde{\theta}_t\| \le \alpha_t K_t. \tag{118}$$

We will show that

$$K_t \le C_5 \left\{ \left(1 + \sqrt{\frac{C_5 d \log^2 n}{n}} \right)^t - 1 \right\}, \tag{119}$$

which, in view of (115), implies the desired

$$K_t \le C_5' \sqrt{\frac{d \log^2 n}{n}}\, t$$

for all $t \le T_\star$.

Next we apply the induction hypothesis to (78) in Lemma 9:

$$\begin{aligned}
&\|\widetilde{\theta}_{t+1}-\theta_{t+1}\| \\
&\quad\le \left(1+\|\theta_*\|^2+\sqrt{\frac{Cd\log n}{n}}\right)\|\widetilde{\theta}_t-\theta_t\|+\sqrt{\frac{Cd\log n}{n}}\alpha_t+\sqrt{\frac{C\log n}{n}}\|\theta_t\| \\
&\quad\overset{(a)}{\le} \alpha_t\left\{K_t\left(1+\|\theta_*\|^2+\sqrt{\frac{Cd\log n}{n}}\right)+\sqrt{\frac{Cd\log n}{n}}\right\}+\sqrt{\frac{C\log n}{n}}\|\theta_t\| \\
&\quad\overset{(b)}{\le} \alpha_t\left\{K_t\left(1+\|\theta_*\|^2+\sqrt{\frac{Cd\log n}{n}}\right)+\sqrt{\frac{C_6 d\log^2 n}{n}}\right\} \\
&\quad\overset{(c)}{\le} \alpha_{t+1}\frac{K_t\left(1+\|\theta_*\|^2+\sqrt{Cd\log(n)/n}\right)+\sqrt{C_6 d\log^2(n)/n}}{1+\|\theta_*\|^2-C_4\sqrt{d\log^2(n)/n}},
\end{aligned}$$

where (a) follows from (118); (b) follows from (75); and (c) follows from (117). This means that K_t satisfies

$$K_{t+1}\le\frac{K_t\left(1+\|\theta_*\|^2+\sqrt{Cd\log(n)/n}\right)+\sqrt{C_6 d\log^2(n)/n}}{1+\|\theta_*\|^2-C_4\sqrt{d\log^2(n)/n}}$$

Since $K_0=0$, in view of the assumption (65), applying Lemma 21 shows that K_t satisfies (119). Thus, we obtain the desired (73) at time $t+1$. ∎

Proof of Lemma 8. First of all, in view of (91) and (45), with probability at least $1-2\exp(-2c_0 d\log n)$, both the main and the auxiliary sequences are bounded, i.e.,

$$\sup_{t\ge0}\|\theta_t\|\le 4(r+1),\qquad \sup_{t\ge0}\|\widetilde{\theta}_t\|\le 4(r+1). \tag{120}$$

Write

$$f_n(\theta_t)=\mathbb{E}_n\big[YY^\top\big]\theta_t+\mathbb{E}_n\big[Y\big(\tanh\langle\theta_t,Y\rangle-\langle\theta_t,Y\rangle\big)\big].$$

Then,

$$\alpha_{t+1}=\underbrace{\mathbb{E}_n\big[Y_1\langle Y,\theta_t\rangle\big]}_{R_1}-\underbrace{\mathbb{E}_n\big[Y_1\big(\langle Y,\theta_t\rangle-\tanh\langle\theta_t,Y\rangle\big)\big]}_{R_2}.$$

We first show that with probability at least $1-O(n^{-1})$,

$$R_1\ge\left(1+\|\theta_*\|^2-\sqrt{\frac{C\log n}{n}}\right)\alpha_t-\sqrt{\frac{C\log n}{n}}\|\theta_{t,\perp}\|-\sqrt{\frac{Cd\log n}{n}}\|\widetilde{\theta}_t-\theta_t\| \tag{121}$$

and

$$|R_2|\le C\alpha_t\|\theta_t\|^2+\sqrt{\frac{C\log n}{n}}\|\theta_{t,\perp}\|+\sqrt{\frac{Cd\log n}{n}}\|\theta_t-\widetilde{\theta}_t\|. \tag{122}$$

Then, the desired result (77) follows from (121) and (122).

For the linear term R_1, we have

$$\begin{aligned} R_1 &= \mathbb{E}_n\big[Y_1\langle Y, \theta_t\rangle\big] \\ &= \frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\big(\alpha_t b_i Y_{i1} + \langle Y_{i\perp}, \theta_{t,\perp}\rangle\big) \\ &= \left(\frac{1}{n}\sum_{i=1}^n Y_{i1}^2\right)\alpha_t + \frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\langle Y_{i\perp}, \theta_{t,\perp}\rangle. \end{aligned} \tag{123}$$

Here, the first term (signal) satisfies

$$\frac{1}{n}\sum_{i=1}^n Y_{i1}^2 \geq 1 + \|\theta_*\|^2 - O\left(\sqrt{\frac{\log n}{n}}\right),$$

in view of (107). For the second term, we cannot afford to take union bound over the d-dimensional sphere. Instead, we resort to the auxiliary iterates $\{\widetilde{\theta}_t\}$. Write

$$\frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\langle Y_{i\perp}, \theta_{t,\perp}\rangle = \frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\langle Y_{i\perp}, \widetilde{\theta}_{t,\perp}\rangle + \frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\langle Y_{i\perp}, \theta_{t,\perp} - \widetilde{\theta}_{t,\perp}\rangle. \tag{124}$$

Using the independence between $(\widetilde{\theta}_t, \{Y_{i,1}\})$ and $\{b_i\}$, for some constants C, C', we have

$$\begin{aligned} &\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\langle Y_{i\perp}, \widetilde{\theta}_{t,\perp}\rangle\right| \geq \sqrt{\frac{C\log n}{n}}\,\|\widetilde{\theta}_{t,\perp}\|\right] \\ &\overset{\text{(a)}}{\leq} 2n^{-1} + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n Y_{i1}^2\langle Y_{i\perp}, \widetilde{\theta}_{t,\perp}\rangle^2 \geq C'\|\widetilde{\theta}_{t,\perp}\|^2\right] \\ &\overset{\text{(b)}}{\leq} 6n^{-1}, \end{aligned} \tag{125}$$

where (a) follows from Lemma 14 and (b) follows from Lemma 13. Furthermore, in the event (120), applying Lemma 15 to q being the identity function, we conclude that, with probability at least $1 - O(n^{-1})$,

$$\left|\frac{1}{n}\sum_{i=1}^n b_i Y_{i1}\langle Y_{i\perp}, \theta_{t,\perp} - \widetilde{\theta}_{t,\perp}\rangle\right| \leq \sqrt{\frac{Cd\log n}{n}}\,\|\theta_t - \widetilde{\theta}_t\|. \tag{126}$$

Combining (123)–(126) and using the triangle inequality yields (121).

For the nonlinear term R_2, define

$$g(x) \triangleq x - \tanh(x),$$

$$T(x,y) \triangleq \frac{1}{2}\big(g(y+x)+g(y-x)\big),$$
$$H(x,y) \triangleq \frac{1}{2}\big(g(y+x)-g(y-x)\big).$$

Then, for any x, y and any $b \in \{\pm 1\}$, we have

$$g(y+bx) = T(x,y) + bH(x,y). \tag{127}$$

Furthermore, we have the following lemma:

Lemma 17. *For any $x, y \in \mathbb{R}$,*

$$0 \le y \cdot T(x,y) \le x^2y^2 + y^4 \tag{128}$$

and

$$|H(x,y)| \le |x|. \tag{129}$$

Then,

$$\begin{aligned}
R_2 &= \mathbb{E}_n\big[Y_1 g(\langle Y, \theta_t\rangle)\big] \\
&= \frac{1}{n}\sum_{i=1}^n b_i Y_{i,1} g\big(\langle Y_i, \theta_t\rangle\big) = \frac{1}{n}\sum_{i=1}^n b_i Y_{i,1} g\big(b_i\alpha_t Y_{i,1} + \langle Y_{i,\perp}, \theta_{t,\perp}\rangle\big) \\
&\overset{\text{(a)}}{=} \frac{1}{n}\sum_{i=1}^n Y_{i,1} g\big(\alpha_t Y_{i,1} + b_i\langle Y_{i,\perp}, \theta_{t,\perp}\rangle\big) \\
&\overset{\text{(b)}}{=} \underbrace{\frac{1}{n}\sum_{i=1}^n T\big(\langle Y_{i,\perp}, \theta_{t,\perp}\rangle, \alpha_t Y_{i,1}\big)Y_{i,1}}_{R_3} + \underbrace{\frac{1}{n}\sum_{i=1}^n H\big(\langle Y_{i,\perp}, \theta_{t,\perp}\rangle, \alpha_t Y_{i,1}\big)Y_{i,1}b_i}_{R_4},
\end{aligned}$$

where (a) is due to $g(\pm x) = \pm g(x)$ and (b) follows from (127). Next we show (122) by proving that, with probability at least $1 - O(n^{-1})$,

$$|R_3| \le C\alpha_t\|\theta_t\|^2, \tag{130}$$

$$|R_4| \le \sqrt{\frac{C\log n}{n}}\|\theta_{t,\perp}\| + \sqrt{\frac{Cd\log n}{n}}\|\theta_t - \widetilde{\theta}_t\|. \tag{131}$$

To prove (130), let us recall that $\alpha_t > 0$ by assumption. Then, with probability at least $1 - O(n^{-1})$,

$$0 \overset{\text{(a)}}{\le} R_3 = \frac{1}{n}\sum_{i=1}^n T\big(\langle Y_{i,\perp}, \theta_{t,\perp}\rangle, \alpha_t Y_{i,1}\big)Y_{i,1}$$

$$\overset{\text{(b)}}{\leq} \alpha_t\left(\frac{1}{n}\sum_{i=1}^{n} Y_{i,1}^2\langle Y_{i,\perp},\theta_{t,\perp}\rangle^2\right) + \alpha_t^3\left(\frac{1}{n}\sum_{i=1}^{n} Y_{i,1}^4\right)$$
$$\overset{\text{(c)}}{\leq} C\alpha_t\|\theta_{t,\perp}\|^2 + C\alpha_t^3$$
$$\overset{\text{(d)}}{=} C\alpha_t\|\theta_t\|^2,$$

where (a) and (b) follow from (128) in Lemma 17; (c) follows from Lemma 13 and (108); and (d) is due to $\|\theta_{t,\perp}\|^2 + |\theta_{t,1}|^2 = \|\theta_t\|^2$. This completes the proof of (130).

To show (131), we again use the auxiliary iterates $\{\widetilde{\theta}_t\}$. For any $\theta = (\theta_1, \theta_\perp) \in \mathbb{R}^d$, define

$$\xi(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} H\big(\langle Y_{i,\perp},\theta_\perp\rangle, \theta_1 Y_{i,1}\big) Y_{i,1} b_i.$$

Then,

$$R_4 = \xi(\theta_t) = \xi(\widetilde{\theta}_t) + \xi(\theta_t) - \xi(\widetilde{\theta}_t), \tag{132}$$

Define

$$\theta_t' \triangleq (-\theta_{t,1}, \theta_{t,\perp}), \quad \widetilde{\theta}_t' \triangleq (-\widetilde{\theta}_{t,1}, \widetilde{\theta}_{t,\perp}), \tag{133}$$

which satisfies $\|\theta_t - \widetilde{\theta}_t\| = \|\theta_t' - \widetilde{\theta}_t'\|$. Then,

$$\xi(\theta_t) - \xi(\widetilde{\theta}_t) = \frac{1}{2n}\sum_{i=1}^{n} Y_{i,1} b_i \big\{g\big(\langle\widetilde{\theta}_t, Y_i\rangle\big) - g\big(\langle\theta_t, Y_i\rangle\big)\big\}$$
$$- \frac{1}{2n}\sum_{i=1}^{n} Y_{i,1} b_i \big\{g\big(\langle\widetilde{\theta}_t', Y_i\rangle\big) - g\big(\langle\theta_t', Y_i\rangle\big)\big\}.$$

In the event (120), applying Lemma 15 to $q = g$ whose first two derivatives are bounded by absolute constants, we conclude that, with probability at least $1-O(n^{-1})$,

$$|\xi(\theta_t) - \xi(\widetilde{\theta}_t)| \leq \sqrt{\frac{Cd\log n}{n}}\big(\|\theta_t - \widetilde{\theta}_t\| + \|\theta_t' - \widetilde{\theta}_t'\|\big)$$
$$= 2\sqrt{\frac{Cd\log n}{n}}\|\theta_t - \widetilde{\theta}_t\|. \tag{134}$$

To bound $\xi(\widetilde{\theta}_t)$, let $\widetilde{x}_i \triangleq H(\langle Y_{i,\perp}, \widetilde{\theta}_{t,\perp}\rangle, \widetilde{\alpha}_t Y_{i,1}) Y_{i,1}$, which are independent of $\{b_i\}$. Then,

$$\mathbb{P}\left[|\xi(\widetilde{\theta}_t)| \geq \sqrt{\frac{Cs}{n}}\|\widetilde{\theta}_{t,\perp}\|\right] = \mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\widetilde{x}_i b_i\right| \geq \sqrt{\frac{Cs}{n}}\|\widetilde{\theta}_{t,\perp}\|\right]$$
$$\overset{\text{(a)}}{\leq} 2\exp(-s/8) + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2 \geq C\|\widetilde{\theta}_{t,\perp}\|^2\right]$$
$$\overset{\text{(b)}}{\leq} 2\exp(-s/8) + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} Y_{i,1}^2\langle Y_{i,\perp}, \widetilde{\theta}_{t,\perp}\rangle^2 \geq C\|\widetilde{\theta}_{t,\perp}\|^2\right]$$

$$\overset{\text{(c)}}{\leq} 2\exp(-s/8) + n^{-3},$$

where (a) follows from Lemma 14; (b) is due to (129) in Lemma 17; and (c) is due to Lemma 13. Setting $s = 8\log n$ yields, with probability at least $1 - O(n^{-1})$,

$$|\xi(\widetilde{\theta}_t)| \leq \sqrt{\frac{C\log n}{n}}\|\widetilde{\theta}_{t,\perp}\| \leq \sqrt{\frac{C\log n}{n}}\big(\|\theta_{t,\perp}\| + \|\theta_t - \widetilde{\theta}_t\|\big). \tag{135}$$

Combining (132) with (134) and (135) completes the proof of (131), and hence the lemma. ■

Proof of Lemma 9. Write

$$\widetilde{\theta}_{t+1} - \theta_{t+1} = \underbrace{f_n(\widetilde{\theta}_t) - f_n(\theta_t)}_{\triangleq \mathcal{E}_1} + \underbrace{\widetilde{f}_n(\widetilde{\theta}_t) - f_n(\widetilde{\theta}_t)}_{\triangleq \mathcal{E}_2}.$$

For the first term, applying Lemma 12 yields that, with probability at least $1 - \exp(-C'd\log n)$,

$$\|\mathcal{E}_1\| = \|f_n(\widetilde{\theta}_t) - f_n(\theta_t)\| \leq \left(1 + \|\theta_*\|^2 + \sqrt{\frac{Cd}{n}}\right)\|\widetilde{\theta}_t - \theta_t\|. \tag{136}$$

Next we proceed to the second term. A trivial yet useful lemma is the following:

Lemma 18. *Assume that $b_i, \widetilde{b}_i \in \{\pm 1\}$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} h(y_i + \widetilde{b}_i x_i) - h(y_i + b_i x_i) = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{b}_i - b_i)B(x_i, y_i),$$

where $B(x, y) \triangleq (h(y + x) - h(y - x))/2$.

Proof. This simply follows from the fact that whenever $b = \pm 1$, we can write

$$h(x + by) = s + b\delta,$$

where

$$s \triangleq \frac{h(x+y) + h(x-y)}{2} \quad \text{and} \quad \delta = \frac{h(x+y) - h(x-y)}{2}.$$ ■

Next we bound $\mathcal{E}_2 = (\mathcal{E}_{2,1}, \mathcal{E}_{2,\perp})$. To bound the orthogonal component $\mathcal{E}_{2,\perp}$, note that $\widetilde{Y}_{i,\perp} = Y_{i,\perp}$. To apply Lemma 18 with $h = \tanh$, we define

$$\begin{aligned} B(x, y) &\triangleq \frac{\tanh(y + x) - \tanh(y - x)}{2}, \\ Q(x, y) &\triangleq \frac{B(x, y)}{x}, \end{aligned} \tag{137}$$

with $Q(0, y)$ understood as $\lim_{x\to 0} Q(x, y) = \operatorname{sech}^2(y)$. The function Q satisfies the following smoothness property:

Lemma 19. *Then, for all $x, y \in \mathbb{R}$, we have*

$$|Q(x, y)| \le 1, \quad |\partial_x Q(x, y)| \le 1/3, \quad |\partial_y Q(x, y)| \le 1.$$

In view of (127), we have

$$\begin{aligned}
\mathcal{E}_{2,\perp} &= \frac{1}{n}\sum_{i=1}^{n} Y_{i,\perp} \tanh\langle\widetilde{\theta}_t, \widetilde{Y}_i\rangle - \frac{1}{n}\sum_{i=1}^{n} Y_{i,\perp} \tanh\langle\widetilde{\theta}_t, Y_i\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_{i,\perp}\big(\tanh\big(\langle\widetilde{\theta}_{t,\perp}, Y_{i,\perp}\rangle + \widetilde{b}_i\widetilde{\theta}_{t,1}Y_{i,1}\big) - \tanh\big(\langle\widetilde{\theta}_{t,\perp}, Y_{i,\perp}\rangle + b_i\widetilde{\theta}_{t,1}Y_{i,1}\big)\big) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\widetilde{b}_i - b_i)Y_{i,\perp}B\big(\widetilde{\theta}_{t,1}Y_{i,1}, \langle\widetilde{\theta}_{t,\perp}, Y_{i,\perp}\rangle\big) \\
&= \widetilde{\theta}_{t,1}\left\{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{b}_i - b_i)Y_{i,1}Y_{i,\perp}Q\big(\widetilde{\theta}_{t,1}Y_{i,1}, \langle\widetilde{\theta}_{t,\perp}, Y_{i,\perp}\rangle\big)\right\},
\end{aligned}$$

where the penultimate step follows from applying Lemma 18 to $h = \tanh$. To apply Lemma 16, first note that the function Q defined in (137) fulfills the bounded derivative condition thanks to Lemma 19. Thus with probability at least $1 - O(n^{-1})$, it holds that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{b}_i - b_i)Y_{i,\perp}Y_{i,1}Q\big(\widetilde{\theta}_{t,1}Y_{i,1}, \langle\widetilde{\theta}_{t,\perp}, Y_{i,\perp}\rangle\big)\right\| \le \sqrt{\frac{Cd\log n}{n}},$$

and hence

$$\|\mathcal{E}_{2,\perp}\| \le |\widetilde{\theta}_{t,1}|\sqrt{\frac{Cd\log n}{n}} \le \big(\alpha_t + \|\widetilde{\theta}_t - \theta_t\|\big)\sqrt{\frac{Cd\log n}{n}}. \tag{138}$$

To bound the first coordinate of $\mathcal{E}_2$, let $\widetilde{x}_i = \widetilde{\theta}_{t,1}Y_{i,1}$, $\widetilde{y}_i = \langle\widetilde{\theta}_{t,\perp}, Y_{i,\perp}\rangle$ and similarly, $x_i = \theta_{t,1}Y_{i,1}$, $y_i = \langle\theta_{t,\perp}, Y_{i,\perp}\rangle$. Then,

$$\begin{aligned}
\mathcal{E}_{2,1} &= \frac{1}{n}\sum_{i=1}^{n} b_iY_{i,1}\tanh(\widetilde{y}_i + b_i\widetilde{x}_i) - \widetilde{b}_iY_{i,1}\tanh(\widetilde{y}_i + \widetilde{b}_i\widetilde{x}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_{i,1}\{\tanh(\widetilde{x}_i + b_i\widetilde{y}_i) - \tanh(\widetilde{x}_i + \widetilde{b}_i\widetilde{y}_i)\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\widetilde{b}_i - b_i)Y_{i,1}B(\widetilde{y}_i, \widetilde{x}_i)
\end{aligned}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\widetilde{b}_i Y_{i,1} B(y_i, x_i)}_{\mathcal{E}_3} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} b_i Y_{i,1} B(\widetilde{y}_i, \widetilde{x}_i)}_{\mathcal{E}_4}$$

$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\widetilde{b}_i Y_{i,1}\left\{B(\widetilde{y}_i, \widetilde{x}_i) - B(y_i, x_i)\right\}}_{\mathcal{E}_5} .$$

The first two terms can be dealt with using the same technology: For $\mathcal{E}_3$, we have

$$\begin{aligned}
\mathbb{P}\left[|\mathcal{E}_3| \geq 4\|\theta_{t,\perp}\|\sqrt{\frac{s}{n}}\right] &\overset{\text{(a)}}{\leq} 2\exp(-s/8) + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} B(y_i, x_i)^2 \geq 16\|\theta_{t,\perp}\|^2\right] \\
&\overset{\text{(b)}}{=} 2\exp(-s/8) + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\langle\theta_{t,\perp}, Y_{i,\perp}\rangle^2 \geq 16\|\theta_{t,\perp}\|^2\right] \\
&\overset{\text{(c)}}{\leq} 2\exp(-s/8) + \exp(-n),
\end{aligned}$$

where (a) follows from Lemma 14; (b) follows from the fact that

$$|B(y, x)| = \frac{|\tanh(x + y) - \tanh(x - y)|}{2} \leq |y|,$$

since tanh is 1-Lipschitz; and (c) follows from (111) in Lemma 14. Choosing $s = 8\log n$ yields

$$|\mathcal{E}_3| \leq \|\theta_t\|\sqrt{\frac{C\log n}{n}} \tag{139}$$

with probability at least $1 - O(n^{-1})$.

Entirely analogously, we have

$$\mathbb{P}\left[|\mathcal{E}_4| \geq 4\|\widetilde{\theta}_{t,\perp}\|\sqrt{\frac{s}{n}}\right] \leq 2\exp(-s/8) + \exp(-n).$$

Choosing $s = 8\log n$ yields

$$|\mathcal{E}_4| \leq \left(\|\theta_t\| + \|\widetilde{\theta}_t - \theta_t\|\right)\sqrt{\frac{C\log n}{n}} \tag{140}$$

with probability at least $1 - O(n^{-1})$.

To bound $\mathcal{E}_5$, recall from (133) the notations

$$\theta_t' = (-\theta_{t,1}, \theta_{t,\perp}) \quad \text{and} \quad \widetilde{\theta}_t' = (-\widetilde{\theta}_{t,1}, \widetilde{\theta}_{t,\perp}),$$

which satisfy $\|\theta_t' - \widetilde{\theta}_t'\| = \|\theta_t - \widetilde{\theta}_t\|$. Then, we have

$$\begin{aligned}\mathcal{E}_5 = \frac{1}{2n}\sum_{i=1}^{n}\widetilde{b}_i Y_{i,1}\big(\tanh\langle\theta_t, Y_i\rangle - \tanh\langle\widetilde{\theta}_t, Y_i\rangle\big)& \\ + \frac{1}{2n}\sum_{i=1}^{n}\widetilde{b}_i Y_{i,1}\big(\tanh\langle\theta_t', Y_i\rangle - \tanh\langle\widetilde{\theta}_t', Y_i\rangle\big).& \qquad (141)\end{aligned}$$

By Lemma 15 (applied to $q = \tanh$), the first term satisfies, with probability at least $1 - O(n^{-1})$,

$$\left|\frac{1}{2n}\sum_{i=1}^{n}\widetilde{b}_i Y_{i,1}\big(\tanh\langle\theta_t, Y_i\rangle - \tanh\langle\widetilde{\theta}_t, Y_i\rangle\big)\right| \leq \sqrt{\frac{C_1 d\log n}{n}}\,\|\theta_t - \widetilde{\theta}_t\|. \qquad (142)$$

Entirely analogously, the second term (and hence $|\mathcal{E}_5|$ itself) in (141) satisfies the same bound since

$$\|\theta_t' - \widetilde{\theta}_t'\| = \|\theta_t - \widetilde{\theta}_t\|.$$

Finally, since

$$\|\widetilde{\theta}_{t+1} - \theta_{t+1}\| \leq \|\mathcal{E}_1\| + \|\mathcal{E}_{2,\perp}\| + |\mathcal{E}_3| + |\mathcal{E}_4| + |\mathcal{E}_5|,$$

the desired (78) follows from combining (121), (136), (138)–(142). ■

11.2. Proof of supporting lemmas

Proof of Lemma 11. Note that $\frac{1}{n}\sum_{i=1}^{n} Y_{i,1}^2$ is equal in distribution to

$$1 + \|\theta_*\|^2 + \frac{\chi_n^2}{n} - 1 + N\left(0, \frac{4\|\theta_*\|^2}{n}\right).$$

Then, (107) follows from the χ^2-distribution tail bound (155) and the Gaussian tail bound. Next, since $Y_{i,1} \overset{\text{i.i.d.}}{\sim} \frac{1}{2}N(-\|\theta_*\|, 1) + \frac{1}{2}N(\|\theta_*\|, 1)$ have finite moments, (108) follows from the Chebyshev inequality. Also, since $\|Y_i\| \leq \|Z_i\| + \|\theta_*\|$, where $\|Z_i\| \sim \chi_d$, (109) follows similarly from the Chebyshev inequality. Finally, (110) follows simply from the union bound. ■

Proof of Lemma 12. The Jacobian of f_n is the following:

$$J_n(\theta) \triangleq \mathbb{E}_n\big[YY^\top \operatorname{sech}^2(\langle\theta, Y\rangle)\big], \qquad (143)$$

which is a (random) PSD matrix. Since $0 \leq \operatorname{sech} \leq 1$, for any u, we have

$$\begin{aligned}0 \leq u^\top J_n(\theta)u &= \mathbb{E}_n\big[\langle u, Y\rangle^2 \operatorname{sech}^2(\langle\theta, Y\rangle)\big] \\ &\leq \mathbb{E}_n\big[\langle u, Y\rangle^2\big] \leq u^\top J_n(0)u = u^\top \Sigma_n u.\end{aligned}$$

Thus, $J_n(\theta) \preceq \Sigma_n$ for any θ. For $\tau \in [0,1]$, define $a_\tau \triangleq (1-\tau)a_0 + \tau a_1$. Then,

$$\begin{aligned} f_n(a_1) - f_n(a_0) &= \frac{1}{n}\sum_{i=1}^n Y_i \int_0^1 d\tau \, \operatorname{sech}^2\big(\langle a_\tau, Y_i\rangle\big)\langle Y_i, a_1 - a_0\rangle \\ &= \left\{\int_0^1 d\tau J_n(a_\tau)\right\}(a_1 - a_0). \end{aligned}$$

Therefore,

$$\begin{aligned} \|f_n(a_1) - f_n(a_0)\| &\le \left\|\int_0^1 d\tau J_n(a_\tau)\right\|_{\text{op}} \|a_1 - a_0\| \\ &\le \sup_\theta \|J_n(\theta)\|_{\text{op}}\|a_1 - a_0\| \\ &\le \|\Sigma_n\|_{\text{op}}\|a_1 - a_0\|. \end{aligned}$$

Finally,

$$\|\Sigma_n\|_{\text{op}} \le \|\Sigma\|_{\text{op}} + \|\Sigma_n - \Sigma\|\|_{\text{op}},$$

where $\|\Sigma\|_{\text{op}} = 1 + \|\theta_*\|^2$. Furthermore, since the entries of Y_i are independent and subgaussian with parameter depending only on $\|\theta_*\| \le r$, by concentration of the sample covariance matrix (cf. [36, Exercise 4.7.3]), we have

$$\|\Sigma_n - \Sigma\|_{\text{op}} \le \sqrt{\frac{Cd\log n}{n}}$$

with probability at least $1 - \exp(-C'd\log n)$ for some constants C and C'. ■

Proof of Lemma 13. Note that $Y_\perp$ is a $(d-1)\times n$ matrix with i.i.d. $N(0,1)$ entries. By the Davidson–Szarek bound [7, Theorem II.7],

$$\mathbb{P}\big[\|Y_\perp\|_{\text{op}} \ge \sqrt{n} + \sqrt{d-1} + t\big] \le e^{-t^2/2},$$

which implies (111) since $n \ge d$.

To prove the second claim, it suffices to bound the operator norm of

$$\frac{1}{n}\sum_{i=1}^n Y_{i,1}^2 Y_{i,\perp} Y_{i,\perp}^\top.$$

We first condition on $Y_{i,1}$'s, which are independent of $Y_{i,\perp}$'s. Take $\mathcal{U}$ to be a $\frac{1}{2}$-net of S^{d-2} and $|\mathcal{U}| \le 5^d$. Then,

$$\left\|\frac{1}{n}\sum_{i=1}^n Y_{i,1}^2 Y_{i,\perp} Y_{i,\perp}^\top\right\|_{\text{op}} \le 2\max_{u\in\mathcal{U}} \frac{1}{n}\sum_{i=1}^n Y_{i,1}^2 \langle Y_{i,\perp}, u\rangle^2.$$

Abbreviate $a_i = Y_{i,1}^2$ and $a = (a_1, \dots, a_n)$. By [23, Lemma 1],

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n a_i\big(\langle Y_{i,\perp}, u\rangle^2 - 1\big)\right| \ge \frac{2\|a\|_2\sqrt{t} + \|a\|_\infty t}{n} \;\Big|\; a\right] \le \exp(-t).$$

Furthermore,

$$\sum_{i=1}^n a_i = \sum_{i=1}^n Y_{i,1}^2, \quad \sum_{i=1}^n a_i^2 = \sum_{i=1}^n Y_{i,1}^4,$$

and

$$\|a\|_\infty = \max_{i\in[n]} |Y_{i,1}|^2,$$

which are controlled by the high-probability events H_2, H_4, H_∞ in Lemma 11, respectively. Choosing $t = d\log n$ in the above display and taking the union bound over $u \in \mathcal{U}$, we have, with probability at least $1 - 3n^{-1} - 5^d n^{-d}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n Y_{i,1}^2 Y_{i,\perp} Y_{i,\perp}^\top\right\|_{\mathrm{op}} \le 2\big(1 + \|\theta_*\|^2\big) + C\left(\sqrt{\frac{d\log n}{n}} + \frac{d\log n}{n}\right),$$

where C only depends on $\|\theta_*\|$. The proof is complete in view of the assumption that $n \ge d\log d$. ■

Proof of Lemma 14. Note that each b_i is Rademacher and hence 4-subgaussian. Thus conditioned on any realization of x, $\langle x, b\rangle$ is $4\|x\|^2$-subgaussian, and hence

$$\mathbb{P}\big[|\langle x, b\rangle| \ge \sqrt{s}\|x\| \mid x\big] \le 2\exp(-s/8)$$

for any t. The desired (112) then follows from

$$\mathbb{P}\big[|\langle x, b\rangle| \ge \sqrt{as}\big] \le \mathbb{P}\big[|\langle x, b\rangle| \ge \sqrt{s}\|x\|\big] + \mathbb{P}\big[\|x\| \ge \sqrt{a}\,\big].$$

Finally, (113) follows analogously from the union bound. ■

Proof of Lemma 15. By dilating q, we can assume without loss of generality that $R = 1$. Recall the global assumption $\|\theta_*\| \le r$. Throughout the proof, unless stated to be absolute, all constants depend only on r and L_0. The Lipschitz constant of D on the unit ball $B(1)$ is given by

$$L = \sup_{\theta\in B(1)} \|\nabla D(\theta)\|.$$

It remains to bound L from above with high probability, i.e.,

$$\sup_{\theta\in B(1)} \|\nabla D(\theta)\| \le \sqrt{\frac{L_2 d\log n}{n}} \tag{144}$$

for some constant L_2. Furthermore, the Hessian of D is given by

$$\nabla^2 D(\theta) = \frac{1}{n}\sum_{i=1}^{n} b_i Y_{i,1} Y_i Y_i^\top q''\big(\langle \theta, Y_i\rangle\big).$$

Since $|q''| \leq L_0$, we have

$$\sup_{\theta\in B(1)} \|\nabla^2 D(\theta)\|_{\rm op} \leq L_0 \max_{i\in[n]} |Y_{i,1}| \|Y_i\|^2.$$

In view of (110),

$$\max_{i\in[n]} |Y_{i,1}| \leq \sqrt{\kappa \log n}$$

with probability at least $1 - n^{-2}$. Furthermore,

$$\|Y_i\|^2 \leq 2\|\theta_*\|^2 + 2\|Z_i\|^2.$$

By Lemma 20, for each i,

$$\mathbb{P}\big[\|Z_i\|^2 \geq d + 2\sqrt{dx} + 2x\big] \leq \exp(-x).$$

Since n/d is at least some absolute constant by assumption,

$$\mathbb{P}\big[\|Z_i\|^2 \geq C_2 d \log n\big] \leq n^{-2}$$

for some absolute constant C_2. Therefore, with probability at least $1 - 2n^{-1}$,

$$\sup_{\theta\in B(1)} \|\nabla^2 D(\theta)\|_{\rm op} \leq L_2 d (\log n)^{3/2} \tag{145}$$

for some constant L_2, i.e., $\theta \mapsto \nabla D(\theta)$ is $L_2 d(\log n)^{3/2}$-Lipschitz. Let Θ be a $1/dn^2$-net of the unit ball $B(1)$, with cardinality [36, Corollary 4.2.13]

$$|\Theta| \leq (1 + 2dn^2)^d \leq (1 + 2n^3)^d. \tag{146}$$

Then, in the event of (145),

$$\sup_{\theta\in B(1)} \|\nabla D(\theta)\| \leq \max_{\theta\in\Theta} \|\nabla D(\theta)\| + \frac{L_2(\log n)^{3/2}}{n^2}. \tag{147}$$

Note that

$$\nabla D(\theta) = \frac{1}{n}\sum_{i=1}^{n} b_i Y_{i,1} Y_i q'\big(\langle \theta, Y_i\rangle\big).$$

Let $\mathcal{U}$ be a $\frac{1}{2}$-net of S^{d-1} with cardinality at most

$$|\mathcal{U}| \leq 5^d. \tag{148}$$

Then,

$$\|\nabla D(\theta)\| \le 2 \max_{u \in \mathcal{U}} \langle u, \nabla D(\theta) \rangle,$$

where

$$\langle u, \nabla D(\theta) \rangle = \frac{1}{n} \sum_{i=1}^{n} b_i Y_{i,1} \langle Y_i, u \rangle q'\big(\langle \theta, Y_i \rangle\big).$$

Since $|q'| \le L_0$, $\langle Y_i, u \rangle q'(\langle \theta, Y_i \rangle)$ is C_0-subgaussian, hence $b_i Y_{i,1} \langle Y_i, u \rangle q'(\langle \theta, Y_i \rangle)$ is C_1-subexponential, for some C_0, C_1 depending on L_0 and $\|\theta_*\|$. By Bernstein's inequality,

$$|\langle u, \nabla D(\theta) \rangle| \le C_2 \sqrt{\frac{d \log n}{n}}$$

with probability at least $1 - \exp(-20d \log n)$. By taking the union bound over $u \in \mathcal{U}$ and $\theta \in \Theta$, the proof is completed in view of (146)–(148). ■

Proof of Lemma 16. The proof is almost identical to that of Lemma 15, so we only mention the part that is different. Without loss of generality, assume that $R = 1$. First note that the Lipschitz constant of $M \colon \mathbb{R}^d \to \mathbb{R}^{d-1}$ (with respect to the Euclidean norm) is bounded by

$$\mathrm{Lip}(M) \le L_0 \frac{1}{n} \sum_{i=1}^{n} \|Y_{i,\perp}\| |Y_{i,1}| \big(\|Y_{i,\perp}\| + |Y_{i,1}|\big).$$

Similar to the argument that leads to (147), we conclude that with probability at least $1 - n^{-1}$ $\mathrm{Lip}(M) \le L_2 d \log n$ for some constant L_2.

Next let Θ be a $1/dn$-net of the unit ball in $\mathbb{R}^d$ and let $\mathcal{U}$ be a $\frac{1}{2}$-net of the unit sphere in $\mathbb{R}^{d-1}$. It suffices to bound $\max_{u \in \mathcal{U}, \theta \in \Theta} \langle u, M(\theta) \rangle$. The rest of the proof is identical to that of Lemma 15. ■

Proof of Lemma 17. Note that $y \mapsto T(x, y)$ is an odd function and $T(x, y) \ge 0$ for $y \ge 0$. For the upper bound, note that

$$\partial_y T(x, y)|_{y=0} = \tanh^2(x)$$

and

$$\partial_y^3 T(x, y) = 3\big(\mathrm{sech}(x + y)^4 + \mathrm{sech}(x - y)^4\big) - 2\big(\mathrm{sech}(x + y)^2 + \mathrm{sech}(x - y)^2\big).$$

Since $\sup_{0 \le r \le 1}(3r^4 - 2r^2) = 1$, we have

$$\partial_y^3 T(x, y) \le 2$$

for all x, y. Thus, (128) follows from the Taylor expansion of $T(x, y)$ at $y = 0$ and the fact that $\tanh(x)^2 \le x^2$. Finally, (129) follows from the 1-Lipschitz continuity of g, since $g'(z) = 1 - \mathrm{sech}^2(z)$. ■

Proof of Lemma 19. Recall that

$$Q(x,y) = \frac{1}{2x}\big(\tanh(y+x) - \tanh(y-x)\big).$$

Then,

$$|Q(x,y)| \le 1 \quad \text{and} \quad |\partial_y Q(x,y)| \le 1,$$

and follow from the 1-Lipschitz continuity of tanh and tanh′, respectively. Finally, by Taylor's theorem, we have

$$\begin{aligned}\tanh(y+x) - \tanh(y-x) &= 2x \tanh'(y) \\ &\quad + x \int_0^1 dz (1-z)\{\tanh''(y+xz) + \tanh''(y-xz)\}.\end{aligned}$$

Therefore,

$$\begin{aligned}\partial_x Q(x,y) &= \frac{1}{2}\frac{\partial}{\partial x}\int_0^1 dz(1-z)\{\tanh''(y+xz) + \tanh''(y-xz)\} \\ &= \frac{1}{2}\int_0^1 dz\, z(1-z)\{\tanh'''(y+xz) - \tanh'''(y-xz)\}.\end{aligned}$$

Since $|\tanh'''| \le 2$, we have

$$|\partial_x Q(x,y)| \le 2\int_0^1 dz\, z(1-z) = \frac{1}{3}. \qquad \blacksquare$$

12. Proofs in Section 6

Proof of Lemma 10. Since $\ell(\theta,\theta_*) \le \delta$, without loss of generality, we can assume that $\|\theta - \theta_*\| \le \delta$. Note that $\nabla^2 \ell_n(\theta) = -I + J_n(\theta)$, where $J_n(\theta)$ is the Jacobian of f_n given in (143). Then,

$$I + \nabla^2 \ell_n(\theta) = J_n(\theta) = \mathbb{E}_n\big[YY^\top \mathrm{sech}^2\langle\theta, Y\rangle\big],$$

which is PSD with probability one. Therefore, it remains to bound the maximum eigenvalue of J_n from above uniformly in a neighborhood of θ_*. We do so in two steps.

Step 1: Population version. By assumption, $\|\theta_*\| \ge 100\delta$ for sufficiently large n, and hence $\langle\theta,\theta_*\rangle \ge 0$. Consider the expectation of J_n:

$$J(\theta) \triangleq \mathbb{E}\big[J_n(\theta)\big] = \mathbb{E}\big[YY^\top \mathrm{sech}^2\langle\theta, Y\rangle\big],$$

which is a PSD matrix. We show that

$$\sup_{\|\theta-\theta_*\|\leq\delta} \sup_{\|u\|=1} u^\top J(\theta)u \leq e^{-c\|\theta_*\|^2}. \tag{149}$$

Consider two cases:

Case 1: $u \perp \theta$. In this case, $|\langle u, \theta_*\rangle| = |\langle u, \theta_* - \theta\rangle| \leq \|\theta - \theta_*\| \leq \delta$. By the independence of $\langle u, Z\rangle$ and $\langle \theta, Z\rangle$, we have

$$u^\top J(\theta)u = \mathbb{E}\big[\langle u, Y\rangle^2 \mathrm{sech}^2\langle \theta, Y\rangle\big] = \mathbb{E}\big[\langle u, Y\rangle^2\big]\mathbb{E}\big[\mathrm{sech}^2\langle \theta, Y\rangle\big]. \tag{150}$$

Here, $\mathbb{E}[\langle u, Y\rangle^2] = \langle u, \theta_*\rangle^2 + 1 \leq 1 + \delta^2$. Furthermore, let $\eta \triangleq \theta/\|\theta\|$. Then,

$$U \triangleq \langle \eta, Y\rangle \sim \frac{1}{2}N(-s, 1) + \frac{1}{2}N(s, 1),$$

where $s = \langle \eta, \theta_*\rangle$ satisfies $|s - \|\theta\|| = |\langle \eta, \theta_* - \theta\rangle| \leq \delta$, and hence $s \geq \|\theta_*\| - 2\delta$. By a change of measure (Lemma 26), we have

$$\begin{aligned}\mathbb{E}\big[\mathrm{sech}^2\langle \theta, Y\rangle\big] &= \mathbb{E}\big[\mathrm{sech}^2\big(\|\theta\|U\big)\big] \\ &= \mathbb{E}\big[\cosh(sW)\mathrm{sech}^2\big(\|\theta\|W\big)\big]e^{-s^2/2}, \quad W \sim N(0,1), \\ &\leq F\big(s, \|\theta\|\big)e^{-\|\theta_*\|^2/4}, \quad W \sim N(0,1). \end{aligned} \tag{151}$$

Next put

$$F(a, b) \triangleq \mathbb{E}\big[\cosh(aW)\mathrm{sech}^2(bW)\big], \quad a, b \geq 0.$$

A straightforward calculation shows that

$$\frac{\partial F(a,b)}{\partial b} \leq 0 \quad \text{and} \quad \frac{\partial F(a,b)}{\partial a} \geq 0,$$

i.e., $F(a, b)$ is increasing in a and decreasing in b. Write $b = \|\theta\|$. Since $|s - b| \leq \delta$, we have

$$\begin{aligned} F(s, b) &\leq F(b + \delta, b) \\ &= \underbrace{\mathbb{E}\big[\cosh(\delta W)\mathrm{sech}(bW)\big]}_{\text{(I)}} + \underbrace{\mathbb{E}\big[\sinh(\delta W)\sinh(bW)\mathrm{sech}^2(bW)\big]}_{\text{(II)}}. \end{aligned}$$

The first term satisfies (I) $\leq \mathbb{E}[\cosh(\delta W)] = e^{\delta^2/2}$. For the second term, using the fact that $\tanh(x) \leq x$ when $x \geq 0$, we get the following bound that is, crucially, proportional to $\|\theta_*\|$:

$$\text{(II)} \leq b\,\mathbb{E}\big[W\sinh(\delta W)\big] = b\delta e^{\delta^2/2} \leq 2\|\theta_*\|\delta e^{\delta^2/2}.$$

Combining the above with (151) and (150), we get

$$\begin{aligned}u^\top J(\theta)u &\le (1+\delta^2)\big(1+2\|\theta_*\|\delta\big)e^{\delta^2/2-\|\theta_*\|^2/4}\\ &\le e^{3\delta^2/2+2\|\theta_*\|\delta-\|\theta_*\|^2/4} \le e^{-\|\theta_*\|^2/16}.\end{aligned}$$

Case 2: $u \mathbin{/\!/} \theta$. Without loss of generality, assume $u=\eta$. Entirely analogously to the previous case, we have

$$u^\top J(\theta)u \le \mathbb{E}\big[W^2\cosh(sW)\mathrm{sech}^2\big(\|\theta\|W\big)\big]e^{-\|\theta_*\|^2/4},$$

and

$$\begin{aligned}&\mathbb{E}\big[W^2\cosh(sW)\mathrm{sech}^2\big(\|\theta\|W\big)\big]\\ &\quad\le \mathbb{E}\big[W^2\cosh\big((\|\theta\|+\delta)W\big)\mathrm{sech}^2\big(\|\theta\|W\big)\big]\\ &\quad= \mathbb{E}\big[W^2\cosh(\delta W)\mathrm{sech}(bW)\big]+\mathbb{E}\big[W^2\sinh(\delta W)\sinh(bW)\mathrm{sech}^2(bW)\big]\\ &\quad\le \mathbb{E}\big[W^2\cosh(\delta W)\big]+b\mathbb{E}\big[W^3\sinh(\delta W)\big]\\ &\quad= (1+\delta^2)e^{\delta^2/2}+\|\theta\|\delta(3+\delta^2)e^{\delta^2/2}.\end{aligned}$$

Therefore, $u^\top J(\theta)u \le e^{-\|\theta_*\|^2/50}$.

Finally, we combine the two cases. For an arbitrary unit vector u, let $u=\cos\phi\eta+\sin\phi v$ for some $v\perp\eta$. Then, $\langle v,Y\rangle$ and $\langle\eta,Y\rangle$ are independent, and hence

$$\begin{aligned}u^\top J(\theta)u &= \cos^2\phi\mathbb{E}\big[\langle\eta,Y\rangle^2\mathrm{sech}^2\langle\theta,Y\rangle\big]+\sin^2\phi\mathbb{E}\big[\langle v,Y\rangle^2\mathrm{sech}^2\langle\theta,Y\rangle\big]\\ &\qquad+2\cos\phi\sin\phi\mathbb{E}\big[\langle v,Y\rangle\langle\eta,Y\rangle\mathrm{sech}^2\langle\theta,Y\rangle\big]\\ &= \cos^2\phi\mathbb{E}\big[\langle\eta,Y\rangle^2\mathrm{sech}^2\langle\theta,Y\rangle\big]+\sin^2\phi\mathbb{E}\big[\langle v,Y\rangle^2\mathrm{sech}^2\langle\theta,Y\rangle\big]\\ &\le e^{-\|\theta_*\|^2/50},\end{aligned}$$

where the second equality follows from

$$\mathbb{E}\big[\langle v,Y\rangle\langle\eta,Y\rangle\mathrm{sech}^2\langle\theta,Y\rangle\big]=\mathbb{E}\big[\langle v,Y\rangle\big]\mathbb{E}\big[\langle\eta,Y\rangle\mathrm{sech}^2\langle\theta,Y\rangle\big]=0$$

thanks to independence. This yields the desired (149).

Step 2: Concentration. We show that with probability at least $1-2n^{-1}$,

$$\sup_{\|\theta-\theta_*\|\le\delta}\|J_n(\theta)-J(\theta)\|_{\mathrm{op}} \le \sqrt{\frac{C_0 d\log n}{n}}. \tag{152}$$

Since sech^2 is 1-Lipschitz, we have

$$\begin{aligned}\|J_n(\theta)-J_n(\theta')\|_{\mathrm{op}} &\le \|\mathbb{E}_n\big[YY^\top|\mathrm{sech}^2\langle\theta,Y\rangle-\mathrm{sech}^2\langle\theta',Y\rangle|\big]\|_{\mathrm{op}}\\ &\le \|\theta-\theta'\|\cdot\|\mathbb{E}_n\big[YY^\top\cdot\|Y\|\big]\|_{\mathrm{op}}\\ &\le \|\theta-\theta'\|\cdot\mathbb{E}_n\big[\|Y\|^3\big].\end{aligned}$$

Therefore, in the event H_3 in (109), which has probability at least $1 - n^{-4}$, $\theta \mapsto J_n(\theta)$ is $C_4 d^{3/2}$-Lipschitz with respect to the ℓ_2-norm and the $\|\cdot\|_{\mathrm{op}}$-norm, where C_4 is a constant depending only on r. Let $\mathcal{E}$ be an ε-net of $B(\theta_*, \delta)$ with

$$\varepsilon = \frac{\delta}{\sqrt{d^3 n}} \quad \text{and} \quad |\mathcal{E}| \leq \left(1 + 2\frac{\delta}{\varepsilon}\right)^d \leq \exp(C_5 d \log n).$$

Let $\mathcal{U}$ be a $\frac{1}{2}$-net of S^{d-1} with cardinality at most $|\mathcal{U}| \leq 5^d$. Then,

$$\sup_{\|\theta - \theta_*\| \leq \delta} \|J_n(\theta) - J(\theta)\|_{\mathrm{op}} \leq 2 \sup_{\theta \in \mathcal{E}} \sup_{u \in \mathcal{U}} u^\top \big(J_n(\theta) - J(\theta)\big) u + \frac{2C_4}{\sqrt{n}}. \tag{153}$$

Fix $u \in \mathcal{U}$ and $\theta \in \mathcal{E}$, put

$$U = \langle u, Y \rangle^2 \mathrm{sech}^2 \langle \theta, Y \rangle \quad \text{and} \quad U_i = \langle u, Y_i \rangle^2 \mathrm{sech}^2 \langle \theta, Y_i \rangle.$$

Note that $\langle u, Y \rangle^2$ is subexponential with $\|\langle u, Y \rangle^2\|_{\psi_1} \leq C_1 = C_1(r)$. By the moment characterization of subexponentiality (cf. [36, Proposition 2.7.1]), since $|\mathrm{sech}| \leq 1$, we conclude that

$$\|U\|_{\psi_1} \leq C_2 = C_2(r).$$

By Bernstein's inequality (cf. [36, Theorem 2.8.1]),

$$\begin{aligned} \mathbb{P}\left[|u^\top \big(J_n(\theta) - J(\theta)\big) u| \geq \frac{t}{\sqrt{n}}\right] &= \mathbb{P}\left[|\mathbb{E}_n[U] - \mathbb{E}[U]| \geq \frac{t}{\sqrt{n}}\right] \\ &\leq 2 \exp\left(-c \min\left\{\frac{t^2}{\|U\|_{\psi_1}^2}, \frac{t\sqrt{n}}{\|U\|_{\psi_1}}\right\}\right) \end{aligned}$$

for some absolute constant c. Choosing $t = \sqrt{C_3 d \log n}$ with $C_3 = C_3(r)$ sufficiently large, and in view of the assumption that $n = \Omega(d \log n)$, we conclude that

$$\mathbb{P}\left[|u^\top (J_n(\theta) - J(\theta)) u| \geq \frac{t}{\sqrt{n}}\right] \leq 2 \exp(-2C_5 d \log n).$$

The proof of (152) is completed by applying the union bound over $\theta \in \mathcal{E}$ and $u \in \mathcal{U}$ in (153).

Finally, since $\|\theta_*\|^2 = \Omega(\sqrt{d \log(n)/n})$, combining (152) with (149) yields the lemma. ■

A. Auxiliary results

Lemma 20 ([23, Lemma 1]). *For any $x \geq 0$,*

$$\mathbb{P}[\chi_n^2 \geq 2n + 3x] \leq \mathbb{P}[\chi_n^2 - n \geq 2\sqrt{nx} + 2x] \leq \exp(-x), \tag{154}$$

$$\mathbb{P}[\chi_n^2 \leq n - 2\sqrt{nx}] \leq \exp(-x). \tag{155}$$

Lemma 21. *Let $\varepsilon, \delta > 0$. Assume that the sequence $\{K_t\}$ satisfies $K_0 = 0$ and $K_{t+1} \leq (1+\varepsilon)K_t + \delta$. Then, for all $t \geq 0$,*

$$K_t \leq \frac{\delta}{\varepsilon}\{(1+\varepsilon)^t - 1\}.$$

Proof. This follows simply from induction on t. ■

The following lemma is useful for analyzing the rate of convergence:

Lemma 22 ([27, Appendix A]). *Let*

$$x_{t+1} \leq x_t - h(x_t), \quad x_0 > 0,$$

where $h\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous increasing function with $h(0) = 0$ and $h(x) < x$ for all $x \in (0, x_0)$. Then, $\{x_t\} \subset \mathbb{R}_+$ is a monotonically decreasing sequence converging to the unique fixed point at zero as $n \to \infty$. Furthermore,

$$x_t \leq G^{-1}(t), \quad t \geq 1,$$

where $G\colon [0,1] \to \mathbb{R}_+$ by $G(x) = \int_x^{x_0} \frac{1}{h(\tau)}\, d\tau$.

The proof of Lemma 3 and Lemma 5 on the properties of the population EM map relies on the following auxiliary results.

Lemma 23. *Let $Y = \alpha V + \beta W$, where $\alpha, \beta \geq 0$ and $W \sim N(0,1)$. Also let $\widehat{V}(y) = \mathbb{E}[V \mid Y = y]$. Then,*

1. *$\widehat{V}$ is an increasing function.*
2. *If V has a symmetric distribution in the sense that $V \overset{\text{law}}{=} -V$, then $\widehat{V}$ is an odd function.*

Proof. By scaling, it suffices to consider $\alpha = \beta = 1$. The first item follows from the well-known fact that $\frac{d}{dy}\widehat{V}(y) = \operatorname{Var}(V \mid Y = y) \geq 0$ (see, e.g., [39, eq. (131)]), while the second is due to the fact that W has a symmetric distribution. ■

We also need the following bound on the Mill's ratio due to Ito and McKean [30, Exercise 1, p. 851]:

Lemma 24. *Let $\varphi(x) \triangleq 1/\sqrt{2\pi}\exp(-x^2/2)$ denote the standard normal density and $\bar{\Phi}(x) = \int_x^\infty \varphi(t)\, dt$ the normal tail probability. Then,*

$$\frac{\bar{\Phi}(x)}{\varphi(x)} \leq \frac{2}{\sqrt{2+x^2}+x}.$$

We will invoke Stein's lemma repeatedly, which is included below for completeness.

Lemma 25. *Let* $W \sim N(0,1)$ *and* f *be a differentiable function such that*

$$\mathbb{E}\big[|f'(W)|\big] < \infty.$$

Then,

$$\mathbb{E}\big[Wf(W)\big] = \mathbb{E}\big[f'(W)\big].$$

The following useful result is simply a change of measure from the symmetric 2-GM to the standard normal:

Lemma 26. *Let* $V \sim P_s = \frac{1}{2}N(-s,1) + \frac{1}{2}N(s,1)$ *as in* (4) *and let* $Z \sim N(0,1)$. *Then,*

$$\mathbb{E}\big[f(V)\big] = \mathbb{E}\big[f(Z)\cosh(sZ)\big]e^{-s^2/2}.$$

Proof. This follows from $p_s(z)/\varphi(z) = \cosh(sz)e^{-s^2/2}$. ■

B. Minimax rates

Theorem 10. *For any* $d \geq 2$ *and* $n \in \mathbb{N}$ *and* $s \geq 0$,

$$\inf_{\widehat{\theta}} \sup_{\|\theta_*\|=s} \mathbb{E}_{\theta^*}\big[\ell(\widehat{\theta},\theta_*)\big] \asymp \min\left\{\frac{1}{s}\left(\frac{d}{n} + \sqrt{\frac{d}{n}}\right) + \sqrt{\frac{d}{n}}, s\right\}. \tag{156}$$

Furthermore, for any $d, n \in \mathbb{N}$ *and* $r \geq 0$,

$$\inf_{\widehat{\theta}} \sup_{\|\theta_*\|\leq r} \mathbb{E}_{\theta^*}\big[\ell(\widehat{\theta},\theta_*)\big] \asymp \min\left\{\left(\frac{d}{n}\right)^{1/4} + \sqrt{\frac{d}{n}}, r\right\}. \tag{157}$$

Before proving Theorem 10, we note that the rate in (156) behaves as

$$\inf_{\widehat{\theta}} \sup_{\|\theta_*\|=s} \mathbb{E}_{\theta^*}\big[\ell(\widehat{\theta},\theta_*)\big] \asymp \begin{cases} s, & s \leq \left(\frac{d}{n}\right)^{1/4}, \\ \frac{1}{s}\sqrt{\frac{d}{n}}, & \left(\frac{d}{n}\right)^{1/4} \leq s \leq 1, \\ \sqrt{\frac{d}{n}}, & s \geq 1 \end{cases} \tag{158}$$

for $d \leq n$ and

$$\inf_{\widehat{\theta}} \sup_{\|\theta_*\|=s} \mathbb{E}_{\theta^*}\big[\ell(\widehat{\theta},\theta_*)\big] \asymp \begin{cases} s, & s \leq \sqrt{\frac{d}{n}}, \\ \sqrt{\frac{d}{n}}, & s \geq \sqrt{\frac{d}{n}} \end{cases} \tag{159}$$

for $d \geq n$. The latter case coincides with the ℓ_2-rate of the Gaussian location model.

Upper bound. As before, denote $s = \|\theta_*\|$ and $\eta_* = \theta_*/s$. Let $\varepsilon \triangleq \max\{\sqrt{d/n}, d/n\}$. Since the trivial estimator $\widehat{\theta} = 0$ achieves $\ell(\widehat{\theta}, \theta_*) = s$, it remains to show the upper bound $C_0\sqrt{\varepsilon}$ under the assumption that $\|\theta_*\| \geq C_1\sqrt{\varepsilon}$, for some universal constants C_0, C_1. Let $\widehat{\lambda}$ and $\widehat{\eta}$ denote the top eigenvalue and the associated eigenvector (of unit norm) of the sample covariance matrix $\widehat{\Sigma} \triangleq \mathbb{E}_n[YY^\top]$. Let $\Sigma = \mathbb{E}[YY^\top] = I_d + \theta_*\theta_*^\top$. Consider the estimator:

$$\widehat{\theta} = \widehat{s}\,\widehat{\eta}, \quad \widehat{s} = \sqrt{(\widehat{\lambda} - 1)_+}, \tag{160}$$

where $(x)_+ \triangleq \max\{0, x\}$ for any $x \in \mathbb{R}$. To analyze its loss, recall that $Y = X\theta^* + Z$, where X is Rademacher and independent of $Z \sim N(0, I_d)$. Since

$$\mathbb{E}_n[YY^\top] = \theta_*\theta_*^\top + \mathbb{E}_n[ZZ^\top] + \theta_*\big(\mathbb{E}_n[XZ]\big)^\top + \big(\mathbb{E}_n[XZ]\big)\theta_*^\top,$$

we have

$$\widehat{\Sigma} - \Sigma \overset{\text{law}}{=} \Delta + \frac{1}{\sqrt{n}}\big(\theta_* w^\top + w\theta_*^\top\big),$$

where $w \sim N(0, I_d)$ and $\Delta \triangleq \mathbb{E}_n[ZZ^\top] - I_d$. Consequently,

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq \|\Delta\|_{\text{op}} + \frac{2}{\sqrt{n}}\|w\|\|\theta_*\|.$$

By Davis–Kahan's perturbation bound, we have

$$\ell(\widehat{\eta}, \eta_*) \leq 4\frac{\|\widehat{\Sigma} - \Sigma\|_{\text{op}}}{s^2}.$$

Furthermore, by Weyl's inequality, $|\widehat{\lambda} - 1 - s^2| \leq \|\widehat{\Sigma} - \Sigma\|_{\text{op}}$, and thus

$$|\widehat{s} - s| = \frac{|(\widehat{\lambda} - 1)_+^2 - s^2|}{(\widehat{\lambda} - 1)_+ + s} \leq \frac{|\widehat{\lambda} - 1 - s^2|}{s} \leq \frac{\|\widehat{\Sigma} - \Sigma\|_{\text{op}}}{s}.$$

Applying the triangle inequality and combining the last two displays, we obtain

$$\ell(\widehat{\theta}, \theta_*) \leq |\widehat{s} - s| + s\ell(\widehat{\eta}, \eta_*) \leq 5\frac{\|\widehat{\Sigma} - \Sigma\|_{\text{op}}}{s}.$$

Finally, since $\mathbb{E}[\|\Delta\|_{\text{op}}] \leq C\varepsilon$ (see [36, Theorem 4.7.1]) for some universal constant C and $\mathbb{E}[\|w\|] \leq \sqrt{d}$, taking expectation on both sides, we have

$$\mathbb{E}\ell(\widehat{\theta}, \theta_*) \leq 5\frac{\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\text{op}}}{s} \leq C'\left(\frac{\varepsilon}{s} + \sqrt{\frac{d}{n}}\right)$$

for some universal constant C'. This proves the upper bound part of (156), and, upon taking the supremum over $s \leq r$, that of (157) (since the estimator (160) does not depend on $\|\theta_*\|$).

Lower bound. Recall that

$$P_\theta = \frac{1}{2}N(-\theta, I_d) + \frac{1}{2}N(\theta, I_d);$$

in particular, $P_0 = N(0, I_d)$. Then, straightforward calculation shows that the χ^2-divergence is

$$\chi^2\big(P_\theta \| P_0\big) = \cosh\big(\|\theta\|^2\big) - 1.$$

Since $D(P\|Q) \leq \log(1 + \chi^2(P\|Q))$, the KL divergence is upper bounded by

$$D\big(P_\theta \| P_0\big) \leq \log\cosh\big(\|\theta\|^2\big).$$

Note that $\log\cosh(x) \asymp \min\{x, x^2\}$ for $x \geq 0$. Applying Le Cam's method (two-point argument) to $\theta_* = 0$ versus $\theta_* = \varepsilon$, with $\varepsilon = c_0 \min\{r, n^{-1/4}\}$ for some sufficiently small constant c_0, we obtain the desired lower bound in (157) for $d = 1$.

Next we consider $d \geq 2$. It suffices to prove the lower bound part of (156), which yields that of (157) by taking the supremum over $s \leq r$. Furthermore, since the rate for the Gaussian location model (which is $s \wedge \sqrt{d/n}$) constitutes a lower bound for the Gaussian mixture model, this proves (159) as well as the last case of (158). So next we focus on $2 \leq d \leq n$ and $s \leq 1$.

Let c_0 be some small absolute constant. Let $v_1, \ldots, v_M$ be a c_0-net for the unit sphere $S^{d-2} \cap \mathbb{R}_+^{d-1}$, such that

(a) $\|v_i\| = 1$;

(b) $\ell(v_i, v_j) = \|v_i - v_j\| \geq c_0$ for any $i \neq j$; and

(c) $M \geq \exp(C_0 d)$ for some absolute constant C_0.

Now define $u_0, \ldots, u_M \in \mathbb{R}^d$ by $u_0 = e_1 = [1, 0, \ldots, 0]$ and $u_i = [1 - \varepsilon^2, \varepsilon v_i]$ for $i \in [M]$, where $\varepsilon = c_1 \min\{1, (1/s^2)\sqrt{d/n}\}$ for some small constant c_1. Then, $\ell(u_i, u_j) \geq c_0\varepsilon$ for any distinct $i, j \in [M]$ and $\ell(u_i, u_0) \leq 2c_0\varepsilon$ for any $i \in [M]$. Finally, let $\theta_i = s u_i$ for $i = 0, \ldots, M$. By the key result, Lemma 27 below, the KL radius of $\{P_{\theta_i} : i \in [M]\}$ is at most

$$\max_{i \in [M]} D\big(P_{\theta_i} \| P_{\theta_0}\big) \leq C_1 s^4 \varepsilon^2$$

for some absolute constant C_1. Applying Fano's method [44] yields a lower bound that is a constant factor of $\varepsilon s \asymp \min\{s, (1/s)\sqrt{d/n}\}$.

It remains to prove the following result on the local behavior of KL divergence in the 2-GM model.

Lemma 27. *Let* $0 \leq s \leq 1$. *Then, there exists a universal constant* C, *such that for any* $d \geq 1$ *and* $u, v \in S^{d-1}$,

$$D\big(P_{su} \| P_{sv}\big) \leq C\ell(u, v)^2 s^4. \tag{161}$$

Remark 4. The result (161) can be interpreted in two ways. First, by the local expansion of the KL divergence, we have

$$D\big(P_{\theta'}\|P_\theta\big) = O\big(\|\theta-\theta'\|^2 I(\theta)\big),$$

where $I(\theta)$ is the Fisher information at θ, which, in the 2-GM model, behaves as $\|\theta\|^2$ for small θ (see Remark 2); however, this intuition does not directly lead to the desired dimension-free bound. Additionally, (161) can be "anticipated" by drawing analogy to the covariance model: Suppose the latent variable in the mixture model is standard normal instead of Rademacher. Then,

$$\begin{aligned} D\big(P_{su}\|P_{sv}\big) &= D\big(N(0, I+s^2uu^\top)\|N(0, I+s^2vv^\top)\big) \\ &= \frac{s^4}{2(1+s^2)}\|uu^\top - vv^\top\|_F^2 \asymp s^4\ell(u,v)^2, \end{aligned}$$

where the second identity is from [3, eq. (52)].

Proof of Lemma 27. First of all, by symmetry, it suffices to show

$$D\big(P_{su}\|P_{sv}\big) \le C\|u-v\|^2 s^4. \tag{162}$$

Next, let $\delta = \|u-v\| \in [0, \sqrt{2}]$. By the rotational invariance of the normal distribution, we can and shall assume $v = e_1$ and u satisfies $|u_1 - 1| \le \delta$ and $\|u_\perp\| \le \delta$, where $u_\perp = (u_2, \dots, u_d)$. Let $Q = Q_{Y_1,\dots,Y_d} = P_{sv}$ and $P = P_{Y_1,\dots,Y_d} = P_{sv}$. Then, $Q = P_s \otimes N(0, I_{d-1})$ is a product distribution, while P is not, since under P, $Y_1, \dots, Y_d$ are dependent through the common label; this is where the majority of the technical difficulty of this proof comes from. Next we use the chain rule to evaluate the KL divergence:

$$D\big(P_{Y_1,\dots,Y_d}\|Q_{Y_1,\dots,Y_d}\big) = \underbrace{D\big(P_{Y_1}\|Q_{Y_1}\big)}_{\text{(I)}} + \underbrace{\mathbb{E}\big[D\big(P_{Y_\perp|Y_1}\|N(0, I_{d-1})\big)\big]}_{\text{(II)}},$$

where we used the fact that $Y_\perp$ is standard normal and independent of Y_1 under Q, and the expectation in (II) is taken over P_{Y_1}. In what follows we show that both terms are $O(s^4\delta^2)$.

Bounding (I). Let $u_1 = s+\varepsilon$, where $|\varepsilon| \le s\delta$. Then, (I) $= D(P_{s+\varepsilon}\|P_s)$. Recall $p_\theta(y)$ given in (5) denotes the density function of P_θ. In one dimension, we have $p_\theta(y) = e^{-\theta^2/2}\varphi(y)\cosh(\theta y)$. Then,

$$\begin{aligned} \text{(I)} &\le \chi^2\big(P_{s+\varepsilon}\|P_s\big) \\ &\overset{\text{(a)}}{\le} e^{s^2/2}\int \varphi(y)\big[e^{-(s+\varepsilon)^2/2}\cosh((s+\varepsilon)y) - e^{-s^2/2}\cosh(sy)\big]^2 \\ &\overset{\text{(b)}}{=} e^{s^2/2}\big(\cosh(s^2) - 2\cosh(s(s+\varepsilon)) + \cosh((s+\varepsilon)^2)\big) \overset{\text{(c)}}{\le} C_1 s^2\varepsilon^2 \le C_1 s^4\delta^2, \end{aligned}$$

where (a) is due to $\cosh \geq 1$; (b) follows from the facts that

$$\int dy \varphi(y) \cosh(sy) = e^{s^2/2}, \quad \int dy \varphi(y) \cosh(sy)^2 = e^{s^2} \cosh(s^2),$$
$$2\cosh(a)\cosh(b) = \cosh(a+b) + \cosh(a-b);$$

and (c) is by Taylor expansion since $0 \leq |\varepsilon| \leq \sqrt{2}s \leq \sqrt{2}$, where C_1 is some universal constant.

Bounding (II). Let $Y = (Y_1, Y_\perp)$ and $Y_\perp = (Y_2, \dots, Y_d)$. Under P, we can write $Y_i = R_i + Z_i$, where $R_i = su_i \cdot B$, B is Rademacher and independent of $Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. Therefore, $P_{Y_\perp|Y_1} = P_{R_\perp|Y_1} * N(0, I_{d-1})$ is a Gaussian location mixture (convolution). Recall the Ingster–Suslina identity [18]: for any distribution μ on $\mathbb{R}^d$,

$$\chi^2\big(\mu * N(0, I_d) \| N(0, I_d)\big) = \mathbb{E}\big[\exp\big(\langle X, \widetilde{X}\rangle\big)\big] - 1,$$

where $X, \widetilde{X} \overset{\text{i.i.d.}}{\sim} \mu$. Then, we have

$$(\mathrm{II}) \leq \mathbb{E}\big[\chi^2\big(P_{Y_\perp|Y_1} \| N(0, I_{d-1})\big)\big] = \mathbb{E}\big[\exp\big(\langle R_\perp, \widetilde{R}_\perp\rangle\big)\big] - 1,$$

where $\widetilde{R}_\perp$ is an independent copy of $R_\perp$ conditioned on Y_1. Note that $\|R_\perp\| \leq s\|u_\perp\| \leq s\delta$ almost surely. Then, $|\langle R_\perp, \widetilde{R}_\perp\rangle| \leq (s\delta)^2 \leq 2$. Therefore, by Taylor expansion, we have

$$\mathbb{E}\big[\exp\big(\langle R_\perp, \widetilde{R}_\perp\rangle\big)\big] - 1 \leq \mathbb{E}\big[\langle R_\perp, \widetilde{R}_\perp\rangle\big] + C_2(s\delta)^4,$$

where C_2 is some universal constant. By linearity, we have

$$\begin{aligned}
\mathbb{E}\big[\langle R_\perp, \widetilde{R}_\perp\rangle\big] &= \sum_{i=2}^{d} \mathbb{E}[R_i \widetilde{R}_i] = \sum_{i=2}^{d} \mathbb{E}\big[\mathbb{E}[R_i|Y_1]\mathbb{E}[\widetilde{R}_i|Y_1]\big] \\
&\overset{(a)}{=} \sum_{i=2}^{d} \mathbb{E}\big[\mathbb{E}[R_i|Y_1]^2\big] \overset{(b)}{=} s^2 \sum_{i=2}^{d} u_i^2 \mathbb{E}\big[\mathbb{E}[B|Y_1]^2\big] \\
&\overset{(c)}{=} s^2\delta^2 \mathbb{E}\big[\tanh(u_1 Y_1)^2\big] \overset{(d)}{\leq} 4s^4(1+4s^2)\delta^2 \leq 40 s^4 \delta^2,
\end{aligned}$$

where (a) is because of $\widetilde{R}_i$ is a conditional independent copy of R_i; (b) is due to $R_i = su_i B$; (c) is by $\|u_\perp\| = \delta$ and the conditional mean is given by (7); and (d) is by $|\tanh(x)| \leq |x|$ and $|u_1| \leq s(1+\delta) \leq 2s$.

Finally, combining (I) and (II) completes the proof of (162). ■

C. Analysis of MLE

In this appendix we provide a crude statistical guarantee for the MLE that is needed for proving the convergence of EM to the MLE in Section 6. Existing analysis of MLE typically relies on bounding the Hellinger bracketing entropy of the class of distributions (see, e.g., [34]). Such program has been carried out for the Gaussian mixture model for both the parametric and nonparametric MLE [13, 14, 17, 29, 45]; however, we found it difficult to bound the bracketing entropy accurately in high dimensions. Instead, we opt for a standard argument involving only the usual metric entropy. To state a general result, let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a parametric family of densities. Let $Y_i \overset{\text{i.i.d.}}{\sim} p_{\theta_*}$. The MLE is defined as any global maximizer of the likelihood:

$$\widehat{\theta}_{\text{MLE}} \in \arg\max_{\theta\in\Theta} \ell_n(\theta), \quad \ell_n(\theta) \triangleq \mathbb{E}_n\big[\log p_\theta(Y)\big] = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(Y_i). \tag{163}$$

The following result is standard:

Theorem 11. *Abbreviate $H(\theta,\theta') \triangleq H(p_\theta, p_{\theta'})$. Denote by $\mathcal{N}(\mathcal{P}, H, \varepsilon)$ the ε-covering number of $\mathcal{P}$ (without bracketing) with respect to the Hellinger distance. Then,*

$$\mathbb{P}\big[H(\widehat{\theta}_{\text{MLE}}, \theta_*) > \varepsilon\big] \le \mathcal{N}\left(\mathcal{P}, H, \left(\frac{\varepsilon^2}{8L}\right)^{1/s}\right) \exp(-n\varepsilon^2/4) + \mathbb{P}\big[\text{Lip}_s(\ell_n) > L\big],$$

where $\text{Lip}_s(\ell_n)$ is the s-Lipschitz constant of the (random) function $\theta \mapsto \ell_n(\theta)$ on Θ with respect to the Hellinger distance for some $s > 0$, i.e.,

$$\text{Lip}_s(\ell_n) = \sup_{\theta,\theta'\in\Theta} \frac{|\ell_n(\theta) - \ell_n(\theta')|}{H(\theta,\theta')^s}.$$

Next we apply Theorem 11 in some high-dimensional parametric models:

Gaussian location model. Consider $\mathcal{P}_{\text{GLM}} = \{N(\theta, I_d) : \|\theta\| \le r\}$, where r is a constant and $d \le n$. Then,

$$H(\theta,\theta')^2 = 2 - 2e^{-\|\theta-\theta'\|^2/8}.$$

Thus, on $\Theta = \{\|\theta\| \le r\}$, we have $H(\theta,\theta') \asymp \|\theta - \theta'\|$. By the usual covering number bound for the Euclidean space, we have

$$\mathcal{N}(\mathcal{P}_{\text{GLM}}, H, \delta) \le \left(\frac{C}{\delta}\right)^{Cd}.$$

Furthermore, the log-likelihood process is given by

$$\ell_n(\theta) = \text{constant} - \frac{1}{2}\mathbb{E}_n\big[(Y-\theta)^2\big],$$

with $\nabla \ell_n(\theta) = \mathbb{E}_n[Y - \theta]$. Thus, with high probability,

$$\sup_{\theta \in \Theta} \|\nabla \ell_n(\theta)\| \le C = C(r).$$

Thus, for $s = 1$, we have $\mathrm{Lip}_1(\ell_n) \le C$ with high probability. Applying Theorem 11 with $s = 1$, $L = C$ and $\varepsilon = \sqrt{Cd \log(n)/n}$, we get

$$H(\widehat{\theta}_{\mathrm{MLE}}, \theta_*) \le \sqrt{\frac{Cd \log n}{n}}$$

with high probability.

Symmetric 2-GM. Consider $\mathcal{P}_{2\text{-GM}} = \{\frac{1}{2}N(-\theta, I_d) + \frac{1}{2}N(\theta, I_d) : \|\theta\| \le r\}$, where r is a constant and $d \le n$; this is the setting of the current paper. On $\Theta = \{\|\theta\| \le r\}$, we have

$$H(\theta, \theta')^2 \asymp \|\theta\theta^\top - \theta'\theta'^\top\|_{\mathrm{F}}^2,$$

which follows from the moment tensor characterization of Hellinger for Gaussian mixtures in [9, Theorem 4.1]. Furthermore, since $\ell(\theta, \theta')^2 \lesssim \|\theta\theta^\top - \theta'\theta'^\top\|_{\mathrm{F}} \lesssim \ell(\theta, \theta')$,[6] we have

$$\ell(\theta, \theta')^2 \lesssim H(\theta, \theta') \lesssim \ell(\theta, \theta'). \tag{164}$$

By the covering number bound for rank-one matrices (or one-dimensional subspaces; see [32]), we have $\mathcal{N}(\mathcal{P}_{2\text{-GM}}, H, \delta) \le (C/\delta)^{Cd}$. The analogous result also holds for general Gaussian mixtures; cf. [9, Lemma 4.5]. Next, recall the relation (13) between EM algorithm and the gradient descent, we have

$$\nabla \ell_n(\theta) = -\theta + \mathbb{E}_n\big[Y \tanh\langle \theta, Y\rangle\big].$$

By Theorem 4, with high probability,

$$\sup_{\theta \in \Theta} \|\nabla \ell_n(\theta)\| \le C\sqrt{d}.$$

[6] For the lower bound, assume that $\|\theta\| \ge \|\theta'\|$. Then,

$$\|\theta\theta^\top - \theta'\theta'^\top\|_{\mathrm{F}} \ge \|\theta\theta^\top - \theta'\theta'^\top\|_{\mathrm{op}} \ge \frac{1}{\|\theta\|^2}\theta^\top(\theta\theta^\top - \theta'\theta'^\top)\theta = \|\theta\|^2 - \frac{1}{\|\theta\|^2}|\langle\theta, \theta'\rangle|^2$$
$$\ge \|\theta\|^2 - |\langle\theta, \theta'\rangle| \ge \frac{1}{2}\big(\|\theta\|^2 + \|\theta'\|^2 - 2|\langle\theta, \theta'\rangle|\big) = \frac{1}{2}\ell(\theta, \theta')^2.$$

For the upper bound, assuming that $\|\theta\|, \|\theta'\| \le r$, we have

$$\|\theta\theta^\top - \theta'\theta'^\top\|_{\mathrm{F}}^2 \le 2\|\theta(\theta - \theta')^\top\|_{\mathrm{F}}^2 + 2\|(\theta - \theta')\theta'^\top\|_{\mathrm{F}}^2 = 4r^2\|\theta - \theta'\|^2.$$

Replacing θ' with $-\theta'$ yields $\|\theta\theta^\top - \theta'\theta'^\top\|_{\mathrm{F}} \le 2r\ell(\theta, \theta')$.

Thus, in view of (164), we have $\mathrm{Lip}_{1/2}(\ell_n) \le C\sqrt{d}$ with high probability. Finally, applying Theorem 11 with $s = 1/2$, $L = C\sqrt{d}$, and $\varepsilon = \sqrt{Cd\log(n)/n}$, we get, with high probability,

$$H(\widehat{\theta}_{\mathrm{MLE}}, \theta_*) \le \sqrt{\frac{Cd\log n}{n}},$$

and consequently, in view of (164),

$$\ell(\widehat{\theta}_{\mathrm{MLE}}, \theta_*) \le \left(\frac{Cd\log n}{n}\right)^{1/4}. \tag{165}$$

Proof of Theorem 11. Rewrite (163) as

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta\in\Theta} L_n(\theta),$$

where $L_n(\theta) \triangleq \mathbb{E}_n[\log(p_\theta/p_{\theta_*})(Y)]$ is the log-likelihood ratio process. Note that

$$\big\{H(\widehat{\theta}_{\mathrm{MLE}}, \theta_*) > \varepsilon\big\} = \Big\{\sup_{H(\theta,\theta_*)>\varepsilon} L_n(\theta) > \sup_{H(\theta,\theta_*)\le\varepsilon} L_n(\theta)\Big\} \subset \Big\{\sup_{H(\theta,\theta_*)\ge\varepsilon} L_n(\theta) > 0\Big\},$$

where we used the fact that $L_n(\theta_*) = 0$. Let Θ' be the minimal δ-covering of Θ in Hellinger distance, where $0 < \delta < \varepsilon/2$ is to be specified. Let $S_\varepsilon = \{\theta : H(\theta, \theta_*) \ge \varepsilon\}$ and $\widetilde{\Theta}_\varepsilon = \Theta' \cap S_\varepsilon$. Then, for any $\theta \in S_\varepsilon$, there exist $\widetilde{\theta} \in \widetilde{\Theta}_\varepsilon$ such that both $H(\theta, \widetilde{\theta}) \le \delta$ and $H(\widetilde{\theta}, \theta_*) \ge \varepsilon/2$ and hold. Furthermore, $|\Theta_\varepsilon| \le |\Theta'| \le \mathcal{N}(\mathcal{P}, H, \delta)$.

In the event that $\mathrm{Lip}_s(\ell_n) \le L$, since $\mathrm{Lip}_s(\ell_n) = \mathrm{Lip}_s(L_n)$, by the (s, L)-Lipschitz continuity of L_n, we have

$$\sup_{H(\theta,\theta_*)\ge\varepsilon} L_n(\theta) \le \sup_{H(\theta,\theta_*)\ge\varepsilon,\theta\in\Theta'} L_n(\theta) + L\delta^s.$$

To complete the proof, applying the union bound yields

$$\begin{aligned}
\mathbb{P}\big[H(\widehat{\theta}_{\mathrm{MLE}}, \theta_*) > \varepsilon, \mathrm{Lip}_s(\ell_n) \le L\big] &= \mathbb{P}\big[\sup_{\theta\in S_\varepsilon} L_n(\theta) > 0, \mathrm{Lip}_s(\ell_n) \le L\big] \\
&= \mathbb{P}\big[\sup_{\theta\in\Theta_\varepsilon} L_n(\theta) \ge -L\delta^s\big] \\
&\overset{\text{(a)}}{\le} \mathcal{N}(\mathcal{P}, H, \delta)\exp\Big(-\frac{n}{2}\Big(\frac{\varepsilon^2}{4} - L\delta^s\Big)\Big) \\
&\overset{\text{(b)}}{=} \mathcal{N}\big(\mathcal{P}, H, (\varepsilon^2/(2L))^{1/s}\big)\exp\Big(-\frac{\varepsilon^2 n}{16}\Big),
\end{aligned}$$

where (a) follows from the following well-known result, and in (b) we chose $\delta^s = \varepsilon^2/(8L)$.

Lemma 28. *Let Y_i be i.i.d. with density p. Then, for any density q,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{q}{p}(Y_i) \geq -\delta\right] \leq \exp\left(-\frac{H^2(P,Q)-\delta}{2}n\right).$$

Proof. By Chernoff bound, we have

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{q}{p}(Y_i) \geq -\delta\right] &= \mathbb{P}\left[\exp\left(\frac{1}{2}\sum_{i=1}^{n}\log\frac{q}{p}(Y_i)\right) \geq \exp(-\delta n/2)\right] \\ &\leq \left(\int\sqrt{pq}\right)^n \exp(\delta n/2) \\ &\leq \exp\big(-H^2(P,Q)n/2+\delta n/2\big),\end{aligned}$$

where the last step uses

$$H^2(P,Q) = 2-2\int\sqrt{pq} \quad \text{and} \quad 1-x \leq \exp(-x) \quad \text{for } x>0. \qquad \blacksquare$$

This completes the proof of Theorem 11. ■

Acknowledgments. The authors are grateful for Yuxin Chen for helpful discussions on [4] and Natalie Doss for pointing out [17]. The authors also thank Nir Weinberger for a very careful reading of the draft and supplying many helpful comments.

Funding. The results of this paper were first presented in part at Joint Statistical Meetings (JSM), Vancouver BC, Canada, July 2018 [41]. Y. Wu is supported in part by the NSF grants CCF-1527105, CCF-1900507, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. H. H. Zhou is supported in part by NSF grants DMS 1811740, DMS 1918925, and NIH grant 1P50MH115716.

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu, Statistical guarantees for the EM algorithm: from population to sample-based analysis. *Ann. Statist.* **45** (2017), no. 1, 77–120 Zbl 1367.62052 MR 3611487

[2] C. Biernacki, G. Celeux, and G. Govaert, Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Recent developments in mixture models (Hamburg, 2001). *Comput. Statist. Data Anal.* **41** (2003), no. 3–4, 561–575 Zbl 1429.62235 MR 1968069

[3] T. T. Cai, Z. Ma, and Y. Wu, Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.* **41** (2013), no. 6, 3074–3110 Zbl 1288.62099 MR 3161458

[4] Y. Chen, Y. Chi, J. Fan, and C. Ma, Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Math. Program.* **176** (2019), no. 1–2, Ser. B, 5–37 Zbl 1415.90086 MR 3960803

[5] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-region methods*. MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, PA, 2000 Zbl 0958.65071 MR 1774899

[6] C. Daskalakis, C. Tzamos, and M. Zampetakis, Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 704–710, Proceedings of Machine Learning Research 65, PMLR, 2017

[7] K. R. Davidson and S. J. Szarek, Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pp. 317–366, North-Holland, Amsterdam, 2001 Zbl 1067.46008 MR 1863696

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** (1977), no. 1, 1–38 Zbl 0364.62022 MR 501537

[9] N. Doss, Y. Wu, P. Yang, and H. H. Zhou, Optimal estimation of high-dimensional Gaussian mixtures. 2020, arXiv:2002.05818

[10] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, Challenges with EM in application to weakly identifiable mixture models. 2019, arXiv:1902.00194

[11] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, Singularity, misspecification and the convergence rate of EM. *Ann. Statist.* **48** (2020), no. 6, 3161–3182 Zbl 1462.62382 MR 4185804

[12] N. Fournier and A. Guillin, On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** (2015), no. 3–4, 707–738 Zbl 1325.60042 MR 3383341

[13] C. R. Genovese and L. Wasserman, Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** (2000), no. 4, 1105–1127 Zbl 1105.62333 MR 1810921

[14] S. Ghosal and A. W. van der Vaart, Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** (2001), no. 5, 1233–1263 Zbl 1043.62025 MR 1873329

[15] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Translated from the Russian, Seventh edn., Academic Press, Amsterdam, 2007 Zbl 1208.65001 MR 2360010

[16] P. Heinrich and J. Kahn, Optimal rates for finite mixture estimation. 2015, arXiv:1507.04313

[17] N. Ho and X. Nguyen, Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Statist.* **44** (2016), no. 6, 2726–2755 Zbl 1359.62076 MR 3576559

[18] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. Lect. Notes Stat. 169, Springer, New York, NY, 2003 Zbl 1013.62049 MR 1991446

[19] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan, Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems 29*, pp. 4116–4124, Curran Associates, 2016

[20] D. Karlis and E. Xekalaki, Choosing initial values for the EM algorithm for finite mixtures. Recent developments in mixture models (Hamburg, 2001). *Comput. Statist. Data Anal.* **41** (2003), no. 3–4, 577–590 Zbl 1429.62082 MR 1968070

[21] J. M. Klusowski and W. D Brinda, Statistical guarantees for estimating the centers of a two-component gaussian mixture by EM. 2016, arXiv:1608.02280

[22] J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis, Global convergence of EM algorithm for mixtures of two component linear regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 2055–2110, Proceedings of Machine Learning Research 99, PMLR, 2019

[23] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (2000), no. 5, 1302–1338 Zbl 1105.62328 MR 1805785

[24] Y. Lu and H. H. Zhou, Statistical and computational guarantees of Lloyd's algorithm and its variants. 2016, arXiv:1612.02099

[25] S. Mei, Y. Bai, and A. Montanari, The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** (2018), no. 6A, 2747–2774 Zbl 1409.62117 MR 3851754

[26] M. Ndaoud, Sharp optimal recovery in the two Gaussian mixture model. 2018, arXiv:1812.08078

[27] Y. Polyanskiy and Y. Wu, Dissipation of information in channels with input constraints. *IEEE Trans. Inform. Theory* **62** (2016), no. 1, 35–55 Zbl 1359.94277 MR 3447966

[28] R. A. Redner and H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** (1984), no. 2, 195–239 Zbl 0536.62021 MR 738930

[29] S. Saha and A. Guntuboyina, On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Ann. Statist.* **48** (2020), no. 2, 738–762 Zbl 1454.62120 MR 4102674

[30] G. R. Shorack and J. A. Wellner, *Empirical processes with applications to statistics*. Classics Appl. Math. 59, SIAM, Philadelphia, PA, 2009 Zbl 1171.62057 MR 3396731

[31] F. J Sigworth, Peter C Doerschuk, J.-M. Carazo, and S. H. W. Scheres, An introduction to maximum-likelihood methods in cryo-EM. In *Cryo-EM, Part B: 3-D Reconstruction*, pp. 263–294, Methods in Enzymology 482, Academic Press, Amsterdam, 2010

[32] S. J. Szarek, Nets of Grassmann manifold and orthogonal group. In *Proceedings of Research Workshop on Banach Space Theory*, pp. 169–185, Univ. Iowa Press, Iowa City, IA, 1982 Zbl 0526.53047 MR 724113

[33] M. Talagrand, The transportation cost from the uniform measure to the empirical measure in dimension ≥ 3. *Ann. Probab.* **22** (1994), no. 2, 919–959 Zbl 0809.60015 MR 1288137

[34] S. van de Geer, *Empirical processes in M-estimation*. Cambridge Univ. Press, Cambridge, 2000

[35] A. W. van der Vaart, *Asymptotic statistics*. Camb. Ser. Stat. Probab. Math. 3, Cambridge Univ. Press, Cambridge, 1998 Zbl 0910.62001 MR 1652247

[36] R. Vershynin, *High-dimensional probability. An introduction with applications in data science*. Camb. Ser. Stat. Probab. Math. 47, Cambridge Univ. Press, Cambridge, 2018 Zbl 1430.60005 MR 3837109

[37] C. Villani, *Topics in optimal transportation*. Grad. Stud. Math. 58, Amer. Math. Soc., Providence, RI, 2003 Zbl 1106.90001 MR 1964483

[38] N. Weinberger and G. Bresler, The EM algorithm is adaptively-optimal for unbalanced symmetric Gaussian mixtures. To appear in *J. Mach. Learn. Res.*

[39] Y. Wu and S. Verdú, Functional properties of minimum mean-square error and mutual information. *IEEE Trans. Inform. Theory* **58** (2012), no. 3, 1289–1301 Zbl 1365.60002 MR 2932809

[40] Y. Wu and P. Yang, Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.* **48** (2020), no. 4, 1981–2007 Zbl 1455.62075 MR 4134783

[41] Y. Wu and H. H. Zhou, EM algorithm achieves the near-optimal rate for two-component symmetric Gaussian mixtures in $O(\sqrt{n})$ iterations. In *JSM Proceedings*, American Statistical Association, Alexandria, VA, 2018

[42] J. Xu, D. J. Hsu, and A. Maleki, Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*, pp. 2676–2684, Curran Associates, 2016

[43] L. Xu and M. I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8** (1996), no. 1, 129–151

[44] Y. Yang and A. Barron, Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** (1999), no. 5, 1564–1599 Zbl 0978.62008 MR 1742500

[45] C.-H. Zhang, Generalized maximum likelihood estimation of normal mixture densities. *Statist. Sinica* **19** (2009), no. 3, 1297–1318 Zbl 1166.62013 MR 2536157

Received 21 March 2020; revised 4 December 2021.

Yihong Wu

Department of Statistics and Data Science, Yale University, Room 235, 10 Hillhouse Avenue,
New Haven, CT 06511, USA; yihong.wu@yale.edu

Harrison H. Zhou

Department of Statistics and Data Science, Yale University, Room 233, 10 Hillhouse Avenue,
New Haven, CT 06511, USA; huibin.zhou@yale.edu