Sharp analysis of EM for learning mixtures of pairwise differences

Abhishek Dhawan ABHISHEKDHAWAN@GATECH.EDU
School of Mathematics, Georgia Institute of Technology

Cheng Mao CHENG.MAO@MATH.GATECH.EDU
School of Mathematics, Georgia Institute of Technology

Ashwin Pananjady ASHWINPM@GATECH.EDU
Schools of Industrial and Systems Engineering and Electrical and Computer Engineering, Georgia Institute of Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider a symmetric mixture of linear regressions with random samples from the pairwise comparison design, which can be seen as a noisy version of a type of Euclidean distance geometry problem. We analyze the expectation-maximization (EM) algorithm locally around the ground truth and establish that the sequence converges linearly, providing an ℓ_∞-norm guarantee on the estimation error of the iterates. Furthermore, we show that the limit of the EM sequence achieves the sharp rate of estimation in the ℓ_2-norm, matching the information-theoretically optimal constant. We also argue through simulation that convergence from a random initialization is much more delicate in this setting, and does not appear to occur in general. Our results show that the EM algorithm can exhibit several unique behaviors when the covariate distribution is suitably structured.

Keywords: Mixture of linear regressions, pairwise comparison, EM algorithm, Euclidean distance geometry problem

1. Introduction

Mixtures of linear regressions are classical methods used to model heterogeneous populations (Jordan and Jacobs, 1994; Xu et al., 1994; Viele and Tong, 2002) and have been widely studied in recent years (see, e.g., Chaganty and Liang (2013); Chen et al. (2017); Balakrishnan et al. (2017); Li and Liang (2018); Kwon et al. (2019); Chen et al. (2020)). A common approach is to apply a spectral algorithm to initialize the parameters of interest, and then run a *locally convergent* nonconvex optimization algorithm; alternatively, one could even run the nonconvex algorithm from a random initialization. These nonconvex optimization algorithms have been extensively analyzed under Gaussian assumptions on the covariates (see, e.g. Balakrishnan et al. (2017); Klusowski et al. (2019); Chen et al. (2019); Kwon and Caramanis (2020)), and it is known that they can exhibit favorable behavior globally in some settings.

In many tasks such as ranking (Bradley and Terry, 1952), crowd-labeling (Dawid and Skene, 1979) and web search (Chen et al., 2013), however, measurements are taken as *pairwise comparisons* between entities, which renders the covariates inherently *discrete*. These designs or covariates are far from Gaussian, and it is natural to ask how standard nonconvex optimization algorithms now perform. Motivated by this question, we study how the celebrated expectation-maximization (EM) algorithm behaves when estimating a mixture of symmetric linear regressions under the pairwise comparison design. More specifically, we consider the following model of a symmetric mixture of

two linear regressions with parameter vectors $\theta^*, -\theta^* \in \mathbb{R}^d$. For $r = 1, \dots, N$, suppose that we observe $(x_r, y_r) \in \mathbb{R}^d \times \mathbb{R}$ satisfying

$$y_r = z_r \, x_r^\top \theta^* + \epsilon_r. \tag{1}$$

Here the covariates x_r follow the pairwise comparison design, i.e., $x_r = e_{i_r} - e_{j_r}$ for $i_r, j_r \in [d] := \{1, \dots, d\}$, where e_i is the ith standard basis vector in $\mathbb{R}^d$; $z_r \in \{-1, 1\}$ is a latent sign indicating whether the observation is from component θ^* or $-\theta^*$; ϵ_r models additive noise, assumed to be i.i.d. Gaussian for theoretical results. Our goal is to estimate the parameter vector $\theta^* \in \mathbb{R}^d$.

The model (1) can also be seen as a special case of the *Euclidean distance geometry problem* (Gower, 1982; Liberti et al., 2014) and multidimensional scaling (Borg and Groenen, 2005), and bears some resemblance to real phase retrieval (see Eldar and Mendelson (2014); Chen et al. (2019) and the references therein). Let $y'_r = |y_r|$ be the observation and $\epsilon'_r = z_r \epsilon_r$ denote noise. Because the sign z_r is unobserved, (1) is equivalent to

$$y'_r = |x_r^\top \theta^* + \epsilon'_r|. \tag{2}$$

Since $x_r = e_{i_r} - e_{j_r}$, we observe $|\theta^*_{i_r} - \theta^*_{j_r} + \epsilon'_r|$ which is the distance between two parameters contaminated by noise[1]. Hence our goal is to reconstruct $\theta^*_1, \dots, \theta^*_d$ from pairwise distances. This is a noisy version of the classical Euclidean distance geometry problem, also known as multidimensional scaling (although the latter is more often used for a class of visualization methods). The above model is also related to the problem of angular or phase synchronization (Singer, 2011; Zhong and Boumal, 2018; Gao and Zhang, 2021), where angles $\theta^*_1, \dots, \theta^*_d \in [0, 2\pi)$ are estimated from noisy measurements of $\theta^*_i - \theta^*_j \bmod 2\pi$. Moreover, a heterogeneous version of angular synchronization (Cucuringu and Tyagi, 2020) aims to reconstruct multiple groups of angles from a mixture of pairwise differences between angles from unknown groups. For all the aforementioned observation models, researchers have considered convex, spectral, and nonconvex fitting approaches.

As mentioned previously, our focus is on the EM algorithm (Dempster et al., 1977)—and we discuss other estimation procedures briefly in Section 4—which has long been used for estimation in latent variable models. Asymptotic convergence guarantees for EM to local optima in general latent variable models are classical (Wu, 1983). For Gaussian covariates x_r in (1), nonasymptotic, local linear convergence of the EM algorithm for symmetric mixtures of linear regressions was established in Balakrishnan et al. (2017). In the same setting, a form of "global convergence"—where the EM algorithm is first initialized with a close relative called "Easy-EM" which is in itself randomly initialized (see the background section to follow)—was established in Kwon et al. (2019). There is a large literature on EM and other iterative algorithms for mixtures of regressions and related models; see, e.g., Daskalakis et al. (2017); Xu et al. (2016); Li and Liang (2018); Klusowski et al. (2019); Wu and Zhou (2021); Dwivedi et al. (2020); Kwon and Caramanis (2020); Kwon et al. (2021) and references therein. EM has also been analyzed in progressively more complex settings, most recently in Gaussian latent tree models (Dagan et al., 2022). Our paper should be seen, in spirit, as extending this line of investigation.

1.1. Contributions and organization

In this paper, we apply the EM algorithm to model (1) and establish its local linear convergence around the ground truth θ^*, proving that once the algorithm is initialized sufficiently close to the

1. It is more natural to have the noise ϵ'_r inside the absolute value because the distance is nonnegative.

ground truth θ^*, it converges linearly fast to a fixed point $\hat{\theta}$. Moreover, we provide optimal bounds on the ℓ_∞ and ℓ_2 estimation errors achieved by the EM fixed point in estimating θ^*. Despite the extensive study of the EM algorithm for mixture models, our results differ from existing works in several crucial ways.

First, we establish *entrywise* guarantees on the EM algorithm—our results hold in the stronger ℓ_∞-norm, not just the ℓ_2-norm. Second, and in contrast to several results in the Gaussian case for mixtures of linear regressions (Kwon et al., 2019; Chandrasekher et al., 2021), we do not assume that each iterate of the EM algorithm takes in a fresh sample, i.e., our convergence guarantee is not based on sample-splitting[2]. Third, our result for the ℓ_2-norm is *sharp* in that it also is optimal in terms of the constant factor, while most previous results for EM only achieve the optimal rate up to constant factors. Fourth, our simulation results show that EM's convergence from a random initialization is quite delicate: convergence does not occur in general and this stands in sharp contrast to the case with Gaussian covariates. On a technical level, our results are enabled by analyzing the finite-sample EM operator *directly*, whereas many previous results proceed through the population update. En route to establishing our results, we analyze the $\ell_{\infty\to\infty}$ operator norm of the psedoinverse of the sample covariance in this problem, a result that may be of independent interest (see Section 5).

The rest of this paper is organized as follows. In Section 2, we formally introduce our assumptions and the EM iteration that we study. In Section 3, we state and discuss the main theorems proved in this paper. In Section 4, we present a host of experiments showing that the conditions appearing in our theorem statements are indeed necessary in some respect. These experiments also confirm that global convergence of EM in this setting is a delicate phenomenon and does not appear to occur in general. Section 4 also discusses other algorithms and the related literature, and discusses the (non-)identifiability of mixtures with more than two components. We conclude with a sketch of the proof techniques in Section 5. Full proofs can be found in Section A.

1.2. Notation

Let d and N be positive integers such that $3 \le d \le N$ throughout the paper. Define $[d] := \{1, \dots, d\}$ and $\binom{[d]}{2} := \{(i,j) : i,j \in [d],\, i < j\}$. Let $\mathbf{1}$ denote the all-ones vector in $\mathbb{R}^d$, and let $\mathcal{H}$ denote the orthogonal complement of $\mathbf{1}$ in $\mathbb{R}^d$. We use the standard asymptotic notation $O(\cdot)$ and $o(\cdot)$ as $N \to \infty$; we use $O_p(\cdot)$ and $o_p(\cdot)$ when the asymptotic relation holds between two random variables with high probability. We also write $a_N \ll b_N$ if $a_N = o(b_N)$, $a_N \gg b_N$ if $b_N = o(a_N)$, and $a_N \lesssim b_N$ if $a_N = O(b_N)$. For a matrix $A \in \mathbb{R}^{d\times d}$, we denote the trace of A by $\mathsf{tr}(A)$.

2. Background and problem formulation

For the results to follow, we consider model (1) with the following assumptions.

Assumption 1 *In (1), the covariates are $x_r = e_{i_r} - e_{j_r}$ where $(i_r, j_r)_{r=1}^N$ are i.i.d. and uniformly random in $\binom{[d]}{2} = \{(i,j) : i,j \in [d],\, i < j\}$. The noise terms $(\epsilon_r)_{r=1}^N$ are i.i.d. $\mathcal{N}(0,\sigma^2)$ and independent from the covariates. Moreover, $\mathbf{1}^\top \theta^* = 0$, i.e., $\theta^* \in \mathcal{H}$ where $\mathcal{H}$ is the orthogonal complement of the all-ones vector $\mathbf{1}$ in $\mathbb{R}^d$.*

The assumption on the covariates states that the comparisons are chosen uniformly at random and with replacement among all pairs of indices. There is no loss of generality in assuming $\mathbf{1}^\top \theta^* = 0$,

2. Some results of Balakrishnan et al. (2017) do not assume sample-splitting but are suboptimal in the low-noise regime.

because the model (1) is invariant under any constant shift of θ^*: If θ^* is replaced by $\theta^* + c\,\mathbf{1}$ for a constant $c \in \mathbb{R}$, the response y_r stays the same. Define the sample covariance matrix of the covariates to be

$$\hat{\Sigma} := \frac{d-1}{2N} \sum_{r=1}^{N} x_r x_r^\top. \tag{3}$$

Note that $\mathbf{1}$ is in the kernel of $\hat{\Sigma}$ and $\mathcal{H}$ is an invariant subspace of $\hat{\Sigma}$. Therefore, $\hat{\Sigma}$ can be viewed as a linear map on $\mathcal{H}$. Let $\hat{\Sigma}^\dagger$ denote the pseudoinverse of $\hat{\Sigma}$; it is the true inverse of $\hat{\Sigma}$ if $\hat{\Sigma}$ is restricted to $\mathcal{H}$ and is invertible. Moreover, the normalization $\frac{d-1}{2N}$ is chosen so that $\mathbb{E}[\hat{\Sigma}] \in \mathbb{R}^{d\times d}$ has entries

$$(\mathbb{E}[\hat{\Sigma}])_{ij} = \begin{cases} (d-1)/d & \text{if } i = j, \\ -1/d & \text{if } i \neq j. \end{cases}$$

It is not hard to check that $\mathbb{E}[\hat{\Sigma}]$ is the identity map on $\mathcal{H}$.

It is convenient to define the so-called "Easy-EM" operator $\bar{Q} : \mathcal{H} \to \mathcal{H}$ (Kwon et al., 2019) and write the EM operator $\hat{Q} : \mathcal{H} \to \mathcal{H}$ in terms of it:

$$\bar{Q}(\theta) := \frac{d-1}{2N} \sum_{r=1}^{N} \tanh\Big(\frac{y_r\, x_r^\top \theta}{\sigma^2}\Big) y_r x_r, \tag{4a}$$

$$\hat{Q}(\theta) := \hat{\Sigma}^\dagger \bar{Q}(\theta). \tag{4b}$$

See Balakrishnan et al. (2017) for a derivation of the EM operator. Starting from an initialization $\theta^{(0)}$, the EM iteration is then given by

$$\theta^{(t+1)} := \hat{Q}(\theta^{(t)}) = \Big(\sum_{r=1}^{N} x_r x_r^\top\Big)^\dagger \sum_{r=1}^{N} \tanh\Big(\frac{y_r\, x_r^\top \theta^{(t)}}{\sigma^2}\Big) y_r x_r, \qquad t \geq 0. \tag{5}$$

3. Main results

It is helpful to define the following class of parameter vectors. For a constant $\beta > 0$, let

$$\Theta(\beta) := \Big\{\theta \in \mathcal{H} : \theta_1 \leq \cdots \leq \theta_d,\ |\theta_i - \theta_j| \geq \beta \frac{|i-j|}{d} \text{ for any } (i,j) \in \binom{[d]}{2}\Big\}. \tag{6}$$

Operationally, the set $\Theta(\beta)$ is a natural family of parameters to consider. First, up to a relabeling of the coordinates, there is no loss of generality in restricting our attention to parameter vectors with nondecreasing entries. Second, each parameter vector in $\Theta(\beta)$ has its ith and jth entries separated by at least $\beta\frac{|i-j|}{d}$. This separation condition is imposed to simplify the statement of our main results. It also appeared in, e.g., Chen et al. (2022), where the pairwise comparison design was studied. Our full results in Section A are more general, assuming a condition weaker than that in (6) (see (36b) in particular). In short, for every $i \in [d]$, we require the set $\{j \in [d] : |\theta_i - \theta_j| \leq c_1\beta\}$ to contain at most $c_2\, d$ elements for some constants $c_1, c_2 > 0$; in other words, θ_i should not have too many "neighbors" θ_j. This weaker condition is satisfied with high probability if θ_i^* are i.i.d. uniform random variables in $[-1, 1]$, for example.

With this setup in hand, we are now ready to state our main results. Our first result shows that the EM sequence $\{\theta^{(t)}\}_{t\geq 0}$ converges linearly to a limit $\hat{\theta}$ around the ground truth θ^*, and provides an estimation error guarantee in the ℓ_∞-norm with high probability.

Theorem 1 *Consider the model in (1) satisfying Assumption 1. For any fixed constants $\rho, D > 0$, there exist constants $N_0, c_1, C_2 > 0$ depending only on ρ and D such that the following holds. For $\beta > 0$, suppose $\theta^* \in \Theta(\beta)$ as defined in (6). Suppose $N \geq \max\{d^{1+\rho}, N_0\}$ and $\sigma \leq c_1\beta$. Fix $\theta^{(0)} \in \mathcal{H}$ such that $\|\theta^{(0)} - \theta^*\|_\infty \leq c_1\beta$. Let $\{\theta^{(t)}\}_{t\geq 0}$ be the EM iterates defined in (5). Moreover, let*

$$\tau := C_2\sigma\sqrt{\frac{d}{N}\log N}, \qquad T := \max\left\{0, \left\lceil \log_{4/3}\left(\frac{\|\theta^{(0)} - \theta^*\|_\infty}{4\tau}\right)\right\rceil\right\}.$$

Then it holds with probability at least $1 - N^{-D}$ that

- *there exists $\hat{\theta} \in \mathcal{H}$ such that $\hat{Q}(\hat{\theta}) = \hat{\theta}$ and $\|\hat{\theta} - \theta^*\|_\infty \leq 4\tau$;*
- *the sequence $\{\theta^{(t)}\}_{t\geq 0}$ converges to $\hat{\theta}$;*
- *$\|\theta^{(T+t)} - \hat{\theta}\,\|_\infty \leq 8\tau/2^t$ for all $t \geq 0$.*

Theorem 1 is proved in Section A.6 and is a result of Proposition 17. We begin with a discussion of the conditions in the theorem. First, the setting is nearly high-dimensional because we only require $N \geq d^{1+\rho}$ for an arbitrarily small $\rho > 0$. We remark that it is possible to improve this condition to $N \geq d \cdot \mathrm{polylog}(d)$ by a more careful application of Proposition 17 at the cost of complicating the conditions on the other parameters (see Section C for details). Focusing on the setting where β is a constant for clarity, we have $|\theta_i^* - \theta_j^*| \asymp |i - j|/d$, and $\theta_i^* = O(1)$ for all $i, j \in [d]$. The constant-sized noise condition $\sigma \leq c_1\beta$ is mild, because it should be compared against the minimum signal strength $|\theta_i^* - \theta_{i+1}^*| \asymp 1/d$ (see Eq. (6)). Next, the initialization $\theta^{(0)}$ has to be close to the true parameter vector θ^* so that $\|\theta^{(0)} - \theta^*\|_\infty \leq c_1\beta$. This is the sense in which the convergence guarantee is local, and we conjecture that some such condition is necessary (see Section 4 for experimental verification of this conjecture).

With these conditions assumed, we now explain the conclusions of Theorem 1. It is useful to note that $\tau \asymp \sigma\sqrt{\frac{d}{N}\log N}$ is the optimal rate[3] we expect to have when estimating each entry θ_i^* of a d-dimensional vector from N samples with probability $1 - N^{-D}$. It is easiest to interpret Theorem 1 as a two-stage result for the EM iteration:

- First, the quantity T indicates that it takes at most a logarithmic number of steps for the EM iteration to get from distance $\|\theta^{(0)} - \theta^*\|_\infty$ to an $O(\tau)$-neighborhood around the ground truth θ^*. By the bounds $\|\hat{\theta} - \theta^*\|_\infty \leq 4\tau$ and $\|\theta^{(T)} - \hat{\theta}\|_\infty \leq 8\tau$ (by taking $t = 0$ in the last statement), we have $\|\theta^{(T)} - \theta^*\|_\infty = O(\tau)$. Hence, $\theta^{(T)}$ already achieves the optimal rate of estimation in the ℓ_∞-norm.
- Second, beyond time T, the EM sequence $\{\theta^{(t)}\}_{t\geq 0}$ eventually converges to a limit $\hat{\theta}$, and the rate of convergence is linear. In other words, the EM operator is locally contractive. Moreover, the limit $\hat{\theta}$ lies in the 4τ-neighborhood of θ^* in the ℓ_∞-norm.

Finally, we once again remark that the theorem does not require any resampling for the EM iteration. The conclusions of the theorem hold simultaneously on a high-probability event.

Our second theorem shows that the limit $\hat{\theta}$ of the EM sequence achieves the sharp estimation error in the ℓ_2-norm with high probability in the low-noise regime, in that the constant is also sharp.

3. In fact, this is the optimal rate even if we consider linear regression without a mixture, i.e., where all the signs z_r are known a priori.

Theorem 2 *In the setting of Theorem 1 (which in particular assumes $N \geq d^{1+\rho}$ for fixed $\rho > 0$), we additionally assume $\sigma \ll \frac{\beta}{(\log N)^2}$ and $d \gg \log N$. Then it holds with probability at least $1 - N^{-D}$ that*

$$\|\hat{\theta} - \theta^*\|_2^2 \leq (1 + o_p(1))\, \sigma^2 \frac{d-1}{2N} \mathsf{tr}(\hat{\Sigma}^\dagger).$$

The proof of Theorem 2 is deferred to Section A.6. Compared to the noise condition $\sigma \leq c_1\beta$ in Theorem 1, here we need σ to be smaller by a polylogarithmic factor, and an additional, mild condition on the dimension. While we assume a random design of x_r and thus the right-hand side of the above equation is random, the result can be understood as follows. Conditional on a typical realization of the covariates $(x_r)_{r=1}^N$, the squared error $\|\hat{\theta} - \theta^*\|_2^2$ is within a $(1 + o(1))$ factor of

$$\sigma^2 \frac{d-1}{2N} \mathsf{tr}(\hat{\Sigma}^\dagger) = \sigma^2 \,\mathsf{tr}\left(\Big(\sum_{r=1}^N x_r x_r^\top\Big)^\dagger\right) \tag{7}$$

with high probability over the randomness of the noise $(\epsilon_r)_{r=1}^N$. Note that (7) is the optimal rate even for vanilla linear regression with a fixed design (see Section D for more discussion). As a result, we obtain the optimal ℓ_2 error (including the sharp constant) for the EM algorithm provided that the initialization is sufficiently informative and the noise level is sufficiently small.

4. Experiments and discussion

In this section, we further discuss our model and results, related methods, and possible extensions. Our discussion is accompanied by numerical experiments.

4.1. Failure of random initialization

Theorem 1 requires that the initialization of the EM algorithm satisfies $\|\theta^{(0)} - \theta^*\|_\infty \leq c\beta$ for a constant $c > 0$. We believe that this is not an artifact of the analysis and that some such condition is necessary. To test this hypothesis, let us numerically test if a "random" initialization suffices for the convergence of EM.

For simplicity, we take $\theta_i^* = \frac{i}{d} - \frac{d+1}{2d}$ for $i \in [d]$ so that $|\theta_i^* - \theta_j^*| = \frac{|i-j|}{d}$. Then $\theta^* \in \Theta(\beta)$ with $\beta = 1$. Let $\theta \in \mathbb{R}^d$ be a random vector whose entries are i.i.d. uniform in $[-0.5, 0.5]$; then define $\theta^R = \theta - a\mathbf{1}$ where the scalar $a \in \mathbb{R}$ is chosen so that $\mathbf{1}^\top \theta^R = 0$. The vector θ^R is a canonical random initialization, because the entries of both θ^R and θ^* are approximately in the range $[-0.5, 0.5]$, yet θ^R is random. For $\eta \in [0, 1]$, define $\theta^{(0)} = \theta^{(0)}(\eta) = (1-\eta)\theta^* + \eta\,\theta^R$, which interpolates between the ground truth θ^* and the random vector θ^R. We use $\theta^{(0)}$ as the initialization of the EM algorithm and study its performance. Note that

$$\|\theta^{(0)} - \theta^*\|_\infty = \eta\, \|\theta^R - \theta^*\|_\infty$$

which is approximately η. Our theory predicts that the EM algorithm performs well when η is a small constant. On the other hand, if a random initialization is not sufficient, then the EM algorithm will have poor performance when η is close to 1. We indeed observe this phenomenon in the left plot of Figure 1.

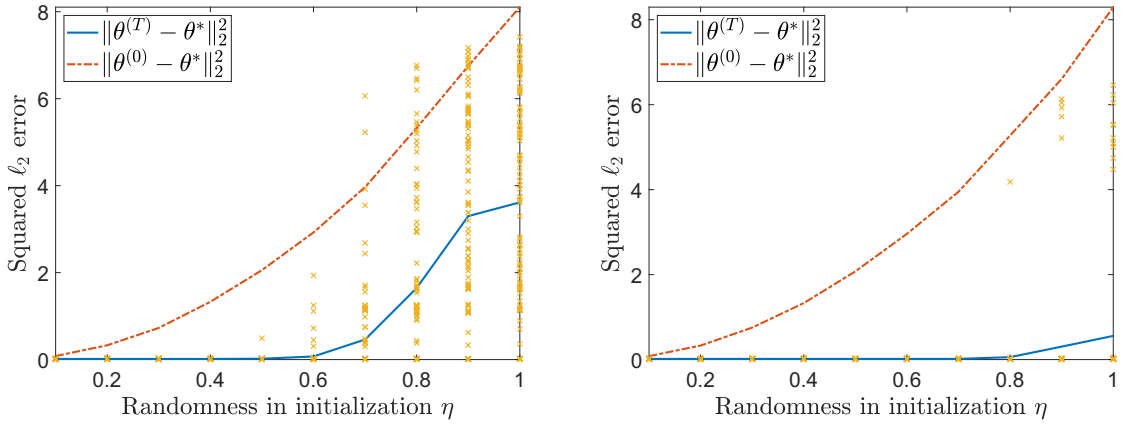

Figure 1. *Left:* Pairwise-comparison design. *Right:* Gaussian design. We plot the initial squared error $\|\theta^{(0)} - \theta^*\|_2^2$ (dashed red line) and the eventual squared error $\|\theta^{(T)} - \theta^*\|_2^2$ (solid blue line) of the EM algorithm against a varying randomness parameter η in the initialization. The experiment for each value of η is averaged over 100 repetitions, and the error $\|\theta^{(T)} - \theta^*\|_2^2$ for each repetition is indicated by a yellow cross.

To be more precise, we set $d = 50$, $N = 1000$, and $\sigma = 0.1$. For $\eta \in \{0.1, 0.2, \ldots, 1\}$, we start from $\theta^{(0)} = \theta^{(0)}(\eta)$ and run the EM algorithm for $T = 100$ steps[4]. We plot the initial squared error $\|\theta^{(0)} - \theta^*\|_2^2$ (dashed red line) and the eventual squared error $\|\theta^{(T)} - \theta^*\|_2^2$ (solid blue line)[5], averaged over 100 independent repetitions for each value of η. The error $\|\theta^{(T)} - \theta^*\|_2^2$ for each repetition is indicated by a yellow cross. As shown by the left plot of Figure 1, an initialization with $\eta \le 0.5$ leads to an extremely small error $\|\theta^{(T)} - \theta^*\|_2^2$ with high probability, while for an initialization with a larger η, the EM algorithm fails to achieve a small error with constant probability.

The failure of random initialization in the setting of the pairwise comparison design stands in sharp contrast to its behavior for Gaussian designs, as confirmed by the right plot of Figure 1. The setup and parameters of this experiment are exactly the same as before, except that we now take i.i.d. covariates $x_r \sim \mathcal{N}(0, \frac{2}{d} I)$, where the normalization $\frac{2}{d}$ is chosen to match the scaling of x_r in Assumption 1. As we see in the figure, the EM algorithm achieves a small error $\|\theta^{(T)} - \theta^*\|_2^2$ with high probability[6] even if η is close to 1. We remark that there is a sizable literature on the success of iterative algorithms with random initialization for a variety of problems, such as phase retrieval (Chen et al., 2019), Gaussian mixtures (Dwivedi et al., 2020; Wu and Zhou, 2021), mixtures of log-concave distributions (Qian et al., 2019), and general regression models with Gaussian covariates (Chandrasekher et al., 2021). However, Gaussianity or continuous density is typically part of the assumption, and analyzing iterative algorithms beyond the Gaussian setting appears to be a generally more challenging problem. See also Dudeja et al. (2022); Wang et al. (2022) for recent work in this direction in the framework of approximate message passing.

4. The EM algorithm typically converges within a few steps, especially when the initialization is close to the ground truth, but we make the conservative choice $T = 100$ to ensure convergence.
5. Since the model (1) is invariant if θ^* is replaced by $-\theta^*$, convergence to a neighborhood of $-\theta^*$ should also be considered as a success. Hence, we actually compute the error $\min\{\|\theta^{(T)} - \theta^*\|_2^2, \|\theta^{(T)} + \theta^*\|_2^2\}$ in all our experiments. For brevity, we do no mention this elsewhere.
6. The occasional failures of random initialization with $\eta = 1$ can be explained by a few factors, such as our moderate choices of d and N, the nontrivial error probability, etc.

An interesting open problem is to offer an explanation for the contrasting behavior of the randomly initialized EM algorithm given the two types of covariates. We speculate that random initialization tends to fail if the covariates are discrete and sparse. Note that our covariates x_r are supported in the highly structured set $\{e_i - e_j : (i,j) \in \binom{[d]}{2}\}$, unlike the Gaussian covariates which are in general positions in $\mathbb{R}^d$. However, we do not have a theoretical explanation for this phenomenon and leave the question to future work.

4.2. Spectral method, initialization, and comparison

Spectral methods form another popular class of algorithms for tackling mixture models and the Euclidean distance geometry problem. One canonical spectral method for localizing points given their pairwise distances is the classical multidimensional scaling (Borg and Groenen, 2005). It takes the following form for model (1). Define a matrix $D \in \mathbb{R}^{d\times d}$ that stores noisy pairwise distances by

$$D_{ij} = D_{ji} = \frac{d(d-1)}{2N} \sum_{r=1}^{N} (y_r^2 - \sigma^2)\, \mathbb{1}\{x_r = e_i - e_j\}.$$

Since $\mathbb{E}[y_r^2 - \sigma^2 \mid x_r = e_i - e_j] = \mathbb{E}[(\theta_i^* - \theta_j^*)^2 + 2\epsilon_r z_r(\theta_i^* - \theta_j^*) + \epsilon_r^2 - \sigma^2 \mid x_r] = (\theta_i^* - \theta_j^*)^2$, we see that

$$\mathbb{E}[D_{ij}] = (\theta_i^* - \theta_j^*)^2.$$

In other words, $\mathbb{E}[D]$ stores the pairwise distances between the entries of θ^*. Let $J = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ where I is the identity matrix in $\mathbb{R}^{d\times d}$ and $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^d$. Using that $\mathbb{E}[D]$ is the pairwise distance matrix, it is not hard to verify that

$$-\frac{1}{2} J\, \mathbb{E}[D] J = \theta^* (\theta^*)^\top,$$

which is the Gram matrix of θ^*. Therefore, if (λ_1, v_1) is the leading eigenpair of the matrix $-\frac{1}{2}JDJ$, then we can use $\tilde{\theta} = \sqrt{\lambda_1}\, v_1$ as an estimator of θ^*.

Using standard spectral perturbation theory and random matrix theory, it is straightforward to analyze the error $\tilde{\theta} - \theta^*$ in either the ℓ_2 or ℓ_∞ norm (see, e.g., Chen et al. (2021)). We do not provide a theoretical analysis of $\tilde{\theta}$ in this work, but let us discuss it qualitatively. There are two sources of error in the spectral estimator $\tilde{\theta}$: the noise ϵ_r, which is inevitable, and the incompleteness of observations. In the regime $N < \binom{d}{2}$, if we never have $x_r = e_i - e_j$ for a pair of indices (i,j), then $D_{ij} = 0$ while $\mathbb{E}[D_{ij}] = (\theta_i^* - \theta_j^*)^2$. As a result, the spectral estimator $\tilde{\theta}$ does not recover θ^* even if $\sigma \to 0$. On the other hand, the limit of the EM algorithm $\hat{\theta}$ recovers θ^* as shown by Theorem 2 in this regime. Nevertheless, this does not mean that the spectral estimator is useless, because it can be used as an initialization of the EM algorithm. The experiment exhibited in the left plot of Figure 2 confirms these observations.

In the left plot of Figure 2, we set $d = 50$, $N = 1000$, and let the noise variance σ^2 vary from 0.002 to 2. Let us first focus on the performance of estimators in the small-noise regime. When $\sigma \to 0$, the squared error $\|\tilde{\theta} - \theta^*\|_2^2$ of the spectral method (dash-dotted red line) does not converge to zero. However, we can set $\theta^{(0)} = \tilde{\theta}$ and run the EM algorithm from this initialization for $T = 20$ steps. The squared error $\|\theta^{(T)} - \theta^*\|_2^2$ (solid blue line) goes to zero as $\sigma \to 0$. Recall that Theorem 2 predicts a sharp optimal rate (7) for the EM algorithm in the small-noise regime. We also plot this optimal error (dashed yellow line) and observe that it is very close to the actual error $\|\theta^{(T)} - \theta^*\|_2^2$

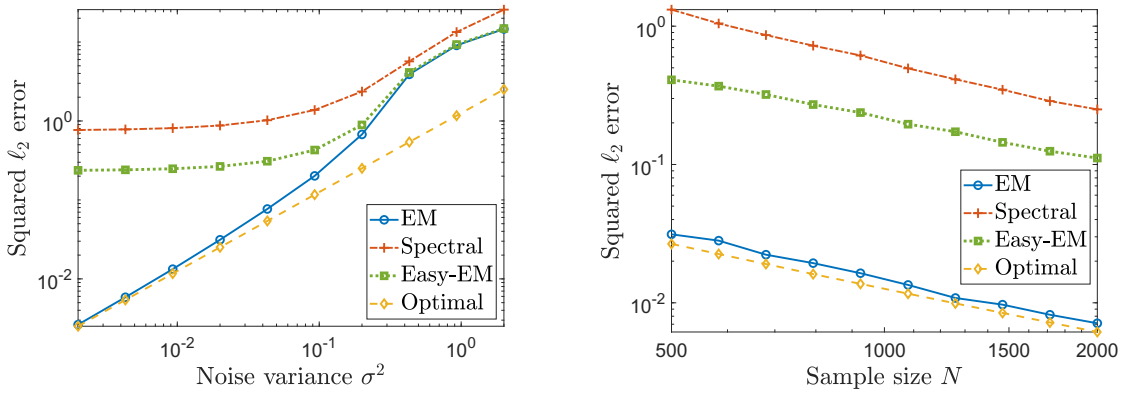

Figure 2. *Left:* Varying noise standard deviation. *Right:* Varying sample size. The log-log plots show the performance of different algorithms against the varying parameters. The squared error $\|\tilde{\theta} - \theta^*\|_2^2$ for the spectral estimator is shown in dash-dotted red lines. The EM and Easy-EM algorithms are initialized with $\theta^{(0)} = \tilde{\theta}$ and run for $T = 20$ steps. Their eventual squared errors $\|\theta^{(T)} - \theta^*\|_2^2$ are shown in solid blue lines and dotted green lines respectively. The experiment for each value of σ^2 or N is averaged over 100 independent repetitions. The theoretically predicted optimal rate is shown in dashed yellow lines.

as $\sigma \to 0$. In addition, we consider the Easy-EM iteration with $\theta^{(t+1)} = \bar{Q}(\theta^{(t)})$ where $\bar{Q}$ is defined in (4a). Compared to the EM iteration, Easy-EM does not have the multiplication by $\hat{\Sigma}^\dagger$ at each step, which turns out to be crucial: We observe that the Easy-EM algorithm (dotted green line) fails to achieve a small error as $\sigma \to 0$.

Next, we consider the regime of large noise in the left plot of Figure 2. The error of the EM algorithm no longer follows the sharp rate predicted by Theorem 2, which suggests that some condition on the noise may be necessary for this theorem (although the condition $\sigma \ll \frac{\beta}{(\log N)^2}$ we impose may not be optimal). Nevertheless, the EM and Easy-EM algorithms still improve upon the spectral initialization, and all the three algorithms appear to differ only by a constant factor from the optimal rate as σ^2 grows.

To further investigate the rate of estimation achieved by these algorithms, we consider another experiment shown in the right plot of Figure 2. We take $d = 50$, $\sigma = 0.1$, and let the sample size N vary from 500 to 2000. We again run the EM and the Easy-EM algorithm from the spectral initialization $\theta^{(0)} = \tilde{\theta}$ for $T = 20$ steps, and plot the squared errors. Given that the noise is small, the optimal rate from Theorem 2 accurately predicts the actual error of the EM algorithm. The spectral estimator and the Easy-EM algorithm clearly have worse performance, but their rates of estimation appear to be optimal as N grows.

4.3. Extensions of the model

There are several directions for extending our model. For example, one may consider a mixture of k linear regressions

$$y_r = x_r^\top \theta^{[\ell_r]} + \epsilon_r,$$

where we have $\theta^{[1]}, \dots, \theta^{[k]} \in \mathbb{R}^d$ and $(\ell_r)_{r=1}^N$ are i.i.d. from the distribution $\sum_{\ell=1}^k w_\ell \delta_\ell$ with weights satisfying $w_\ell \geq 0$ and $\sum_{\ell=1}^k w_\ell = 1$. When the covariates x_r are from a discrete distribution, even the identifiability of the mixture model is a nontrivial open problem.

To see the difficulty of the problem, let us consider a negative example. Suppose that we have a mixture of $k = 3$ equally weighted linear regressions, and the covariates are pairwise comparisons $x_r = e_{i_r} - e_{j_r}$ for $i_r, j_r \in [d]$. For simplicity, suppose that the noise ϵ_r is zero. Let

$$\theta^{[1]} = \begin{bmatrix} 1 \\ 2 \\ \theta_3 \\ \vdots \\ \theta_d \end{bmatrix}, \quad \theta^{[2]} = \begin{bmatrix} 3 \\ 3 \\ \theta_3 \\ \vdots \\ \theta_d \end{bmatrix}, \quad \theta^{[3]} = \begin{bmatrix} 2 \\ 4 \\ \theta_3 \\ \vdots \\ \theta_d \end{bmatrix}, \quad \tilde{\theta}^{[1]} = \begin{bmatrix} 2 \\ 2 \\ \theta_3 \\ \vdots \\ \theta_d \end{bmatrix}, \quad \tilde{\theta}^{[2]} = \begin{bmatrix} 1 \\ 3 \\ \theta_3 \\ \vdots \\ \theta_d \end{bmatrix}, \quad \tilde{\theta}^{[3]} = \begin{bmatrix} 3 \\ 4 \\ \theta_3 \\ \vdots \\ \theta_d \end{bmatrix},$$

where the unspecified entries $\theta_3, \dots, \theta_d$ can take any values. It is not hard to see that, even if we observe the sets

$$\left\{ \theta_i^{[\ell]} - \theta_j^{[\ell]} : \ell = 1, 2, 3 \right\} \quad \text{for all } (i, j) \in \binom{[d]}{2}$$

with no noise, the two mixtures $\{\theta^{[\ell]} : \ell = 1, 2, 3\}$ and $\{\tilde{\theta}^{[\ell]} : \ell = 1, 2, 3\}$ yield the same sets of observations. As a consequence, the mixture model with 3 components is not identifiable with the pairwise comparison design.

Therefore, to have an identifiable mixture model with multiple components, it is necessary to assume a richer design of covariates x_r. Instead of pairwise comparisons, one may consider comparisons between multiple objects as in the Plackett–Luce model (Luce, 1959; Plackett, 1975) or groups of pairwise comparisons used for mixtures of permutations (Mao and Wu, 2022). We leave this interesting topic to future research.

Beyond a mixture of linear regressions with additive noise, one may consider a mixture of generalized linear regressions (see, e.g., Heumann et al. (2008); Sun et al. (2014); Sedghi et al. (2016)), and in particular, the case of binary response is of practical interest in classification tasks. Similar to the linear setting, existing theories for mixtures of generalized linear models are often based on a Gaussian or continuous design. In the discrete setting, there have been a series of recent works on mixtures of the Placket–Luce or multinomial logit models motivated by ranking tasks (see, e.g., Zhao et al. (2016); Mollica and Tardella (2017); Chierichetti et al. (2018); Liu et al. (2019); Zhao et al. (2022)). In particular, Zhang et al. (2022) show that a mixture of two Bradley–Terry–Luce models (i.e., logistic regressions with the pairwise comparison design) are identifiable except when the parameters are in a set of measure zero. Very recently, Nguyen and Zhang (2023) provide a spectral initialization and an accurate implementation of the EM algorithm for learning a mixture of Plackett–Luce models. It would be interesting to analyze the convergence of iterative algorithms such as the EM algorithm for these models.

Along another direction, one may extend the model (2) to the noisy Euclidean distance geometry problem in dimension $m > 1$. Namely, we aim to estimate vectors $\theta_1^*, \dots, \theta_d^* \in \mathbb{R}^m$ based on incomplete, noisy observations of pairwise distances $\|\theta_i^* - \theta_j^*\|_2$. While this problem and its variants have long been studied (see the survey by Liberti et al. (2014)), progress in understanding the sample complexity has been made only in recent years using tools of matrix completion (Lai and Li, 2017; Tasissa and Lai, 2018), and the theory is even less complete in the noisy setting (Drusvyatskiy et al., 2017; Sremac et al., 2019). Compared to semidefinite programs used in many of these works, iterative algorithms are faster to run, but remain to be understood for this problem.

5. Proof techniques

We now informally discuss the strategies and key ingredients in the proofs of our main results; full proofs are provided in Section A.

5.1. Contraction in the ℓ_∞ norm

The first main result of this work is the linear convergence of the EM iteration in the ℓ_∞-norm in Theorem 1. To shed light on this result, let us first consider the zero-noise limit of the EM algorithm for simplicity. Suppose that the noise ϵ_r is zero in (1). Letting $\sigma \to 0$ in (5), we see that the EM iteration becomes

$$\theta^{(t+1)} = \left(\sum_{r=1}^N x_r x_r^\top \right)^\dagger \sum_{r=1}^N \mathsf{sign}(y_r \, x_r^\top \theta^{(t)}) \, y_r x_r \tag{8}$$

for $t \geq 0$, where $\mathsf{sign}(a) = 1$ if $a > 0$, $\mathsf{sign}(a) = 0$ if $a = 0$, and $\mathsf{sign}(a) = -1$ if $a < 0$. In fact, this is equivalent to the alternating minimization procedure where we iteratively compute

- $z_r^{(t)} := \min_{z \in \{-1,1\}} (y_r - z \, x_r^\top \theta^{(t)})^2 = \mathsf{sign}(y_r \, x_r^\top \theta^{(t)})$ for $r \in [N]$;
- $\theta^{(t+1)} := \min_{\theta \in \mathcal{H}} \sum_{r=1}^N (y_r - z_r^{(t)} x_r^\top \theta)^2 = \left(\sum_{r=1}^N x_r x_r^\top \right)^\dagger \sum_{r=1}^N z_r^{(t)} y_r x_r$.

We now explain why the iteration (8) contracts to θ^* locally. By slightly abusing the notation, we denote (only in this subsection) the noiseless Easy-EM and EM operators respectively by

$$\bar{Q}(\theta) = \frac{d-1}{2N} \sum_{r=1}^N \mathsf{sign}(y_r \, x_r^\top \theta) \, y_r x_r, \qquad \hat{Q}(\theta) = \hat{\Sigma}^\dagger \bar{Q}(\theta),$$

where the sample covariance $\hat{\Sigma}$ is defined in (3). We first note that θ^* is a fixed point of the operator $\hat{Q}$. Indeed, since $y_r = z_r x_r^\top \theta^*$, we have

$$\hat{Q}(\theta^*) = \left(\sum_{r=1}^N x_r x_r^\top \right)^\dagger \sum_{r=1}^N \mathsf{sign}\left(z_r (x_r^\top \theta^*)^2 \right) \, z_r (x_r^\top \theta^*) x_r = \left(\sum_{r=1}^N x_r x_r^\top \right)^\dagger \sum_{r=1}^N x_r x_r^\top \theta^* = \theta^*.$$

Next, fix $\theta \in \mathcal{H}$ in an ℓ_∞ neighborhood of θ^*. We have $\hat{Q}(\theta) - \theta^* = \hat{\Sigma}^\dagger \left(\bar{Q}(\theta) - \bar{Q}(\theta^*) \right)$, so

$$\|\hat{Q}(\theta) - \theta^*\|_\infty \leq \|\hat{\Sigma}^\dagger\|_\infty \left\| \bar{Q}(\theta) - \bar{Q}(\theta^*) \right\|_\infty, \tag{9}$$

where $\|\hat{\Sigma}^\dagger\|_\infty := \max_{v \in \mathcal{H}: \|v\|_\infty = 1} \|\hat{\Sigma}^\dagger v\|_\infty$. It remains to show, for example, that for a quantity $L > 0$, we have

$$\|\hat{\Sigma}^\dagger\|_\infty \leq L, \qquad \left\| \bar{Q}(\theta) - \bar{Q}(\theta^*) \right\|_\infty \leq \frac{1}{2L},$$

so that we have the desired contraction $\|\hat{Q}(\theta) - \theta^*\|_\infty \leq \frac{1}{2} \|\theta - \theta^*\|_\infty$.

In fact, bounding $\|\hat{\Sigma}^\dagger\|_\infty$ is one of the most important and technical steps of our proof. We discuss this step in more detail in Section 5.3 below and establish a bound formally in Proposition 11. For the remaining portion of the proof, we must bound $\left\| \bar{Q}(\theta) - \bar{Q}(\theta^*) \right\|_\infty$. Note that

$$\bar{Q}(\theta) - \bar{Q}(\theta^*) = \frac{d-1}{2N} \sum_{r=1}^N \Big[\mathsf{sign}\left(y_r x_r^\top \theta \right) - \mathsf{sign}\left(y_r x_r^\top \theta^* \right) \Big] y_r x_r. \tag{10}$$

If θ is in an ℓ_∞ neighborhood of θ^*, i.e., θ is close to θ^* entrywise, then $x_r^\top \theta = \theta_{i_r} - \theta_{j_r}$ is close to $x_r^\top \theta^* = \theta^*_{i_r} - \theta^*_{j_r}$. As a result, $\mathsf{sign}\left(y_r x_r^\top \theta\right)$ will coincide with $\mathsf{sign}\left(y_r x_r^\top \theta^*\right)$ for most $r \in [N]$, and a majority of terms in the above sum will vanish. Based on this intuition, we can perform an entrywise analysis of $\bar{Q}(\theta) - \bar{Q}(\theta^*)$ and obtain a bound on $\|\bar{Q}(\theta) - \bar{Q}(\theta^*)\|_\infty$. We omit the details.

In the noisy setting, the strategy for analyzing the EM iteration (5) is similar to the above noiseless analysis. However, there are complications due to the presence of noise. First, the fixed point of the EM operator $\hat{Q}$ is the random vector $\hat{\theta}$, not the deterministic vector θ^*. Therefore, we cannot directly show $Q(\hat{\theta}) = \hat{\theta}$; rather, it follows as a consequence of $\hat{Q}$ being a contraction locally around θ^*. Second, to prove that $\hat{Q}$ is contractive, we again reduce it to analyzing the Easy-EM operator $\bar{Q}$ via (9), but now for $\theta, \theta' \in \mathcal{H}$,

$$\bar{Q}(\theta) - \bar{Q}(\theta') = \frac{d-1}{2N} \sum_{r=1}^{N} \left[\tanh\left(\frac{y_r\, x_r^\top \theta}{\sigma^2}\right) - \tanh\left(\frac{y_r\, x_r^\top \theta'}{\sigma^2}\right) \right] y_r x_r.$$

As the function $\tanh(\cdot)$ is a soft approximation of the $\mathsf{sign}(\cdot)$ function in (10), the intuition for why we can control $\bar{Q}(\theta) - \bar{Q}(\theta')$ remains the same, but a more quantitative analysis is necessary. See Section A.4 for details.

5.2. Expansion around the ground truth

Our second main result is the sharp ℓ_2 bound for the EM fixed point $\hat{\theta}$ in Theorem 2. The strategy for proving this result consists in an expansion of $\hat{\theta}$ around the ground truth θ^*. Recall the definitions of $\bar{Q}$ and $\hat{Q}$ in (4). Since $\hat{\theta}$ is a fixed point of $\hat{Q}$, we have

$$\begin{aligned}
\hat{\theta} - \theta^* &= \hat{Q}(\hat{\theta}) - \hat{Q}(\theta^*) + \hat{Q}(\theta^*) - \theta^* \\
&= \hat{\Sigma}^\dagger \frac{d-1}{2N} \sum_{r=1}^{N} \left(\tanh\left(\frac{y_r x_r^\top \hat{\theta}}{\sigma^2}\right) - \tanh\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right) \right) y_r x_r + \hat{Q}(\theta^*) - \theta^*.
\end{aligned}$$

Applying Taylor's expansion of the function $\tanh(\cdot)$, we obtain

$$\hat{\theta} - \theta^* \approx \hat{\Sigma}^\dagger A(\hat{\theta} - \theta^*) + \hat{Q}(\theta^*) - \theta^*,$$

where $A := \frac{d-1}{2N} \sum_{r=1}^{N} \frac{y_r^2}{\sigma^2} \tanh'\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right) x_r x_r^\top$. By analyzing the matrix A, we then show that $\|\hat{\Sigma}^\dagger A(\hat{\theta} - \theta^*)\|_2 \ll \|\hat{\theta} - \theta^*\|_2$ and so

$$\|\hat{\theta} - \theta^*\|_2 \approx \|\hat{Q}(\theta^*) - \theta^*\|_2.$$

It remains to study how far the one-step EM iterate $\hat{Q}(\theta^*)$ moves away from the ground truth θ^*. Since $\hat{Q}(\theta^*)$ is an independent sum conditional on the covariates $x_1, \ldots, x_N$, and θ^* is deterministic, we can apply tools from matrix concentration to obtain the desired sharp bound

$$\|\hat{Q}(\theta^*) - \theta^*\|_2^2 \le (1 + o(1))\, \sigma^2 \frac{d-1}{2N}\, \mathsf{tr}(\hat{\Sigma}^\dagger).$$

5.3. Infinity operator norm of the inverse covariance

As discussed in Section 5.1 above, a key ingredient in our analysis of the EM iteration is an upper bound on $\|\hat{\Sigma}^\dagger\|_\infty := \max_{v\in\mathcal{H}:\|v\|_\infty=1} \|\hat{\Sigma}^\dagger v\|_\infty$. Recall that $\mathbb{E}[\hat{\Sigma}]$ is the identity map on $\mathcal{H}$ by the discussion after the definition of $\hat{\Sigma}$ in (3). Hence, it is plausible that $\hat{\Sigma}$ and $\hat{\Sigma}^\dagger$ have bounded norms. To prove $\|\hat{\Sigma}^\dagger\|_\infty \le L$ for a quantity $L > 0$, we start with the relation (see Lemma 6)

$$\|\hat{\Sigma}^\dagger\|_\infty = \Big(\min_{u\in\mathcal{H}:\|u\|_\infty=1} \|\hat{\Sigma}u\|_\infty\Big)^{-1}.$$

It therefore suffices to prove

$$\min_{u\in\mathcal{H}:\|u\|_\infty=1} \|\hat{\Sigma}u\|_\infty \ge 1/L. \tag{11}$$

The proof of (11) is the most technical step in our analysis and spans Lemmas 7, 8, 9, 10, and Proposition 11. In short, it is not sufficient to control $\|\hat{\Sigma}u\|_\infty$ for each fixed u and then apply a union bound. Instead, we carefully split the constraint set $\{u \in \mathcal{H} : \|u\|_\infty = 1\}$ into a union of subsets according to the sizes of entries $u_1, \dots, u_d$. On each subset consisting of vectors u, we use the specific properties of sizes of entries $u_1, \dots, u_d$ to bound $\|\hat{\Sigma}u\|_\infty$ simultaneously for all u in that subset. Finally, we combine all the subsets using a union bound. See Section A.2 for details.

While (11) arises as a technical ingredient in our work, we believe the analysis of the quantity $\min_{u\in\mathcal{H}:\|u\|_\infty=1} \|\hat{\Sigma}u\|_\infty$ is interesting in its own right in view of its connection to the literature. First, consider the graph G with vertex set $[d]$ and edge set $\{(i_r, j_r) : x_r = e_{i_r} - e_{j_r},\, r \in [N]\}$. Then the graph G resembles an Erdős–Rényi graph $G\big(d, N/\binom{d}{2}\big)$ except that the edges are sampled with replacement. Moreover, it is not hard to see that $\hat{\Sigma}$ is a rescaled version of the Laplacian matrix of the graph G, which is a well-researched object. In fact, if the ℓ_∞-norms in (11) were replaced by the ℓ_2-norms, then the quantity $\min_{u\in\mathcal{H}:\|u\|_2=1} \|\hat{\Sigma}u\|_2$ would be the second-smallest eigenvalue of the Laplacian matrix, known as the algebraic connectivity or the Fiedler value. This eigenvalue for an Erdős–Rényi graph has long been studied; see, e.g., Feige and Ofek (2005); Coja-Oghlan (2007); Chung and Radcliffe (2011); Kolokolnikov et al. (2014). However, we are not aware of an existing bound for the ℓ_∞-norm that would imply (11).

Moreover, if the minimization problem in (11) were over $u \in \{-1, 1\}^d$, then it would be related to the vast literature on discrepancy theory: $\min_{u\in\{-1,1\}^d} \|\hat{\Sigma}u\|_\infty$ is the combinatorial discrepancy of the matrix $\hat{\Sigma}$. We refer the interested reader to the celebrated work by Spencer (1985) which has generated extensive research on this topic, and see, e.g., Perkins and Xu (2021); Abbe et al. (2022); Altschuler (2022) for recent breakthroughs. The relaxation that we study is thus a natural object.

5.4. Proof organization

Finally, we overview how the proofs are organized in Section A. Recall that our results do not require a fresh sample for each step of the EM iteration. To achieve this, we first condition on a high-probability event globally, and then show that the EM operator is contractive deterministically on this event. More precisely, this high-probability event consists of two sets of conditions: Section A.1 proves that certain preliminary conditions hold for the observations $(x_r, y_r)_{r=1}^N$ with high probability (Lemma 3); Section A.2 mainly establishes a high probability bound on $\|\hat{\Sigma}^\dagger\|_\infty$ (Proposition 11) together with a few other conditions on $\hat{\Sigma}$. With these conditions assumed, Lemma 16 in Section A.4 shows that the EM operator is contractive locally around θ^*, and note that this lemma is completely deterministic. Finally, using an additional bound on $\|\hat{Q}(\theta^*) - \theta^*\|_\infty$ in Section A.3,

Proposition 17 in Section A.4 summarizes the convergence of the EM sequence and directly leads to Theorem 1.

For Theorem 2, we first need a sharp analysis of $\|\hat{Q}(\theta^*) - \theta^*\|_2$ in Section A.3: its expectation and concentration are controlled by Lemmas 14 and 15 respectively. Then, in Section A.5, we analyze the lower order terms and conclude with Proposition 20, which immediately implies Theorem 2.

Acknowledgments

We are grateful to the anonymous referees for helpful suggestions. Additionally, we acknowledge funding received by each author. AD was supported in part by NSF grant DMS-2053333. CM was supported in part by NSF grants DMS-2053333 and DMS-2210734. AP was supported in part by NSF grants CCF-2107455 and DMS-2210734, and by Research Awards from Adobe and Amazon. CM thanks Dylan J. Altschuler and Anderson Ye Zhang for helpful discussions.

References

Emmanuel Abbe, Shuangning Li, and Allan Sly. Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 327–338. IEEE, 2022.

Radoslaw Adamczak. A note on the Hanson–Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

Dylan J Altschuler. Critical window of the symmetric perceptron. *arXiv preprint arXiv:2205.02319*, 2022.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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

Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization: A Gaussian process perspective. *arXiv preprint arXiv:2109.09859*, 2021.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805, 2022.

Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via Fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202, 2013.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory*, 64(3):1738–1766, 2017.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Learning a mixture of two multinomial logits. In *International Conference on Machine Learning*, pages 961–969. PMLR, 2018.

Fan Chung and Mary Radcliffe. On the spectra of general random graphs. *the electronic journal of combinatorics*, pages P215–P215, 2011.

Amin Coja-Oghlan. On the Laplacian eigenvalues of G(n, p). *Combinatorics, Probability and Computing*, 16(6):923–946, 2007.

Mihai Cucuringu and Hemant Tyagi. An extension of the angular synchronization problem to the heterogeneous setting. *arXiv preprint arXiv:2012.14932*, 2020.

Yuval Dagan, Vardis Kandiros, and Constantinos Daskalakis. EM's convergence in Gaussian latent tree models. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2597–2667. PMLR, 02–05 Jul 2022.

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

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

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

Dmitriy Drusvyatskiy, Nathan Krislock, Y-L Voronin, and Henry Wolkowicz. Noisy euclidean distance realization: robust facial reduction and the pareto frontier. *SIAM Journal on Optimization*, 27(4):2301–2331, 2017.

Rishabh Dudeja, Yue M Lu, and Subhabrata Sen. Universality of approximate message passing with semi-random matrices. *arXiv preprint arXiv:2204.04281*, 2022.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan, and Bin Yu. Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48(6): 3161 – 3182, 2020.

Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Chao Gao and Anderson Y Zhang. Exact minimax estimation for phase synchronization. *IEEE Transactions on Information Theory*, 67(12):8236–8247, 2021.

John Clifford Gower. Euclidean distance geometry. *Math. Sci*, 7(1):1–14, 1982.

Christian Heumann, Bettina Grün, and Friedrich Leisch. Finite mixtures of generalized linear regression models. *Recent advances in linear models and related areas: essays in honour of helge toutenburg*, pages 205–230, 2008.

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

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65 (6):3515–3524, 2019.

Theodore Kolokolnikov, Braxton Osting, and James Von Brecht. Algebraic connectivity of Erdös–Rényi graphs near the connectivity threshold. *Manuscript in preparation*, 2014.

Jeongyeol Kwon and Constantine Caramanis. EM converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pages 1727–1736. PMLR, 2020.

Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pages 2055–2110. PMLR, 2019.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

Rongjie Lai and Jia Li. Solving partial differential equations on manifolds from incomplete interpoint distance. *SIAM Journal on Scientific Computing*, 39(5):A2231–A2256, 2017.

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

Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications. *SIAM review*, 56(1):3–69, 2014.

Ao Liu, Zhibing Zhao, Chao Liao, Pinyan Lu, and Lirong Xia. Learning Plackett–Luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4328–4335, 2019.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Wiley, 1959.

Cheng Mao and Yihong Wu. Learning mixtures of permutations: Groups of pairwise comparisons and combinatorial method of moments. *The Annals of Statistics*, 50(4):2231–2255, 2022.

Cristina Mollica and Luca Tardella. Bayesian Plackett–Luce mixture models for partially ranked data. *Psychometrika*, 82(2):442–458, 2017.

Duc Nguyen and Anderson Y Zhang. Efficient and accurate learning of mixtures of Plackett–Luce models. *arXiv preprint arXiv:2302.05343*, 2023.

Will Perkins and Changji Xu. Frozen 1-rsb structure of the symmetric ising perceptron. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1579–1588, 2021.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

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

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

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

Stefan Sremac, Fei Wang, Henry Wolkowicz, and Lucas Pettersson. Noisy Euclidean distance matrix completion with a single missing node. *Journal of Global Optimization*, 75:973–1002, 2019.

Yuekai Sun, Stratis Ioannidis, and Andrea Montanari. Learning mixtures of linear classifiers. In *International Conference on Machine Learning*, pages 721–729. PMLR, 2014.

Abiy Tasissa and Rongjie Lai. Exact reconstruction of Euclidean distance geometry problem using low-rank matrix completion. *IEEE Transactions on Information Theory*, 65(5):3124–3144, 2018.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12:315–330, 2002.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022.

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

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

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

Lei Xu, Michael Jordan, and Geoffrey E Hinton. An alternative model for mixtures of experts. *Advances in neural information processing systems*, 7, 1994.

Xiaomin Zhang, Xucheng Zhang, Po-Ling Loh, and Yingyu Liang. On the identifiability of mixtures of ranking models. *arXiv preprint arXiv:2201.13132*, 2022.

Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of Plackett–Luce models. In *International Conference on Machine Learning*, pages 2906–2914. PMLR, 2016.

Zhibing Zhao, Ao Liu, and Lirong Xia. Learning mixtures of random utility models with features from incomplete preferences. *arXiv preprint arXiv:2006.03869*, 2022.

Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

Appendix A. Analysis of the EM algorithm

Throughout the proof section, we use C, C_1, C', etc., to denote universal, positive constants that may change from line to line. Recall the definitions of $\bar{Q}$ and $\hat{Q}$ in (4) and that $y_r = z_r\, x_r^\top \theta^* + \epsilon_r$ where $z_r = \pm 1$. Since $\tanh(\cdot)$ is an odd function, we may assume without loss of generality that $z_r = 1$ for all $r = 1, \dots, N$, which does not change the EM operator. Therefore, we take the liberty of assuming that the observations take the form $y_r = x_r^\top \theta^* + \epsilon_r$ when analyzing the EM operator.

A.1. Preliminaries

For $i \in [d]$, define $\mathcal{R}_i$ to be the set of indices $r \in [N]$ for which the entry $(x_r)_i$ is non-zero, i.e.,

$$\mathcal{R}_i := \big\{r \in [N] : x_r = \pm(e_i - e_j),\, j \in [d] \setminus \{i\}\big\}, \tag{12}$$

where $\pm$ is $+$ if $i < j$ and is $-$ if $i > j$. In addition, for a fixed separation parameter $\Delta > 0$, define

$$S_i(\Delta) := \{j \in [d] \setminus \{i\} : |\theta_i^* - \theta_j^*| \le \Delta\} \quad \text{and} \tag{13a}$$

$$\mathcal{R}_i(\Delta) := \{r \in [N] : x_r = \pm(e_i - e_j),\, j \in S_i(\Delta)\}. \tag{13b}$$

In other words, $S_i(\Delta)$ is the set of indices j such that θ_j^* is close to θ_i^*, and $\mathcal{R}_i(\Delta)$ is a subset of $\mathcal{R}_i$ consisting of indices $r \in [N]$ for which x_r compares θ_i^* to a nearby parameter θ_j^*.

Lemma 3 *There is an absolute constant $C > 0$ such that the following holds for any $\delta \in (0, 0.1)$. Suppose that $N \ge Cd \log \frac{d}{\delta}$. It holds with probability at least $1 - \delta$ that, for all $i \in [d]$ and all $\Delta \ge \sigma$:*

- $\frac{1.9N}{d} \le |\mathcal{R}_i| \le \frac{2.1N}{d}$;
- $\sum_{r \in \mathcal{R}_i(\Delta)} |y_r| \le C\Delta \left(|S_i(\Delta)| \frac{N}{d^2} + \log \frac{d}{\delta}\right)$;
- $\sum_{r \in \mathcal{R}_i(\Delta)} y_r^2 \le C\Delta^2 \left(|S_i(\Delta)| \frac{N}{d^2} + \log \frac{d}{\delta}\right)$.

Proof For a fixed $i \in [d]$, each x_r is equal to $\pm(e_i - e_j)$ for some $j \in [d] \setminus \{i\}$ with probability $(d-1)/\binom{d}{2} = 2/d$. Hence $|\mathcal{R}_i| \sim \mathsf{Bin}\,(N, 2/d)$. It then follows from a binomial tail bound for any $\delta \in (0, 0.1)$ that

$$\left||\mathcal{R}_i| - \frac{2N}{d}\right| \le C_1 \sqrt{\frac{N}{d} \log \frac{d}{\delta}} + C_1 \log \frac{d}{\delta}$$

with probability at least $1 - \frac{\delta}{d}$ for an absolute constant $C_1 > 0$. Taking a union bound over $i \in [d]$ and using the condition $N \ge Cd \log \frac{d}{\delta}$, we obtain the desired bound on $|\mathcal{R}_i|$.

Similarly, $|\mathcal{R}_i(\Delta)| \sim \mathsf{Bin}\left(N, \frac{2|S_i(\Delta)|}{d(d-1)}\right)$ and

$$|\mathcal{R}_i(\Delta)| \le \frac{2N|S_i(\Delta)|}{d(d-1)} + C_2 \sqrt{\frac{N|S_i(\Delta)|}{d(d-1)} \log \frac{d}{\delta}} + C_2 \log \frac{d}{\delta} \le \frac{3N|S_i(\Delta)|}{d(d-1)} + C_3 \log \frac{d}{\delta}$$

with probability at least $1 - \frac{\delta}{d^2}$ for absolute constants $C_2, C_3 > 0$.

Condition on $\mathcal{R}_i(\Delta)$ henceforth. For $r \in \mathcal{R}_i(\Delta)$, we have $|y_r| = |x_r^\top \theta^* + \epsilon_r| \le \Delta + |\epsilon_r|$ and so $y_r^2 \le 2\Delta^2 + 2\epsilon_r^2$. Consequently, by a sub-Gaussian tail bound on $\sum_{r \in \mathcal{R}_i(\Delta)} |\epsilon_r|$, we obtain

$$\sum_{r \in \mathcal{R}_i(\Delta)} |y_r| \le \Delta\, |\mathcal{R}_i(\Delta)| + \sum_{r \in \mathcal{R}_i(\Delta)} |\epsilon_r| \le \Delta\, |\mathcal{R}_i(\Delta)| + \sqrt{\frac{2}{\pi}}\sigma\, |\mathcal{R}_i(\Delta)| + C_4\sigma \sqrt{|\mathcal{R}_i(\Delta)| \log \frac{d}{\delta}}$$

with probability at least $1 - \frac{\delta}{d^2}$ for an absolute constant $C_4 > 0$. Moreover, by a χ^2 tail bound on $\sum_{r \in \mathcal{R}_i(\Delta)} \epsilon_r^2$, we obtain

$$\sum_{r \in \mathcal{R}_i(\Delta)} y_r^2 \le 2\Delta^2\, |\mathcal{R}_i(\Delta)| + 2 \sum_{r \in \mathcal{R}_i(\Delta)} \epsilon_r^2 \le 2\Delta^2\, |\mathcal{R}_i(\Delta)| + 2\sigma^2\, |\mathcal{R}_i(\Delta)| + C_5\sigma^2 \Big(\sqrt{|\mathcal{R}_i(\Delta)| \log \frac{d}{\delta}} + \log \frac{d}{\delta}\Big)$$

with probability at least $1 - \frac{\delta}{d^2}$ for an absolute constant $C_5 > 0$.

Combining the above three bounds and using the condition $\Delta \geq \sigma$, we obtain

$$\sum_{r \in \mathcal{R}_i(\Delta)} |y_r| \leq 2\Delta \left(\frac{3N|S_i(\Delta)|}{d(d-1)} + C_3 \log \frac{d}{\delta} \right) + C_4 \sigma \sqrt{\left(\frac{3N|S_i(\Delta)|}{d(d-1)} + C_3 \log \frac{d}{\delta} \right) \log \frac{d}{\delta}}$$
$$\leq \frac{7N\Delta|S_i(\Delta)|}{d(d-1)} + C_6 \Delta \log \frac{d}{\delta}$$

for an absolute constant $C_6 > 0$ and

$$\sum_{r \in \mathcal{R}_i(\Delta)} y_r^2 \leq 4\Delta^2 \left(\frac{3N|S_i(\Delta)|}{d(d-1)} + C_3 \log \frac{d}{\delta} \right) + C_5 \sigma^2 \left(\sqrt{\left(\frac{3N|S_i(\Delta)|}{d(d-1)} + C_3 \log \frac{d}{\delta} \right) \log \frac{d}{\delta}} + \log \frac{d}{\delta} \right)$$
$$\leq \frac{13N\Delta^2|S_i(\Delta)|}{d(d-1)} + C_7 \Delta^2 \log \frac{d}{\delta}$$

for an absolute constant $C_7 > 0$.

Finally, note that the sets $S_i(\Delta)$ are nested as Δ varies, and $S_i(\Delta)$ can take at most $d-1$ values; hence, the same statements are true for $\mathcal{R}_i(\Delta)$. We can then take a union bound over all ($\leq d-1$) possibilities for $\mathcal{R}_i(\Delta)$ as Δ varies, together with a union bound over $i \in [d]$ to conclude. ■

A.2. Sample covariance

In this section, we study the sample covariance matrix (3). Recall that $\mathcal{H}$ denotes the orthogonal complement of $\mathbf{1}$ in $\mathbb{R}^d$, and the pseudoinverse $\hat{\Sigma}^\dagger$ is the inverse of $\hat{\Sigma}$ when viewed as a map on $\mathcal{H}$. We first bound the spectral norms of $\hat{\Sigma}$ and $\hat{\Sigma}^\dagger$.

Proposition 4 *There is an absolute constant $C > 0$ such that the following holds for any $\delta \in (0, 0.1)$. If $N \geq Cd \log \frac{d}{\delta}$, then with probability at least $1 - \delta$,*

- $\|\hat{\Sigma}\|_{op} \leq 3$*;*
- $\hat{\Sigma}$ *has $d-1$ nonzero eigenvalues;*
- $\|\hat{\Sigma}^\dagger\|_{op} \leq 5$.

Proof Since $\hat{\Sigma}$ is positive semidefinite, we have $\|\hat{\Sigma}\|_{op} = \lambda_{\max}(\hat{\Sigma})$. Let us first upper-bound $\lambda_{\max}(\hat{\Sigma})$. Let $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$. By the definition (3), we have

$$v^\top \hat{\Sigma} v = \frac{d-1}{2N} \sum_{r=1}^{N} (x_r^\top v)^2 = \frac{d-1}{2N} \sum_{r=1}^{N} (v_{i_r} - v_{j_r})^2 \leq \frac{d-1}{N} \sum_{r=1}^{N} (v_{i_r}^2 + v_{j_r}^2) = \frac{d-1}{N} \sum_{i=1}^{d} |\mathcal{R}_i| v_i^2,$$

where the inequality follows since $(a-b)^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$, and $\mathcal{R}_i$ is defined in (12). By the bound on $|\mathcal{R}_i|$ from Lemma 3, we conclude that with probability at least $1 - \delta$,

$$\|\hat{\Sigma}\|_{op} = \max_{v \in \mathbb{R}^d : \|v\|_2 = 1} v^\top \hat{\Sigma} v \leq \frac{d-1}{N} \cdot \frac{2.1N}{d} \leq 3.$$

Note since $\hat{\Sigma}^\dagger$ is also positive semidefinite, we have $\|\hat{\Sigma}^\dagger\|_{op} = \lambda_{\max}(\hat{\Sigma}^\dagger)$. Furthermore, we have

$$\lambda_{\max}(\hat{\Sigma}^\dagger) = 1/\lambda,$$

where λ is the minimum nonzero eigenvalue of $\hat{\Sigma}$. It suffices to lower-bound the smallest eigenvalue of $\hat{\Sigma}$ as a map on $\mathcal{H}$, and then this eigenvalue is precisely λ.

To this end, let us make a few definitions. First define the set

$$S := \{v \in \mathcal{H} \,:\, \|v\|_2 = 1\}.$$

Then define the random matrices

$$D \in \mathbb{R}^{d\times d}, \text{ such that } D_{ij} = \begin{cases} |\mathcal{R}_i| & i = j, \\ 0 & i \neq j, \end{cases}$$

$$A \in \mathbb{R}^{d\times d}, \text{ such that } A_{ij} = \begin{cases} \sum_{r=1}^N \mathbb{1}\{x_r = \pm(e_i - e_j)\} & i \neq j, \\ 0 & i = j, \end{cases}$$

$$M = \frac{N}{\binom{d}{2}}(J - I) - A,$$

where $J = \mathbf{1}\mathbf{1}^\top$. With these in hand, we have the following

$$\lambda = \min_{v\in S} v^\top \hat{\Sigma} v = \frac{d-1}{2N} \min_{v\in S} v^\top (D - A) v \geq \frac{d-1}{2N}\left(\min_{v\in S} v^\top D v - \max_{u\in S} u^\top A u\right). \tag{14}$$

Let us consider the first term. We have

$$\min_{v\in S} v^\top D v = \min_{v\in S} \sum_{i=1}^d |\mathcal{R}_i| v_i^2 \geq \min_{v\in S} \sum_{i=1}^d \frac{1.9N}{d} v_i^2 = \frac{1.9N}{d}, \tag{15}$$

where the inequality holds with probability at least $1 - \delta$ by Lemma 3.

Let us now consider the second term. Note that $(I - J)u = u$ for $u \in S$. By the definition of A, we have

$$\begin{aligned} \max_{u\in S} u^\top A u &= \max_{u\in S} u^\top (-M) u + \frac{N}{\binom{d}{2}} u^\top (I - J) u \\ &\leq \max_{u,v\in S} u^\top (-M) v + \frac{N}{\binom{d}{2}} \\ &\leq \|(-M)\|_{op} + \frac{N}{\binom{d}{2}} = \|M\|_{op} + \frac{N}{\binom{d}{2}}. \end{aligned} \tag{16}$$

It remains to bound $\|M\|_{op}$ using Lemma 24. To this end, let us define random matrices

$$X_r = \frac{1}{\binom{d}{2}}(J - I) - \sum_{1\leq i<j\leq d} \mathbb{1}\{x_r = \pm(e_i - e_j)\}(e_i e_j^\top + e_j e_i^\top),$$

and then $M = \sum_{r=1}^N X_r$. First, let us show $\mathbb{E}[X_r] = 0$. We have

$$\mathbb{E}[X_r] = \frac{1}{\binom{d}{2}}(J - I) - \sum_{1\leq i<j\leq d} \frac{1}{\binom{d}{2}}(e_i e_j^\top + e_j e_i^\top) = \frac{1}{\binom{d}{2}}\left(J - I - \sum_{i\neq j} e_i e_j^\top\right) = 0.$$

Next, we find a bound on $\|X_r\|_{op}$. Note that $\sum_{1\le i<j\le d} \mathbb{1}\{x_r = \pm(e_i - e_j)\} = 1$ for every $r \in [N]$. With this in mind, we have

$$\|X_r\|_{op} \le \frac{1}{\binom{d}{2}}(\|J\|_{op} + \|I\|_{op}) + \sum_{1\le i<j\le d} \mathbb{1}\{x_r = \pm(e_i - e_j)\} \|e_i e_j^\top + e_j e_i^\top\|_{op} = \frac{d+1}{\binom{d}{2}} + 1 \le 2.$$

Finally, we need a bound on $\|\sum_{r=1}^N \mathbb{E}[X_r^2]\|_{op}$. To this end, we compute

$$\begin{aligned} X_r^2 = \frac{1}{\binom{d}{2}^2}((d-2)J + I) &+ \sum_{1\le i<j\le d} \mathbb{1}\{x_r = \pm(e_i - e_j)\}(e_i e_i^\top + e_j e_j^\top) \\ &- \frac{1}{\binom{d}{2}} \sum_{1\le i<j\le d} \mathbb{1}\{x_r = \pm(e_i - e_j)\} \left(\mathbf{1}e_j^\top + \mathbf{1}e_i^\top + e_i\mathbf{1}^\top + e_j\mathbf{1}^\top - 2(e_i e_j^\top + e_j e_i^\top)\right), \end{aligned}$$

which implies

$$\begin{aligned} \mathbb{E}[X_r^2] &= \frac{1}{\binom{d}{2}^2}((d-2)J + I) + \frac{1}{\binom{d}{2}} \sum_{1\le i<j\le d} (e_i e_i^\top + e_j e_j^\top) \\ &\quad - \frac{1}{\binom{d}{2}^2} \sum_{1\le i<j\le d} \left(\mathbf{1}e_j^\top + \mathbf{1}e_i^\top + e_i\mathbf{1}^\top + e_j\mathbf{1}^\top - 2(e_i e_j^\top + e_j e_i^\top)\right) \\ &= \frac{1}{\binom{d}{2}^2}((d-2)J + I) + \frac{d-1}{\binom{d}{2}} I - \frac{2}{\binom{d}{2}^2}((d-2)J + I) \\ &= \frac{d-1}{\binom{d}{2}} I - \frac{1}{\binom{d}{2}^2}((d-2)J + I). \end{aligned}$$

Finally, we have

$$v(M) = \left\|N\,\mathbb{E}[X_1^2]\right\|_{op} \le N\left(\frac{2}{d} + \frac{4(d-1)^2}{d^2(d-1)^2}\right) \le \frac{4N}{d}.$$

By Lemma 24, we have that for some absolute constant $C' > 0$,

$$\|M\|_{op} \le C'\left(\sqrt{\frac{N}{d}\log\frac{d}{\delta}} + \log\frac{d}{\delta}\right),$$

with probability at least $1 - \delta$.

Putting the above together with (15) and (16), we have

$$\min_{v\in S} v^\top D v - \max_{u\in S} u^\top A u \ge \frac{1.9N}{d} - \frac{N}{\binom{d}{2}} - C'\left(\sqrt{\frac{N}{d}\log\frac{d}{\delta}} + \log\frac{d}{\delta}\right),$$

with probability at least $1 - 2\delta$. Since $d \ge 3$, combining this with (14) yields that for some absolute constant $C'' > 0$,

$$\lambda \ge 0.3 - C''\left(\sqrt{\frac{d}{N}\log\frac{d}{\delta}} + \frac{d}{N}\log\frac{d}{\delta}\right),$$

with probability at least $1 - 2\delta$. In particular, for $N \geq C\, d \log \frac{d}{\delta}$, we have $\lambda \geq 0.2$ and

$$\|\hat{\Sigma}^\dagger\|_{op} = 1/\lambda \leq 5$$

with probability at least $1 - 2\delta$. ■

Lemma 5 *There is an absolute constant $C > 0$ such that the following holds for any $\delta \in (0, 0.1)$. Suppose that $N \geq Cd \log \frac{d}{\delta}$. With probability at least $1 - \delta$, we have*

$$\mathsf{tr}(\hat{\Sigma}^\dagger) \geq (d-1)/3.$$

Proof Condition on the event where the bounds in Proposition 4 hold. Let $\lambda_1 \geq \cdots \geq \lambda_{d-1} > 0$ be the nonzero eigenvalues of $\hat{\Sigma}$. Then we have $\lambda_1 \leq 3$ and so $\mathsf{tr}(\hat{\Sigma}^\dagger) = \sum_{i=1}^{d-1} \frac{1}{\lambda_i} \geq \frac{d-1}{3}$. ■

For any linear map $A : \mathcal{H} \to \mathcal{H}$, define the norm

$$\|A\|_\infty := \max_{v \in \mathcal{H} : \|v\|_\infty = 1} \|Av\|_\infty.$$

We bound the norm $\|\hat{\Sigma}^\dagger\|_\infty$ in the rest of this section: starting with Lemmas 7, 8, 9, and 10, the bound is completed in Proposition 11.

Lemma 6 *Suppose that*

$$\min_{u \in \mathcal{H} : \|u\|_\infty = 1} \|\hat{\Sigma} u\|_\infty > 0.$$

Then we have

$$\|\hat{\Sigma}^\dagger\|_\infty = \left(\min_{u \in \mathcal{H} : \|u\|_\infty = 1} \|\hat{\Sigma} u\|_\infty \right)^{-1}.$$

Proof By the assumption, there is no vector $u \in \mathcal{H}$ such that $\hat{\Sigma} u = 0$, so $\hat{\Sigma}$ is invertible on $\mathcal{H}$. Then we have

$$\|\hat{\Sigma}^\dagger\|_\infty = \max_{v \in \mathcal{H} : \|v\|_\infty = 1} \|\hat{\Sigma}^\dagger v\|_\infty = \max_{u \in \mathcal{H} : \|\hat{\Sigma} u\|_\infty = 1} \|u\|_\infty = \max_{u \in \mathcal{H} : \|\hat{\Sigma} u\|_\infty > 0} \frac{\|u\|_\infty}{\|\hat{\Sigma} u\|_\infty} = \max_{u \in \mathcal{H} : \|u\|_\infty = 1} \frac{1}{\|\hat{\Sigma} u\|_\infty}$$

which yields the desired result. ■

Lemma 7 *There exists an absolute constant $C > 0$ such that the following holds for any fixed $\delta \in (0, 0.1)$, $\kappa \in (0, 1]$, and $u \in \mathcal{H}$ with $\|u\|_\infty = 1$. Define $I := \{i \in [d] : u_i \geq \kappa\}$. Suppose that $N|I| \geq d \log \frac{1}{\delta}$. With probability at least $1 - \delta$, we have*

$$\|\hat{\Sigma} u\|_\infty \geq \kappa - C \sqrt{\frac{d}{N|I|} \log \frac{1}{\delta}}.$$

Proof For any $i \in [d]$, we have

$$(\hat{\Sigma}u)_i = \sum_{j=1}^{d} \hat{\Sigma}_{ij} u_j = \sum_{j=1}^{d} \hat{\Sigma}_{ij}(u_j - u_i),$$

since $\sum_{j=1}^{d} \hat{\Sigma}_{ij} = (\hat{\Sigma}\mathbf{1})_i = 0$. Then, by the definition of $\hat{\Sigma}$,

$$(\hat{\Sigma}u)_i = \frac{d-1}{2N} \sum_{j \in [d]\setminus\{i\}} \sum_{r=1}^{N} \left(x_r x_r^\top\right)_{ij} (u_j - u_i) = \frac{d-1}{2N} \sum_{r=1}^{N} \sum_{j \in [d]\setminus\{i\}} (u_i - u_j) \mathbb{1}\{x_r = \pm(e_i - e_j)\}. \tag{17}$$

Then it holds that

$$\|\hat{\Sigma}u\|_\infty \geq \frac{1}{|I|} \sum_{i \in I} (\hat{\Sigma}u)_i = \frac{d-1}{2N|I|} \sum_{r=1}^{N} \sum_{i \in I} \sum_{j \in [d]\setminus\{i\}} (u_i - u_j) \mathbb{1}\{x_r = \pm(e_i - e_j)\}. \tag{18}$$

For any $r \in [N]$, define

$$X_r := \frac{d-1}{2} \sum_{i \in I} \sum_{j \in [d]\setminus\{i\}} (u_i - u_j) \mathbb{1}\{x_r = \pm(e_i - e_j)\}.$$

Then we have

$$\mathbb{E}[X_r] = \frac{d-1}{2} \sum_{i \in I} \sum_{j \in [d]\setminus\{i\}} (u_i - u_j) \frac{2}{d(d-1)} = \sum_{i \in I} \frac{d u_i - \sum_{j=1}^{d} u_j}{d} = \sum_{i \in I} u_i \in \left[\kappa\,|I|, |I|\right]$$

since $\sum_{j=1}^{d} u_j = 0$ and $\kappa \leq u_i \leq \|u\|_\infty = 1$ for $i \in I$. Moreover, it follows that

$$|X_r - \mathbb{E}[X_r]| \leq |X_r| + \mathbb{E}[X_r] \leq d \cdot 2\|u\|_\infty + |I| \leq 3d$$

and that

$$\mathsf{Var}(X_r) \leq \mathbb{E}[X_r^2] = \frac{(d-1)^2}{4} \sum_{i \in I} \sum_{j \in [d]\setminus\{i\}} (u_i - u_j)^2 \frac{2}{d(d-1)} \leq 2d\,|I|.$$

Since $X_1, \ldots, X_N$ are independent, Bernstein's inequality together with (18) implies that

$$\|\hat{\Sigma}u\|_\infty \geq \frac{1}{N|I|} \sum_{r=1}^{N} X_r \geq \kappa - C_1 \left(\sqrt{\frac{d}{N|I|} \log \frac{1}{\delta}} + \frac{d}{N|I|} \log \frac{1}{\delta} \right)$$

with probability at least $1 - \delta$ for an absolute constant $C_1 > 0$. This completes the proof since $N|I| \geq d \log \frac{1}{\delta}$ by assumption. ■

Lemma 8 *There exists an absolute constant $C > 0$ such that the following holds for any fixed $\delta \in (0, 0.1)$ and $\alpha, \kappa \in (0, 1]$. Suppose that $N \geq C \frac{d}{\alpha\kappa^2} \log \frac{2}{\kappa\delta}$. Define*

$$\mathcal{U} := \left\{ u \in \mathcal{H} : \|u\|_\infty = 1,\ |\{i \in [d] \,:\, u_i \geq \kappa\}| \geq \alpha d \right\}.$$

With probability at least $1 - \delta$, we have

$$\min_{u \in \mathcal{U}} \|\hat{\Sigma}u\|_\infty \geq \kappa/2.$$

Proof For $\epsilon \in (0,1)$, the set

$$\mathcal{N} := \{u \in \mathcal{U} : u_i = 0, \pm\epsilon, \pm 2\epsilon, \ldots, \pm\lfloor 1/\epsilon \rfloor \epsilon \text{ for all } i \in [d]\}$$

is clearly an ϵ-net of $\mathcal{U}$ in the ℓ_∞ norm, and it has cardinality $|\mathcal{N}| \le (2/\epsilon)^d$.

For any $u \in \mathcal{U}$, let $v \in \mathcal{N}$ be such that $\|u - v\|_\infty \le \epsilon$. Then we have

$$\|\hat{\Sigma} v\|_\infty = \|\hat{\Sigma}(v - u + u)\|_\infty \le \|\hat{\Sigma}\|_\infty \|v - u\|_\infty + \|\hat{\Sigma} u\|_\infty \le \epsilon \|\hat{\Sigma}\|_\infty + \|\hat{\Sigma} u\|_\infty. \tag{19}$$

By the definition of $\hat{\Sigma}$ and the fact that $\sum_{j=1}^d \hat{\Sigma}_{ij} = 0$ for any $i \in [d]$, we obtain

$$\begin{aligned}\|\hat{\Sigma}\|_\infty = \max_{i \in [d]} \sum_{j=1}^d |\hat{\Sigma}_{ij}| = 2 \max_{i \in [d]} \sum_{j \in [d] \setminus \{i\}} |\hat{\Sigma}_{ij}| &= \frac{d-1}{N} \max_{i \in [d]} \sum_{j \in [d] \setminus \{i\}} \left| \sum_{r=1}^N (x_r x_r^\top)_{ij} \right| \\ &= \frac{d-1}{N} \max_{i \in [d]} \sum_{j \in [d] \setminus \{i\}} \sum_{r=1}^N \mathbb{1}\{x_r = \pm(e_i - e_j)\}.\end{aligned}$$

It then follows from the definition of $\mathcal{R}_i$ in (12) and Lemma 3 that

$$\|\hat{\Sigma}\|_\infty = \frac{d-1}{N} \max_{i \in [d]} |\mathcal{R}_i| \le 3$$

with probability at least $1 - \delta/2$. Combining this bound with (19) gives

$$\min_{u \in \mathcal{U}} \|\hat{\Sigma} u\|_\infty \ge \min_{v \in \mathcal{N}} \|\hat{\Sigma} v\|_\infty - 3\epsilon.$$

Recall that $|\{i \in [d] \,:\, v_i \ge \kappa\}| \ge \alpha d$ for every $v \in \mathcal{N} \subset \mathcal{U}$. Let $\epsilon = \kappa/10$. We apply Lemma 7 with δ replaced by $\frac{\delta}{2(2/\epsilon)^d}$ and a union bound over all $v \in \mathcal{N}$ to obtain that, with probability at least $1 - \delta/2$,

$$\min_{v \in \mathcal{N}} \|\hat{\Sigma} v\|_\infty \ge \kappa - C_1 \sqrt{\frac{d}{N \alpha d} \log \frac{2(2/\epsilon)^d}{\delta}} \ge \kappa - C_2 \sqrt{\frac{d}{N\alpha} \log \frac{2}{\kappa \delta}}$$

for absolute constants $C_1, C_2 > 0$.

In view of the above two displays together with the condition $N \ge C \frac{d}{\alpha \kappa^2} \log \frac{2}{\kappa\delta}$ for a large constant $C > 0$, we conclude that $\min_{u \in \mathcal{U}} \|\hat{\Sigma} u\|_\infty \ge \kappa/2$ with probability at least $1 - \delta$. ■

Lemma 9 *There exists an absolute constant $C > 0$ such that the following holds for any fixed $\delta \in (0, 0.1)$, $0 \le \lambda < \kappa \le 1$, and nonempty subsets $I \subset J \subset [d]$. With probability at least $1 - \delta$, we have that*

$$\|\hat{\Sigma} u\|_\infty \ge (\kappa - \lambda) \frac{d - |J|}{d} - \frac{|I|}{d} - C\left(\sqrt{\frac{d}{N|I|} \log \frac{1}{\delta}} + \frac{d}{N|I|} \log \frac{1}{\delta} \right)$$

for all $u \in \mathcal{H}$ with $\|u\|_\infty = 1$, $\{i \in [d] : u_i \ge \kappa\} = I$, and $\{i \in [d] : u_i \ge \lambda\} = J$.

Proof By (17), we have

$$\begin{aligned}\|\hat{\Sigma}u\|_\infty \geq \frac{1}{|I|}\sum_{i\in I}(\hat{\Sigma}u)_i &= \frac{d-1}{2N|I|}\sum_{r=1}^{N}\sum_{i\in I}\sum_{j\in[d]\setminus\{i\}}(u_i-u_j)\mathbb{1}\{x_r=\pm(e_i-e_j)\}\\ &= \frac{d-1}{2N|I|}\sum_{r=1}^{N}\Bigg(\sum_{i\in I}\sum_{j\in I\setminus\{i\}}(u_i-u_j)\mathbb{1}\{x_r=\pm(e_i-e_j)\}\\ &\qquad +\sum_{i\in I}\sum_{j\in J\setminus I}(u_i-u_j)\mathbb{1}\{x_r=\pm(e_i-e_j)\}\\ &\qquad +\sum_{i\in I}\sum_{j\in[d]\setminus J}(u_i-u_j)\mathbb{1}\{x_r=\pm(e_i-e_j)\}\Bigg).\end{aligned}$$

In view of the assumptions $\|u\|_\infty = 1$, $\{i\in[d]: u_i\geq\kappa\} = I$, and $\{i\in[d]: u_i\geq\lambda\} = J$, we see that for $i\in I$,

$$u_i - u_j \geq \begin{cases} -1 & \text{if } j\in I\setminus\{i\},\\ 0 & \text{if } j\in J\setminus I,\\ \kappa-\lambda & \text{if } j\in[d]\setminus J.\end{cases}$$

It then follows that

$$\|\hat{\Sigma}u\|_\infty \geq -\frac{1}{N|I|}\sum_{r=1}^{N}\frac{d-1}{2}\sum_{i\in I}\sum_{j\in I\setminus\{i\}}\mathbb{1}\{x_r=\pm(e_i-e_j)\} \tag{20}$$

$$+\frac{\kappa-\lambda}{N|I|}\sum_{r=1}^{N}\frac{d-1}{2}\sum_{i\in I}\sum_{j\in[d]\setminus J}\mathbb{1}\{x_r=\pm(e_i-e_j)\}. \tag{21}$$

We now bound the above two terms. To control (20), we define

$$X_r := \frac{d-1}{2}\sum_{i\in I}\sum_{j\in I\setminus\{i\}}\mathbb{1}\{x_r=\pm(e_i-e_j)\}$$

for $r\in[N]$. Then we have

$$\mathbb{E}[X_r] = \frac{d-1}{2}\sum_{i\in I}\sum_{j\in I\setminus\{i\}}\frac{2}{d(d-1)} = \frac{|I|(|I|-1)}{d}\leq\frac{|I|^2}{d}, \qquad |X_r-\mathbb{E}[X_r]|\leq d,$$

and

$$\mathsf{Var}(X_r)\leq\mathbb{E}[X_r^2] = \frac{(d-1)^2}{4}\sum_{i\in I}\sum_{j\in I\setminus\{i\}}\frac{2}{d(d-1)}\leq\frac{|I|^2}{2}.$$

Since $X_1,\ldots,X_N$ are independent, Bernstein's inequality implies that

$$\sum_{r=1}^{N}X_r\leq N\frac{|I|^2}{d}+C_1\left(\sqrt{N|I|^2\log\frac{1}{\delta}}+d\log\frac{1}{\delta}\right)$$

with probability at least $1-\delta/2$ for an absolute constant $C_1 > 0$ and any $\delta \in (0, 0.1)$. Consequently, the term (20) can be bounded as

$$\frac{1}{N|I|}\sum_{r=1}^{N}\frac{d-1}{2}\sum_{i\in I}\sum_{j\in I\setminus\{i\}}\mathbb{1}\{x_r = \pm(e_i - e_j)\} \le \frac{|I|}{d} + C_1\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta}} + \frac{d}{N|I|}\log\frac{1}{\delta}\right).$$

Analogously, if we define

$$X'_r := \frac{d-1}{2}\sum_{i\in I}\sum_{j\in[d]\setminus J}\mathbb{1}\{x_r = \pm(e_i - e_j)\},$$

we can apply Bernstein's inequality to obtain that

$$\sum_{r=1}^{N} X'_r \ge N\frac{|I|(d-|J|)}{d} - C_2\left(\sqrt{N|I|(d-|J|)\log\frac{1}{\delta}} + d\log\frac{1}{\delta}\right)$$

with probability at least $1-\delta/2$ for an absolute constant $C_2 > 0$ and any $\delta \in (0, 0.1)$. Then the term (21) can be bounded as

$$\frac{\kappa-\lambda}{N|I|}\sum_{r=1}^{N}\frac{d-1}{2}\sum_{i\in I}\sum_{j\in[d]\setminus J}\mathbb{1}\{x_r = \pm(e_i - e_j)\} \ge (\kappa-\lambda)\frac{d-|J|}{d} - C_2\left(\sqrt{\frac{d}{N|I|}\log\frac{1}{\delta}} + \frac{d}{N|I|}\log\frac{1}{\delta}\right).$$

Plugging the above bounds into (20) and (21) respectively completes the proof. ■

Lemma 10 *There exists an absolute constant $C > 0$ such that the following holds for any fixed $\delta \in (0, 0.1)$ and $0 \le \lambda < \kappa \le 1$. For $u \in \mathbb{R}^d$, define*

$$\ell(u) := |\{i \in [d] : u_i \ge \kappa\}|, \qquad m(u) := |\{i \in [d] : u_i \ge \lambda\}|.$$

With probability at least $1-\delta$, we have that

$$\|\hat{\Sigma}u\|_\infty \ge \frac{\kappa-\lambda}{2}$$

for all $u \in \mathcal{H}$ with $\|u\|_\infty = 1$, $m(u) \le \frac{d}{4}$, $\ell(u) \le (\kappa-\lambda)\frac{d}{8}$, and $N \ge \frac{Cd\,m(u)}{(\kappa-\lambda)^2\ell(u)}\log\frac{d}{\delta}$.

Proof For fixed positive integers $\ell \le m \le d$, there are at most $\binom{d}{m}\binom{m}{\ell} \le d^m$ pairs of subsets (I, J) such that $|I| = \ell$, $|J| = m$, and $I \subset J \subset [d]$. Therefore, by Lemma 9 with δ replaced by δ/d^m and a union bound over all possible (I, J), we obtain the following: With probability at least $1-\delta$,

$$\|\hat{\Sigma}u\|_\infty \ge (\kappa-\lambda)\frac{d-m}{d} - \frac{\ell}{d} - C_1\left(\sqrt{\frac{dm}{N\ell}\log\frac{d}{\delta}} + \frac{dm}{N\ell}\log\frac{d}{\delta}\right) \tag{22}$$

for all $u \in \mathcal{H}$ with $\|u\|_\infty = 1$, $\ell(u) = \ell$, and $m(u) = m$, where $C_1 > 0$ is an absolute constant, and $\ell(u)$ and $m(u)$ are defined as in the lemma. Moreover, there are at most d^2 possible choices for the pair of integers (ℓ, m). Thus, by a further union bound over (ℓ, m), the bound (22) holds for all $u \in \mathcal{H}$ with $\|u\|_\infty = 1$, provided that the constant C_1 is replaced by a larger absolute constant. The conclusion then follows from the imposed conditions $m \le \frac{d}{4}$, $\ell \le (\kappa-\lambda)\frac{d}{8}$, and $N \ge \frac{Cd\,m}{(\kappa-\lambda)^2\ell}\log\frac{d}{\delta}$. ■

Proposition 11 *There exists an absolute constant $C > 0$ such that the following holds for any $\delta \in (0, 0.1)$. Fix an integer $L \geq 2$. Suppose that*

$$N \geq C \max\left\{L^2 d^{\frac{L}{L-1}} \log \frac{dL}{\delta}, L^3 d \log \frac{L}{\delta}\right\}.$$

Then it holds with probability at least $1 - \delta$ that,

$$\min_{u \in \mathcal{H}: \|u\|_\infty = 1} \|\hat{\Sigma} u\|_\infty \geq \frac{1}{2L}$$

and, as a result,

$$\|\hat{\Sigma}^\dagger\|_\infty \leq 2L.$$

Proof By Lemma 6, it suffices to prove the lower bound on $\|\hat{\Sigma} u\|_\infty$ for $u \in \mathcal{H}$ with $\|u\|_\infty = 1$. For $t = 0, 1, \ldots, L$, define

$$\kappa_t := 1 - t/L, \qquad \ell_t(u) := |\{i \in [d] : u_i \geq \kappa_t\}|.$$

For $t = 1, 2, \ldots, L-1$, define

$$\mathcal{U}_t := \left\{u \in \mathcal{H} : \|u\|_\infty = 1,\, \ell_t(u) \leq \frac{d}{4},\, \ell_{t-1}(u) \leq \frac{d}{8L},\, N \geq \frac{C_1 d L^2 \ell_t(u)}{\ell_{t-1}(u)} \log \frac{dL}{\delta}\right\},$$

where $C_1 > 0$ is a constant to be chosen. In addition, define

$$\mathcal{U}_L := \left\{u \in \mathcal{H} : \|u\|_\infty = 1,\, \ell_{L-1}(u) \geq \frac{d}{8L}\right\}.$$

We claim that

$$\bigcup_{t=1}^{L} \mathcal{U}_t = \left\{u \in \mathcal{H} : \|u\|_\infty = 1,\, \max_{i \in [d]} u_i = 1\right\}. \tag{23}$$

We first finish the proof based on the above claim. For $u \in \bigcup_{t=1}^{L-1} \mathcal{U}_t$, we apply Lemma 10 with $\kappa = \kappa_{t-1}$, $\lambda = \kappa_t$, $\ell(u) = \ell_{t-1}(u)$, $m(u) = \ell_t(u)$, and δ replaced by $\frac{\delta}{2L}$, together with a union bound over $t \in [L-1]$. Note that $\kappa - \lambda = 1/L$. Hence we obtain that, with probability at least $1 - \delta/2$,

$$\min_{u \in \bigcup_{t=1}^{L-1} \mathcal{U}_t} \|\hat{\Sigma} u\|_\infty \geq \frac{1}{2L},$$

provided that C_1 is a sufficiently large absolute constant. Moreover, for $u \in \mathcal{U}_L$, we apply Lemma 8 with $\kappa = K_{L-1} = 1/L$ and $\alpha = 1/(8L)$ to obtain that, with probability at least $1 - \delta/2$,

$$\min_{u \in \mathcal{U}_L} \|\hat{\Sigma} u\|_\infty \geq \frac{1}{2L},$$

provided that $N \geq C_2 L^3 d \log \frac{L}{\delta}$ for a large constant $C_2 > 0$.

The above two bounds together with (23) imply that, with probability at least $1 - \delta$, we have $\|\hat{\Sigma} u\|_\infty \geq 1/(2L)$ for all $u \in \mathcal{H}$ with $\|u\|_\infty = 1$ and $\max_{i \in [d]} u_i = 1$. Finally, for $u \in \mathcal{H}$ with $\|u\|_\infty = 1$ and $\min_{i \in [d]} u_i = -1$, it suffices to apply the bound to $-u$. ■

Proof [Proof of Claim (23)] Consider $u \in \mathcal{H}$ with $\|u\|_\infty = 1$ and $\max_{i\in[d]} u_i = 1$. Suppose that $u \notin \bigcup_{t=1}^{L-1} \mathcal{U}_t$. Then for any $t \in [L-1]$, we have $\ell_t(u) > \frac{d}{4}$, $\ell_{t-1}(u) > \frac{d}{8L}$, or $N < \frac{C_1 dL^2 \ell_t(u)}{\ell_{t-1}(u)} \log \frac{dL}{\delta}$. If either of the first two conditions holds, then we have $\ell_t(u) > \frac{d}{8L}$ for some $t \in \{0, 1, \dots, L-1\}$. Since $\kappa_{L-1} \le \kappa_t$, by the definition of $\ell_t(u)$, we have $\ell_{L-1}(u) \ge \ell_t(u) > \frac{d}{8L}$. Thus $u \in \mathcal{U}_L$. Let us further suppose that $u \notin \mathcal{U}_L$. Then, for all $t \in [L-1]$, we must have

$$N < \frac{C_1 dL^2 \ell_t(u)}{\ell_{t-1}(u)} \log \frac{dL}{\delta} \qquad \Longleftrightarrow \qquad \ell_t(u) > \ell_{t-1}(u) \cdot \frac{N}{C_1 L^2 d \log(dL/\delta)}.$$

Since $\kappa_0 = 1$ and $\ell_0(u) = |\{i \in [d] : u_i = 1\}| \ge 1$, we apply the above bound iteratively to obtain

$$\ell_{L-1}(u) \ge \left(\frac{N}{C_1 L^2 d \log(dL/\delta)} \right)^{L-1} \ge d > \frac{d}{8L},$$

where the second inequality holds by the assumption $N \ge CL^2 d^{\frac{L}{L-1}} \log \frac{dL}{\delta}$. As a result, $u \in \mathcal{U}_L$, finishing the proof. ■

A.3. One-step iterate from the ground truth

In this subsection, we study the difference between the ground truth θ^* and the one-step EM iterate $\hat{Q}(\theta^*)$ from it. All the results in this subsection are conditional on a realization of $x_1, \dots, x_N$. For brevity, we often take the liberty of using $\mathbb{E}$ and $\mathbb{P}$ to denote the conditional expectation and probability on $x_1, \dots, x_N$ respectively. Let us start with a lemma.

Lemma 12 *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then we have*

$$\mathbb{E}\left[X \tanh\left(\frac{\mu X}{\sigma^2} \right) \right] = \mu, \tag{24}$$

and

$$\mathsf{Var}\left(X \tanh\left(\frac{\mu X}{\sigma^2} \right) \right) \le \sigma^2. \tag{25}$$

Proof Equation (24) is well-known, but we provide a proof for completeness. By a change of variable $\nu = \frac{\mu}{\sigma}$ and $Z = \frac{X}{\sigma} \sim \mathcal{N}(\nu, 1)$, it suffices to show that

$$\mathbb{E}\left[Z \tanh(\nu Z) \right] = \nu. \tag{26}$$

We have $\mathbb{E}[Z] = \nu$ and

$$\mathbb{E}\left[Z \left(\tanh(\nu Z) - 1 \right) \right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z \frac{-2e^{-\nu z}}{e^{\nu z} + e^{-\nu z}} e^{-(z-\nu)^2/2} \, dz = -\frac{2e^{-\nu^2/2}}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{z e^{-z^2/2}}{e^{\nu z} + e^{-\nu z}} \, dz = 0, \tag{27}$$

because the integrand is an odd function. Then (26) immediately follows.

Next, it holds that

$$\mathsf{Var}\left(X \tanh(\mu X/\sigma^2) \right) = \sigma^2 \, \mathsf{Var}\left(Z \tanh(\nu Z) \right) = \sigma^2 \, \mathbb{E}\left[\left(Z \tanh(\nu Z) - \nu \right)^2 \right]. \tag{28}$$

Furthermore, we have

$$\begin{aligned}(Z\tanh(\nu Z)-\nu)^2 &= (Z-\nu)^2 - 2vZ(\tanh(\nu Z)-1) + Z^2(\tanh(\nu Z)^2-1)\\ &\le (Z-\nu)^2 - 2vZ(\tanh(\nu Z)-1) + Z^2\tanh(\nu Z)\,(\tanh(\nu Z)-1)\end{aligned}$$

since $\tanh(\nu Z)\le 1$. Let us compute the expectations of these three terms. We have

$$\mathbb{E}[(Z-\nu)^2]=1, \qquad \mathbb{E}[2vZ(\tanh(\nu Z)-1)]=0,$$

by $Z\sim\mathcal{N}(\nu,1)$ and (27) respectively. For the third term,

$$\begin{aligned}\mathbb{E}[Z^2\tanh(\nu Z)\,(\tanh(\nu Z)-1)] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} z^2\frac{(e^{\nu z}-e^{-\nu z})(-2e^{-\nu z})}{(e^{\nu z}+e^{-\nu z})^2}e^{-(z-\nu)^2/2}\,dz\\ &= \frac{-2e^{-v^2/2}}{\sqrt{2\pi}}\int_{\mathbb{R}} z^2 e^{-z^2/2}\frac{e^{\nu z}-e^{-\nu z}}{(e^{\nu z}+e^{-\nu z})^2}\,dz=0,\end{aligned}$$

because the integrand is an odd function. Combining the three terms with (28) proves (25). ■

We now control $\hat{Q}(\theta^*)-\theta^*$ in the ℓ_∞-norm.

Lemma 13 *Condition on a realization of $x_1,\dots,x_N$ such that $\|\hat{\Sigma}^\dagger\|_\infty\le 2L$. There is an absolute constant $C>0$ such that for any $\delta\in(0,0.1)$, it holds with (conditional) probability at least $1-\delta$ that*

$$\|\hat{Q}(\theta^*)-\theta^*\|_\infty \le C\sigma\sqrt{\frac{L\,d}{N}\log\frac{d}{\delta}}.$$

Proof Let us first check that $\mathbb{E}[\hat{Q}(\theta^*)]=\theta^*$. By (24), we have

$$\mathbb{E}_{y_r\sim\mathcal{N}(x_r^\top\theta^*,\sigma^2)}\left[y_r\tanh\left(\frac{y_r x_r^\top\theta^*}{\sigma^2}\right)\right]=x_r^\top\theta^*. \tag{29}$$

Then, it follows that

$$\mathbb{E}[\bar{Q}(\theta^*)]=\frac{d-1}{2N}\sum_{r=1}^N x_r x_r^\top\theta^*=\hat{\Sigma}\,\theta^*, \qquad \mathbb{E}[\hat{Q}(\theta^*)]=\theta^*. \tag{30}$$

Next, we study $\hat{Q}(\theta^*)-\theta^*$, whose ith entry is

$$e_i^\top\big(\hat{Q}(\theta^*)-\theta^*\big)=e_i^\top\hat{\Sigma}^\dagger\big(\bar{Q}(\theta^*)-\mathbb{E}[\bar{Q}(\theta^*)]\big). \tag{31}$$

For $w\in\mathcal{H}$, define

$$f_w(y):=w^\top\bar{Q}(\theta^*)=\frac{d-1}{2N}\sum_{r=1}^N y_r x_r^\top w\tanh\left(\frac{y_r x_r^\top\theta^*}{\sigma^2}\right). \tag{32}$$

We will show that $f_w(y)-\mathbb{E}_y[f_w(y)]$ is sub-Gaussian with variance parameter of order $\frac{\sigma^2 L\,d}{N}$ when $w^\top$ is a row of $\hat{\Sigma}^\dagger$. The proof is similar to that of Proposition 11 of Kwon et al. (2019), with the main tool being Lemma 25, a standard concentration result for Gaussian random variables. To this

end, let $v_r = \frac{y_r - x_r^\top \theta^*}{\sigma} = \frac{\epsilon_r}{\sigma}$ be i.i.d. standard Gaussians and introduce another independent set of i.i.d. standard Gaussians $\zeta_1, \ldots, \zeta_N$. Let us define a new function

$$g_w(v) := f_w(y) = \frac{d-1}{2N} \sum_{r=1}^{N} (\sigma v_r + x_r^\top \theta^*) x_r^\top w \tanh\left(\frac{(\sigma v_r + x_r^\top \theta^*) x_r^\top \theta^*}{\sigma^2} \right).$$

For each $r \in [N]$, we have

$$\begin{aligned}
(\nabla g_w(v))_r &= \frac{d-1}{2N} x_r^\top w \Bigg[\sigma \tanh\left(\frac{(\sigma v_r + x_r^\top \theta^*) x_r^\top \theta^*}{\sigma^2} \right) \\
&\qquad + (\sigma v_r + x_r^\top \theta^*) \frac{x_r^\top \theta^*}{\sigma} \tanh'\left(\frac{(\sigma v_r + x_r^\top \theta^*) x_r^\top \theta^*}{\sigma^2} \right) \Bigg] \\
&= \frac{d-1}{2N} x_r^\top w \sigma \cdot h\left(\frac{(\sigma v_r + x_r^\top \theta^*) x_r^\top \theta^*}{\sigma^2} \right),
\end{aligned}$$

where $h(t) := \tanh(t) + t \cdot \tanh'(t)$. We will use the inequality $|h(t)| \le 2$ for all $t \in \mathbb{R}$. By Lemma 25,

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda(f_w(y) - \mathbb{E}[f_w(y)]))] = \mathbb{E}[\exp(\lambda(g_w(v) - \mathbb{E}[g_w(v)]))] \\
&\le \mathbb{E}_{v,\zeta}\left[\exp\left(\frac{\lambda \pi}{2} \langle \nabla g, \zeta \rangle \right) \right] \\
&= \mathbb{E}_v \mathbb{E}_\zeta \left[\exp\left(\frac{\lambda \pi}{2} \left(\frac{d-1}{2N} \sum_{r=1}^{N} \zeta_r x_r^\top w \sigma \cdot h\left(\frac{(\sigma v_r + x_r^\top \theta^*) x_r^\top \theta^*}{\sigma^2} \right) \right) \right) \right] \\
&= \mathbb{E}_v \left[\exp\left(\left(\frac{\lambda^2 \sigma^2 \pi^2 (d-1)}{16N} \right) \left(\frac{d-1}{2N} \sum_{r=1}^{N} \left(x_r^\top w \cdot h((\sigma v_r + x_r^\top \theta^*) x_r^\top \theta^*) \right)^2 \right) \right) \right] \\
&\le \exp\left(\left(\frac{\lambda^2 \sigma^2 \pi^2 (d-1)}{4N} \right) \left(\frac{d-1}{2N} \sum_{r=1}^{N} (x_r^\top w)^2 \right) \right) \\
&\le \exp\left(\frac{\lambda^2 \sigma^2 \pi^2 d}{4N} w^\top \hat{\Sigma} w \right).
\end{aligned}$$

If $w^\top$ is the ith row of $\hat{\Sigma}^\dagger$, we have $w = \hat{\Sigma}^\dagger e_i$ and

$$w^\top \hat{\Sigma} w = e_i^\top \hat{\Sigma}^\dagger e_i \le \|\hat{\Sigma}^\dagger\|_\infty \le 2L$$

by assumption. It follows that

$$\mathbb{E}[\exp(\lambda(f_w(y) - \mathbb{E}[f_w(y)]))] \le \exp\left(\frac{\lambda^2 \sigma^2 \pi^2 d L}{2N} \right),$$

so $f_w(y)$ is sub-Gaussian with variance parameter $\frac{\sigma^2 \pi^2 d L}{N}$. Therefore, it holds with probability at least $1 - \delta/d$ that

$$\left| f_w(y) - \mathbb{E}[f_w(y)] \right| \le C\sigma \sqrt{\frac{Ld}{N} \log \frac{d}{\delta}}$$

for a constant $C > 0$. In view of (31) and (32), the left-hand side of the above bound is precisely $|e_i^\top(\hat{Q}(\theta^*) - \theta^*)|$ when $w^\top$ is the ith row of $\hat{\Sigma}^\dagger$. Taking a union bound over $i \in [d]$ proves the desired result on $\|\hat{Q}(\theta^*) - \theta^*\|_\infty$. ■

The following two lemmas provide a sharp control on $\hat{Q}(\theta^*) - \theta^*$ in the ℓ_2-norm: they respectively bound the expectation and deviation of $\|\hat{Q}(\theta^*) - \theta^*\|_2^2$.

Lemma 14 *Condition on a realization of $x_1, \dots, x_N$ such that $\hat{\Sigma}^\dagger$ is the inverse of $\hat{\Sigma}$ as a map on $\mathcal{H}$. Let $\mathbb{E}[\cdot]$ denote the conditional expectation. Then we have*

$$\mathbb{E}\big[\|\hat{Q}(\theta^*) - \theta^*\|_2^2\big] \le \sigma^2 \frac{d-1}{2N} \mathsf{tr}(\hat{\Sigma}^\dagger).$$

Proof By the definition of $\hat{Q}$ in (4b), it is not difficult to see that

$$\hat{Q}(\theta^*) - \theta^* = \bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger \sum_{r=1}^N \bigg(\tanh\Big(\frac{y_r\, x_r^\top \theta^*}{\sigma^2}\Big) y_r - x_r^\top \theta^*\bigg) x_r.$$

Then it follows that

$$\begin{aligned}
&\mathbb{E}\left[(\hat{Q}(\theta^*) - \theta^*)(\hat{Q}(\theta^*) - \theta^*)^\top\right] \\
&= \bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger \bigg(\sum_{r=1}^N \mathbb{E}\left[\bigg(\tanh\Big(\frac{y_r\, x_r^\top \theta^*}{\sigma^2}\Big) y_r - x_r^\top \theta^*\bigg)^2\right] x_r x_r^\top\bigg) \bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger,
\end{aligned}$$

where the cross terms vanish thanks to (29). By (25), we have

$$\sigma_r := \mathbb{E}\left[\bigg(\tanh\Big(\frac{y_r\, x_r^\top \theta^*}{\sigma^2}\Big) y_r - x_r^\top \theta^*\bigg)^2\right] \le \sigma^2.$$

As a result, the matrix

$$\sigma^2\bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger - \mathbb{E}\left[(\hat{Q}(\theta^*) - \theta^*)(\hat{Q}(\theta^*) - \theta^*)^\top\right] = \bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger \bigg(\sum_{r=1}^N (\sigma^2 - \sigma_r^2) x_r x_r^\top\bigg) \bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger$$

is positive semi-definite. In other words,

$$\mathbb{E}\left[(\hat{Q}(\theta^*) - \theta^*)(\hat{Q}(\theta^*) - \theta^*)^\top\right] \preccurlyeq \sigma^2 \bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger$$

in the Loewner order. Since the trace operator is monotone in the Loewner order, we obtain

$$\mathbb{E}\big[\|\hat{Q}(\theta^*) - \theta^*\|_2^2\big] = \mathsf{tr}\left(\mathbb{E}\left[(\hat{Q}(\theta^*) - \theta^*)(\hat{Q}(\theta^*) - \theta^*)^\top\right]\right) \le \sigma^2\, \mathsf{tr}\left(\bigg(\sum_{r=1}^N x_r x_r^\top\bigg)^\dagger\right),$$

which completes the proof. ■

Lemma 15 *Condition on a realization of $x_1, \dots, x_N$ such that $\|\hat{\Sigma}\|_{op} \le 3$ and $\|\hat{\Sigma}^\dagger\|_{op} \le 5$. Then there exists an absolute constant $C > 0$ such that for any $\delta \in (0, 0.1)$, we have*

$$\left| \|\hat{Q}(\theta^*) - \theta^*\|_2^2 - \mathbb{E}\big[\|\hat{Q}(\theta^*) - \theta^*\|_2^2\big] \right| \le C\sigma^2 \left(\frac{d^{3/2}}{N} \sqrt{\log \frac{1}{\delta}} + \frac{d}{N} \log \frac{1}{\delta} \right)$$

with (conditional) probability at least $1 - \delta$.

Proof Let $\epsilon' := \epsilon/\sigma \sim \mathcal{N}(0, I_N)$, and then $y_r = x_r^\top \theta^* + \sigma\epsilon'_r$. Consider the random vector

$$f(\epsilon') := \hat{Q}(\theta^*) - \theta^* = \hat{\Sigma}^\dagger \left(\frac{d-1}{2N} \sum_{r=1}^N \tanh\left(\frac{(x_r^\top \theta^* + \sigma\epsilon'_r) x_r^\top \theta^*}{\sigma^2} \right) (x_r^\top \theta^* + \sigma\epsilon'_r) x_r \right) - \theta^*.$$

We claim that the function $f(\cdot)$ is $15\sigma\sqrt{d/N}$-Lipschitz in the Euclidean distance.

Assuming the claim for a moment, we now show that the random vector $f(\epsilon')$ satisfies the convex concentration property in Definition 22. To this end, let ϕ be an arbitrary 1-Lipschitz function. (Note that we do not require ϕ to be convex and hence prove something stronger.) Then $\phi \circ f$ is $15\sigma\sqrt{d/N}$-Lipschitz, so by Lemma 21,

$$\mathbb{P}\left\{ \left| \phi(f(\epsilon')) - \mathbb{E}[\phi(f(\epsilon'))] \right| \ge t \right\} \le 2 \exp\left(\frac{-t^2}{C_1^2 \sigma^2 d/N} \right)$$

for an absolute constant $C_1 > 0$. Therefore, $f(\epsilon')$ satisfies the convex concentration property for $K = C_1\, \sigma \sqrt{d/N}$. We have seen in (30) that $\mathbb{E}[f(\epsilon')] = 0$. Then Lemma 23 with $A = I_d$ yields that for an absolute constant $C_2 > 0$,

$$\Pr\left\{ \left| \|\hat{Q}(\theta^*) - \theta^*\|_2^2 - \mathbb{E}[\|\hat{Q}(\theta^*) - \theta^*\|_2^2] \right| \ge t \right\} \le 2 \exp\left(-\frac{1}{C_2} \min\left(\frac{N^2 t^2}{\sigma^4 d^3}, \frac{Nt}{\sigma^2 d} \right) \right).$$

From here, the lemma follows.

It remains to prove the claim that the function $f(\cdot)$ is $15\sigma\sqrt{d/N}$-Lipschitz. Let $\epsilon', \epsilon'' \in \mathbb{R}^N$. By Taylor's theorem, we have for some $t_r \in (0, 1)$,

$$\begin{aligned}
&\tanh\left(\frac{(x_r^\top \theta^* + \sigma\epsilon''_r) x_r^\top \theta^*}{\sigma^2} \right) (x_r^\top \theta^* + \sigma\epsilon''_r) - \tanh\left(\frac{(x_r^\top \theta^* + \sigma\epsilon'_r) x_r^\top \theta^*}{\sigma^2} \right) (x_r^\top \theta^* + \sigma\epsilon'_r) \\
&= (\epsilon''_r - \epsilon'_r) \frac{d}{d\tilde{\epsilon}} \left(\tanh\left(\frac{(x_r^\top \theta^* + \sigma\tilde{\epsilon}) x_r^\top \theta^*}{\sigma^2} \right) (x_r^\top \theta^* + \sigma\tilde{\epsilon}) \right) \Big|_{\tilde{\epsilon} = \epsilon'_r + t_r(\epsilon''_r - \epsilon'_r)} \\
&= (\epsilon''_r - \epsilon'_r)\, \sigma\, h\left(\frac{(x_r^\top \theta^* + \sigma\tilde{\epsilon}_r) x_r^\top \theta^*}{\sigma^2} \right),
\end{aligned}$$

where $h(t) := \tanh(t) + t \tanh'(t)$ for $t \in \mathbb{R}$, and $\tilde{\epsilon}_r := \epsilon'_r + t_r(\epsilon''_r - \epsilon'_r)$. Then we have

$$f(\epsilon'') - f(\epsilon') = \hat{\Sigma}^\dagger \left(\frac{d-1}{2N} \sum_{r=1}^N (\epsilon''_r - \epsilon'_r)\, \sigma\, h\left(\frac{(x_r^\top \theta^* + \sigma\tilde{\epsilon}_r) x_r^\top \theta^*}{\sigma^2} \right) x_r \right) = \sigma \hat{\Sigma}^\dagger \tilde{X} D (\epsilon'' - \epsilon'),$$

where $\tilde{X} \in \mathbb{R}^{d \times N}$ is the matrix whose rth column is $\frac{d-1}{2N}\, x_r$, and $D \in \mathbb{R}^{N \times N}$ is the diagonal matrix defined by

$$D_{rr} := h\left(\frac{(y_r + t_r(y_r - y_r)) x_r^\top \theta^*}{\sigma^2} \right).$$

As a result,

$$\|f(\epsilon'') - f(\epsilon')\|_2 \le \sigma \|\hat{\Sigma}^\dagger\|_{op} \|\tilde{X}\|_{op} \|D\|_{op} \|\epsilon' - \epsilon''\|_2.$$

Note that $\tilde{X}\tilde{X}^\top = \frac{d-1}{2N}\hat{\Sigma}$ and so we have

$$\|\tilde{X}\|_{op} = \sqrt{\frac{d-1}{2N}\|\hat{\Sigma}\|_{op}} \le \sqrt{\frac{3d}{2N}}$$

by the conditioning. Moreover, we have $\|\hat{\Sigma}^\dagger\|_{op} \le 5$ by the conditioning, and $\|D\|_{op} \le 2$ by the fact that $|h(t)| = |\tanh(t) + t\tanh'(t)| \le 2$ for all $t \in \mathbb{R}$. Combining all the bounds gives

$$\|f(\epsilon'') - f(\epsilon')\|_2 \le 15\,\sigma \sqrt{\frac{d}{N}}\,\|\epsilon'' - \epsilon'\|_2,$$

completing the proof. ■

A.4. Convergence analysis

Assuming the bounds in Lemma 3 and the control on $\|\hat{\Sigma}^\dagger\|_\infty$ in Proposition 11, the following result shows that the EM operator is contractive locally around θ^*, deterministically.

Lemma 16 *There exist absolute constants $C, C' > 0$ such that the following holds. Suppose that all the bounds in Lemma 3 hold with $C > 0$ and $\delta \in (0, 0.1)$. Consider $\theta, \theta' \in \mathcal{H}$, $\lambda \in (0,1)$, and $\Delta := \max\{\frac{1}{\lambda}\|\theta - \theta'\|_\infty, \frac{3\sigma}{\sqrt{\lambda}}\}$, such that*

$$\max\{\|\theta - \theta^*\|_\infty,\, \|\theta' - \theta^*\|_\infty\} \le \frac{\Delta}{6}, \qquad N \ge C'\frac{d}{\lambda^2}\log\frac{d}{\delta}, \qquad \max_{i\in[d]}|S_i(\Delta)| \le \frac{\lambda^2}{C'}d,$$

where $S_i(\cdot)$ is defined as in (13a). Then we have

$$\|\bar{Q}(\theta) - \bar{Q}(\theta')\|_\infty \le \lambda\|\theta - \theta'\|_\infty.$$

If, in addition, $\|\hat{\Sigma}^\dagger\|_\infty \le \frac{1}{2\lambda}$, then

$$\|\hat{Q}(\theta) - \hat{Q}(\theta')\|_\infty \le \frac{1}{2}\|\theta - \theta'\|_\infty.$$

Proof For notational simplicity, let us consider iteration of the first entry of each vector and bound $|\bar{Q}(\theta)_1 - \bar{Q}(\theta')_1|$. The same analysis applies to any other entry with straightforward modification. We can write the iteration of the first entry as

$$\bar{Q}(\theta)_1 = \frac{d-1}{2N}\sum_{r\in\mathcal{R}_1}\tanh\Big(\frac{y_r\,x_r^\top\theta}{\sigma^2}\Big)y_r,$$

where $\mathcal{R}_1$ is defined as in (12). It follows that

$$|\bar{Q}(\theta)_1 - \bar{Q}(\theta')_1| \le \frac{d-1}{2N}\sum_{r\in\mathcal{R}_1}\left|\tanh\Big(\frac{y_r\,x_r^\top\theta}{\sigma^2}\Big) - \tanh\Big(\frac{y_r\,x_r^\top\theta'}{\sigma^2}\Big)\right||y_r|.$$

Define $\theta(z) := \theta' + z(\theta - \theta')$ for $z \in [0,1]$. By the fundamental theorem of calculus, we have

$$\begin{aligned}\left|\tanh\left(\frac{y_r\, x_r^\top \theta}{\sigma^2}\right) - \tanh\left(\frac{y_r\, x_r^\top \theta'}{\sigma^2}\right)\right| &= \left|\int_0^1 \tanh'\left(\frac{y_r\, x_r^\top \theta(z)}{\sigma^2}\right)\left(\frac{y_r\, x_r^\top (\theta - \theta')}{\sigma^2}\right) dz\right| \\ &\le \frac{|y_r\, x_r^\top (\theta - \theta')|}{\sigma^2}\int_0^1 4\exp\left(-2\frac{|y_r\, x_r^\top \theta(z)|}{\sigma^2}\right) dz \\ &\le 4\frac{|y_r\, x_r^\top (\theta - \theta')|}{\sigma^2}\exp\left(-2\min_{0\le z\le 1}\frac{|y_r\, x_r^\top \theta(z)|}{\sigma^2}\right),\end{aligned}$$

where we used the fact $0 \le \tanh'(t) = (\cosh(t))^{-2} \le 4\exp(-2|t|)$ for all $t \in \mathbb{R}$. Combining the above bounds with the trivial bound $|\tanh(t)| \le 1$ for $t \in \mathbb{R}$, we obtain

$$|\bar{Q}(\theta)_1 - \bar{Q}(\theta')_1| \le \frac{d}{N}\sum_{r\in\mathcal{R}_1} \min\left\{|y_r|,\, 2\frac{y_r^2\,|x_r^\top(\theta-\theta')|}{\sigma^2}\exp\left(-2\min_{0\le z\le 1}\frac{|y_r\, x_r^\top\theta(z)|}{\sigma^2}\right)\right\}. \quad (33)$$

For brevity, we let $\eta := \|\theta - \theta'\|_\infty$ in the sequel. Then $\Delta = \max\{\frac{\eta}{\lambda}, \frac{3\sigma}{\sqrt{\lambda}}\}$ by definition. Recall that the sets $S_1(\Delta)$ and $\mathcal{R}_1(\Delta)$ are defined in (13a) and (13b) respectively. We now analyze the summands in (33) for $r \in \mathcal{R}_1(\Delta)$ and $r \in \mathcal{R}_1 \setminus \mathcal{R}_1(\Delta)$ separately.

First, for $r \in \mathcal{R}_1 \setminus \mathcal{R}_1(\Delta)$, we have $x_r = e_1 - e_j$ such that $|x_r^\top \theta^*| = |\theta_1^* - \theta_j^*| > \Delta$. Moreover, $\|\theta - \theta^*\|_\infty \le \Delta/6$ and $\|\theta' - \theta^*\|_\infty \le \Delta/6$ by assumption. It follows that

$$\begin{aligned}|x_r^\top\theta(z)| = |\theta(z)_1 - \theta(z)_j| &\ge |\theta_1^* - \theta_j^*| - |\theta(z)_1 - \theta_1^*| - |\theta(z)_j - \theta_j^*| \\ &\ge |\theta_1^* - \theta_j^*| - z|\theta_1 - \theta_1^*| - (1-z)|\theta_1' - \theta_1^*| - z|\theta_j - \theta_j^*| - (1-z)|\theta_j' - \theta_j^*| \\ &\ge |x_r^\top\theta^*| - \frac{\Delta}{3} > \frac{2}{3}|x_r^\top\theta^*|\end{aligned}$$

for any $z \in [0,1]$. Moreover, it holds that $|x_r^\top(\theta - \theta')| \le 2\|\theta - \theta'\|_\infty = 2\eta$. Therefore,

$$2\frac{y_r^2\,|x_r^\top(\theta-\theta')|}{\sigma^2}\exp\left(-2\min_{0\le z\le 1}\frac{|y_r\, x_r^\top\theta(z)|}{\sigma^2}\right) \le 4\eta\frac{y_r^2}{\sigma^2}\exp\left(\frac{-4\,|y_r|\,|x_r^\top\theta^*|}{3\sigma^2}\right).$$

It is not hard to see that $\max_{a\ge 0} a^2\exp(-ab) = 4/(eb)^2$ for $b > 0$. Taking $a = |y_r|$ and $b = \frac{4\,|x_r^\top\theta^*|}{3\sigma^2}$ in the above bound then yields

$$2\frac{y_r^2\,|x_r^\top(\theta-\theta')|}{\sigma^2}\exp\left(-2\min_{0\le z\le 1}\frac{|y_r\, x_r^\top\theta(z)|}{\sigma^2}\right) \le \frac{9\sigma^2}{e^2\,|x_r^\top\theta^*|^2}\eta \le \left(\frac{3\sigma}{e\Delta}\right)^2\eta.$$

As a result,

$$\frac{d}{N}\sum_{r\in\mathcal{R}_1\setminus\mathcal{R}_1(\Delta)} 2\frac{y_r^2\,|x_r^\top(\theta-\theta')|}{\sigma^2}\exp\left(-2\min_{0\le z\le 1}\frac{|y_r\, x_r^\top\theta(z)|}{\sigma^2}\right) \le \frac{d}{N}|\mathcal{R}_1|\left(\frac{3\sigma}{e\Delta}\right)^2\eta \le \frac{\lambda}{2}\eta \quad (34)$$

by the bound $|\mathcal{R}_1| \le 3N/d$ in Lemma 3 and the condition $\Delta \ge \frac{3\sigma}{\sqrt{\lambda}}$.

Second, for $r \in \mathcal{R}_1(\Delta)$, since $\exp(-t) \le 1$ for $t \ge 0$ and $|x_r^\top(\theta - \theta')| \le 2\eta$, we have

$$\frac{d}{N} \sum_{r \in \mathcal{R}_1(\Delta)} \min\left\{ |y_r|,\, 2\frac{y_r^2\,|x_r^\top(\theta-\theta')|}{\sigma^2} \exp\left(-2 \min_{0\le z\le 1} \frac{|y_r\, x_r^\top \theta(z)|}{\sigma^2}\right)\right\}$$
$$\le \frac{d}{N} \sum_{r \in \mathcal{R}_1(\Delta)} \min\left\{ |y_r|,\, 4\eta\frac{y_r^2}{\sigma^2}\right\} \le \frac{d}{N} \min\left\{ \sum_{r \in \mathcal{R}_1(\Delta)} |y_r|,\, \frac{4\eta}{\sigma^2} \sum_{r \in \mathcal{R}_1(\Delta)} y_r^2\right\}.$$

Plugging the bounds from Lemma 3 into the above then yields

$$\frac{d}{N} \sum_{r \in \mathcal{R}_1(\Delta)} \min\left\{ |y_r|,\, 2\frac{y_r^2\,|x_r^\top(\theta-\theta^*)|}{\sigma^2} \exp\left(-2 \min_{0\le z\le 1} \frac{|y_r\, x_r^\top \theta(z)|}{\sigma^2}\right)\right\}$$
$$\le C\frac{d}{N}\left(|S_1(\Delta)|\frac{N}{d^2} + \log\frac{d}{\delta}\right) \min\left\{\Delta,\, \frac{4\eta}{\sigma^2}\Delta^2\right\}. \tag{35}$$

Lastly, combining (34) and (35) with (33) gives

$$|\bar{Q}(\theta)_1 - \bar{Q}(\theta')_1| \le \frac{\lambda}{2}\eta + C\left(\frac{|S_1(\Delta)|}{d} + \frac{d}{N}\log\frac{d}{\delta}\right) \min\left\{\Delta,\, \frac{4\eta}{\sigma^2}\Delta^2\right\}.$$

Since $\Delta = \max\{\frac{\eta}{\lambda}, \frac{3\sigma}{\sqrt{\lambda}}\}$, we have

$$\min\left\{\Delta,\, \frac{4\eta}{\sigma^2}\Delta^2\right\} \le \max\left\{\frac{\eta}{\lambda},\, \frac{4\eta}{\sigma^2}\left(\frac{3\sigma}{\sqrt{\lambda}}\right)^2\right\} \le 36\frac{\eta}{\lambda}.$$

The above two bounds together with the assumptions $N \ge C'\frac{d}{\lambda^2}\log\frac{d}{\delta}$ and $|S_1(\Delta)| \le \frac{\lambda^2}{C'}d$ yield

$$|\bar{Q}(\theta)_1 - \bar{Q}(\theta')_1| \le \frac{\lambda}{2}\eta + C\left(\frac{\lambda^2}{C'} + \frac{\lambda^2}{C'}\right)\cdot 64\frac{\eta}{\lambda} \le \lambda\eta,$$

once C' is chosen to be sufficiently large.

For the final statement of the lemma, note that $\hat{Q}(\theta) = \hat{\Sigma}^\dagger \bar{Q}(\theta)$, so we have

$$\|\hat{Q}(\theta) - \hat{Q}(\theta')\|_\infty \le \|\hat{\Sigma}^\dagger\|_\infty \|\bar{Q}(\theta) - \bar{Q}(\theta')\|_\infty,$$

from which the conclusion follows. ■

We summarize the convergence results for the EM sequence in the following proposition.

Proposition 17 *There exist absolute constants $C, C' > 0$ such that the following holds for any $\delta \in (0, 0.1)$ and any integer $L \ge 2$. Fix $\theta^{(0)} \in \mathcal{H}$ and let $\Delta^{(0)} := \max\{4L\|\theta^{(0)} - \theta^*\|_\infty, 6\sqrt{L}\,\sigma\}$. Suppose that*

$$N \ge C\max\left\{L^2 d^{\frac{L}{L-1}}\log\frac{dL}{\delta}, L^3 d\log\frac{L}{\delta}\right\}, \tag{36a}$$

$$\max_{i\in[d]} |S_i(\Delta^{(0)})| \le \frac{d}{CL^2}, \tag{36b}$$

where $S_i(\cdot)$ is defined in (13a). Let $\{\theta^{(t)}\}_{t\geq 0}$ be the EM iterates defined in (5). Moreover, let

$$\tau := C'\sigma\sqrt{\frac{L\,d}{N}\log\frac{d}{\delta}}, \qquad T := \max\left\{0, \left\lceil \log_{4/3}\left(\frac{\|\theta^{(0)}-\theta^*\|_\infty}{4\tau}\right)\right\rceil\right\}.$$

Then it holds with probability at least $1-\delta$ that

- *there exists $\hat{\theta}\in\mathcal{H}$ such that $\hat{Q}(\hat{\theta})=\hat{\theta}$ and $\|\hat{\theta}-\theta^*\|_\infty \leq 4\tau$;*
- *the sequence $\{\theta^{(t)}\}_{t\geq 0}$ converges to $\hat{\theta}$;*
- *$\|\theta^{(T+t)}-\hat{\theta}\,\|_\infty \leq 8\tau/2^t$ for all $t\geq 0$.*

Proof By Proposition 11 and the assumption on N, it holds with probability at least $1-\delta/3$ that

$$\|\hat{\Sigma}^\dagger\|_\infty \leq 2L.$$

By Lemma 13, it holds with probability at least $1-\delta/3$ that

$$\|\hat{Q}(\theta^*)-\theta^*\|_\infty \leq C'\sigma\sqrt{\frac{L\,d}{N}\log\frac{d}{\delta}} = \tau \tag{37}$$

for a constant $C'>0$. Applying Lemma 3 and Lemma 16 with $\lambda=\frac{1}{4L}$, we see that with probability at least $1-\delta/3$, the following holds: For all $\theta,\theta'\in\mathcal{H}$ and $\Delta := \max\{4L\|\theta-\theta'\|_\infty, 6\sqrt{L}\,\sigma\}$ such that

$$\max\{\|\theta-\theta^*\|_\infty,\, \|\theta'-\theta^*\|_\infty\} \leq \frac{\Delta}{6}, \qquad \max_{i\in[d]}|S_i(\Delta)| \leq \frac{d}{CL^2}, \tag{38}$$

we have

$$\|\hat{Q}(\theta)-\hat{Q}(\theta')\|_\infty \leq \frac{1}{2}\|\theta-\theta'\|_\infty. \tag{39}$$

In the sequel, we condition on the event of probability at least $1-\delta$ that all the above bounds hold. We split the rest of the proof into two parts as we will use the above conclusion twice.

Part I: Convergence to a neighborhood of the ground truth. Let us take $\theta=\theta^{(t)}$ for $t\geq 0$ and $\theta'=\theta^*$ in (38) and (39). Let $\Delta^{(t)} := \max\{4L\|\theta^{(t)}-\theta^*\|_\infty, 6\sqrt{L}\,\sigma\}$. Note that we have $\|\theta^{(t)}-\theta^*\|_\infty \leq \frac{4L}{6}\|\theta^{(t)}-\theta^*\|_\infty \leq \frac{\Delta^{(t)}}{6}$, so the first condition in (38) holds. We now show inductively that, for $t\geq 0$,

$$\Delta^{(t)} \leq \Delta^{(0)}, \tag{40}$$

and furthermore,

$$\|\theta^{(t+1)}-\theta^*\|_\infty \leq \frac{1}{2}\|\theta^{(t)}-\theta^*\|_\infty + \tau. \tag{41}$$

First, (40) is trivial for $t=0$. Suppose that (40) holds. Then, by $\max_{i\in[d]}|S_i(\Delta^{(0)})| \leq \frac{d}{CL^2}$ and that $S_i(\cdot)$ is a nondecreasing function defined in (13a), we see that $\max_{i\in[d]}|S_i(\Delta^{(t)})| \leq \frac{d}{CL^2}$. Hence (38) is satisfied, and then (39) implies

$$\|\theta^{(t+1)}-\hat{Q}(\theta^*)\|_\infty \leq \frac{1}{2}\|\theta^{(t)}-\theta^*\|_\infty.$$

Combining this bound with (37) gives (41).

Next, suppose that (41) holds. If $\|\theta^{(t)} - \theta^*\|_\infty \geq 2\tau$, then (41) implies that $\|\theta^{(t+1)} - \theta^*\|_\infty \leq \|\theta^{(t)} - \theta^*\|_\infty$. As a result, $\Delta^{(t+1)} \leq \Delta^{(t)}$, so (40) holds for $t+1$ in place of t. On the other hand, if $\|\theta^{(t)} - \theta^*\|_\infty \leq 2\tau$, then (41) implies that $\|\theta^{(t+1)} - \theta^*\|_\infty \leq 2\tau$. By the definition of τ and our assumption on N, it is clear that $4L\|\theta^{(t+1)} - \theta^*\|_\infty \leq 8L\tau \leq 6\sqrt{L}\,\sigma$. As a result, $\Delta^{(t+1)} \leq 6\sqrt{L}\,\sigma \leq \Delta^{(0)}$, so again (40) holds for $t+1$ in place of t. This completes the induction.

To conclude this part, note that by (41), the EM iterates $\{\theta^{(t)}\}_{t\geq 0}$ satisfy

$$\|\theta^{(t+1)} - \theta^*\|_\infty \leq \frac{3}{4}\|\theta^{(t)} - \theta^*\|_\infty \tag{42}$$

if $\|\theta^{(t)} - \theta^*\|_\infty > 4\tau$. Moreover, define

$$\mathcal{B} := \{\theta \in \mathcal{H} : \|\theta - \theta^*\|_\infty \leq 4\tau\}, \qquad T_1 := \min\big\{t \geq 0 : \theta^{(t)} \in \mathcal{B}\big\}.$$

By (42), we have

$$T_1 \leq \max\left\{0, \left\lceil \log_{4/3}\left(\frac{\|\theta^{(0)} - \theta^*\|_\infty}{4\tau}\right)\right\rceil\right\} = T.$$

Finally, $\theta^{(t)} \in \mathcal{B}$ for all $t \geq T_1$ by (41).

Part II: Convergence to a fixed point. We now focus on the set $\mathcal{B}$ defined above. Fix any $\theta, \theta' \in \mathcal{B}$. We have $\|\theta - \theta'\|_\infty \leq 8\tau$. By the definition of τ and our assumption on N, it holds that $4L\|\theta - \theta'\|_\infty \leq 16L\tau \leq 6\sqrt{L}\,\sigma$. Hence, $\Delta = \max\{4L\|\theta - \theta'\|_\infty, 6\sqrt{L}\,\sigma\} = 6\sqrt{L}\,\sigma \leq \Delta^{(0)}$. By our assumption on $\Delta^{(0)}$ and that $S_i(\cdot)$ is a nondecreasing function, we obtain $\max_{i\in[d]}|S_i(\Delta)| \leq \frac{d}{CL^2}$. In addition, $\max\{\|\theta - \theta^*\|_\infty, \|\theta' - \theta^*\|_\infty\} \leq 4\tau \leq \sqrt{L}\,\sigma = \frac{\Delta}{6}$. Therefore, (38) is satisfied, and so (39) holds.

Similar to Part I, taking $\theta' = \theta^*$ in (39), we obtain

$$\|\hat{Q}(\theta) - \theta^*\|_\infty \leq \|\hat{Q}(\theta) - \hat{Q}(\theta^*)\|_\infty + \|\hat{Q}(\theta^*) - \theta^*\|_\infty \leq \frac{1}{2}\|\theta - \theta^*\|_\infty + \tau \leq 4\tau,$$

so the EM operator $\hat{Q}$ can be viewed as a map on $\mathcal{B}$. Furthermore, (39) says that $\hat{Q}$ is a contraction on $\mathcal{B}$. By the Banach fixed-point theorem, $\hat{Q}$ has a unique fixed point $\hat{\theta}$ in $\mathcal{B}$, and the EM sequence $\{\theta^{(t)}\}_{t\geq 0}$ converges to $\hat{\theta}$ with

$$\|\theta^{(t+1)} - \hat{\theta}\,\|_\infty \leq \frac{1}{2}\,\|\theta^{(t)} - \hat{\theta}\,\|_\infty$$

for all $t \geq T_1$. Since $\|\theta^{(T_1)} - \hat{\theta}\|_\infty \leq 8\tau$, the conclusion follows immediately. ∎

A.5. Sharp results in the low-noise regime

Before proving the sharp bound on $\|\hat{\theta} - \theta^*\|_2$, let us start with two lemmas that control the lower order terms.

Lemma 18 *There exist absolute constants $C, C' > 0$ such as the following holds for any $\delta \in (0, 0.1)$. Suppose that all the bounds in Lemma 3 hold with the constant C, and that $|y_r - x_r^\top \theta^*| \leq C\sigma\sqrt{\log\frac{N}{\delta}}$ for all $r \in [N]$. Define*

$$A := \frac{d-1}{2N}\sum_{r=1}^{N}\frac{y_r^2}{\sigma^2}\tanh'\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right)x_r x_r^\top. \tag{43}$$

Moreover, let $\Delta := 2C\sigma\sqrt{\log\frac{N}{\delta}}$, *and define* $S_i(\Delta)$ *as in* (13a). *Then we have*

$$\|A\|_\infty \le C'\left(\max_{i\in[d]}\frac{|S_i(\Delta)|}{d} + \frac{d}{N}\log\frac{d}{\delta}\right)\log\frac{N}{\delta}.$$

Proof We have $\|A\|_\infty = \max_{i\in[d]}\sum_{j=1}^d |A_{ij}|$. If $x_r = e_i - e_j$, then $x_r x_r^\top$ is the matrix with entries $(x_r x_r^\top)_{ii} = (x_r x_r^\top)_{jj} = 1$, $(x_r x_r^\top)_{ij} = (x_r x_r^\top)_{ji} = -1$, and 0 elsewhere. Also, note that $\frac{y_r^2}{\sigma^2}\tanh'\left(\frac{y_r x_r^\top\theta^*}{\sigma^2}\right) \ge 0$. Then, by the definition of A, it is not hard to see that

$$\sum_{j=1}^d |A_{ij}| = \frac{d-1}{N}\sum_{r\in\mathcal{R}_i}\frac{y_r^2}{\sigma^2}\tanh'\left(\frac{y_r x_r^\top\theta^*}{\sigma^2}\right) \le \frac{4d}{N}\sum_{r\in\mathcal{R}_i}\frac{y_r^2}{\sigma^2}\exp\left(-2\left|\frac{y_r x_r^\top\theta^*}{\sigma^2}\right|\right) \tag{44}$$

where $\mathcal{R}_i$ is defined in (12), and the inequality follows from the fact $0 \le \tanh'(t) = (\cosh(t))^{-2} \le 4\exp(-2|t|)$ for all $t\in\mathbb{R}$.

For $\Delta = 2C\sigma\sqrt{\log\frac{N}{\delta}}$, let $\mathcal{R}_i(\Delta)$ be defined in (13b). Similar to the proof of Lemma 16, we split the analysis of the sum in (44) into two parts. First, let us consider $r \in \mathcal{R}_i\setminus\mathcal{R}_i(\Delta)$. Then $|x_r^\top\theta^*| > \Delta$, and by the assumption $|y_r - x_r^\top\theta^*| \le C\sigma\sqrt{\log\frac{N}{\delta}} = \Delta/2$, we have

$$\left|\frac{y_r x_r^\top\theta^*}{\sigma^2}\right| \ge \frac{(x_r^\top\theta^*)^2}{2\sigma^2}.$$

Consequently,

$$\begin{aligned}\frac{4d}{N}\sum_{r\in\mathcal{R}_i\setminus\mathcal{R}_i(\Delta)}\frac{y_r^2}{\sigma^2}\exp\left(-2\left|\frac{y_r x_r^\top\theta^*}{\sigma^2}\right|\right) &\le \frac{4d}{N}N\frac{4(x_r^\top\theta^*)^2}{\sigma^2}\exp\left(-\frac{(x_r^\top\theta^*)^2}{\sigma^2}\right)\\ &\overset{(a)}{\le} 16\,d\frac{\Delta^2}{\sigma^2}\exp\left(-\frac{\Delta^2}{\sigma^2}\right) \overset{(b)}{\le} \frac{1}{N^{10}},\end{aligned}$$

where (a) holds as the function $t\exp(-t)$ is decreasing for $t \ge 1$, and (b) holds by the definition of Δ provided that the constant C is sufficiently large.

Next, consider $r \in \mathcal{R}_i(\Delta)$. We have

$$\begin{aligned}\frac{4d}{N}\sum_{r\in\mathcal{R}_i(\Delta)}\frac{y_r^2}{\sigma^2}\exp\left(-2\left|\frac{y_r x_r^\top\theta^*}{\sigma^2}\right|\right) &\le \frac{4d}{\sigma^2 N}\sum_{r\in\mathcal{R}_i(\Delta)} y_r^2\\ &\le \frac{4d}{\sigma^2 N}C\Delta^2\left(|S_i(\Delta)|\frac{N}{d^2} + \log\frac{d}{\delta}\right)\\ &= 16C^3\left(\frac{|S_i(\Delta)|}{d} + \frac{d}{N}\log\frac{d}{\delta}\right)\log\frac{N}{\delta},\end{aligned}$$

where the second inequality follows from Lemma 3.

Finally, combining the above two bounds with (44) finishes the proof. ■

Lemma 19 *There exist absolute constants $C, C' > 0$ such as the following holds for any $\delta \in (0, 0.1)$. Suppose that all the bounds in Lemma 3 hold with the constant C, and that $|y_r - x_r^\top \theta^*| \le C\sigma\sqrt{\log \frac{N}{\delta}}$ for all $r \in [N]$. For $\theta \in \mathcal{H}$, $t \in [0,1]^d$, and $r \in [N]$, define*

$$\omega(\theta, r, t_r) := \frac{y_r^2}{2\sigma^4} \tanh''\left(\frac{y_r x_r^\top (\theta^* + t_r(\theta - \theta^*))}{\sigma^2}\right)(x_r^\top(\theta - \theta^*))^2, \tag{45a}$$

$$\omega(\theta, t) := \frac{d-1}{2N} \sum_{r=1}^N \omega(\theta, r, t_r)\, y_r\, x_r. \tag{45b}$$

Suppose further that $\|\theta - \theta^*\|_\infty \le \frac{C}{4}\sigma\sqrt{\log \frac{N}{\delta}}$. Then we have

$$\|\omega(\theta, t)\|_\infty \le \frac{C'}{\sigma}\|\theta - \theta^*\|_\infty^2 \Big(\log \frac{N}{\delta}\Big)^{3/2}.$$

Proof Fix $i \in [d]$. By (45) and the definition of $\mathcal{R}_i$ in (13b), we have

$$|\omega(\theta, t)_i| \le \frac{d-1}{2N} \sum_{r \in \mathcal{R}_i} |\omega(\theta, r, t_r)\, y_r| \le \frac{d}{4N} \sum_{r \in \mathcal{R}_i} \frac{|y_r^3|}{\sigma^4} \left|\tanh''\left(\frac{y_r x_r^\top (\theta^* + t_r(\theta - \theta^*))}{\sigma^2}\right)\right| (x_r^\top(\theta - \theta^*))^2.$$

Using that $\big|\tanh''(t)\big| \le 8\exp(-2|t|)$ for all $t \in \mathbb{R}$ and $(x_r^\top(\theta - \theta^*))^2 \le 4\|\theta - \theta^*\|_\infty^2$, we obtain

$$|\omega(\theta, t)_i| \le \frac{8d}{\sigma N}\|\theta - \theta^*\|_\infty^2 \sum_{r \in \mathcal{R}_i} \frac{|y_r^3|}{\sigma^3} \exp\left(-2\left|\frac{y_r x_r^\top (\theta^* + t_r(\theta - \theta^*))}{\sigma^2}\right|\right). \tag{46}$$

Similar to the proof of the previous lemma, we again split the analysis into two parts. Define $\Delta := 2C\sigma\sqrt{\log \frac{N}{\delta}}$. First, consider $r \in \mathcal{R}_i \setminus \mathcal{R}_i(\Delta)$. Then $|x_r^\top \theta^*| > \Delta$, and by the assumptions $|y_r - x_r^\top \theta^*| \le \Delta/2$ and $\|\theta - \theta^*\|_\infty \le \Delta/4$, we have $t_r |x_r^\top(\theta - \theta^*)| \le 2\|\theta - \theta^*\|_\infty \le \Delta/2$ and

$$\left|\frac{y_r x_r^\top (\theta^* + t_r(\theta - \theta^*))}{\sigma^2}\right| \ge \frac{(x_r^\top \theta^*)^2}{4\sigma^2}.$$

It follows that

$$\frac{|y_r^3|}{\sigma^3} \exp\left(-2\left|\frac{y_r x_r^\top (\theta^* + t_r(\theta - \theta^*))}{\sigma^2}\right|\right) \le 8\frac{|x_r^\top \theta^*|^3}{\sigma^3} \exp\left(-\frac{(x_r^\top \theta^*)^2}{2\sigma^2}\right) \le 16$$

since the function $t^3 \exp(-t^2/2) \le 2$ for $t \in \mathbb{R}$. Next, consider $r \in \mathcal{R}_i(\Delta)$. Then $|x_r^\top \theta^*| \le \Delta$ and so $|y_r| \le 2\Delta$. Hence we have

$$\frac{|y_r^3|}{\sigma^3} \exp\left(-2\left|\frac{y_r x_r^\top (\theta^* + t_r(\theta - \theta^*))}{\sigma^2}\right|\right) \le 8\frac{\Delta^3}{\sigma^3} = 64C^3\Big(\log \frac{N}{\delta}\Big)^{3/2}.$$

Putting the two cases together with (46), we obtain

$$|\omega(\theta, t)_i| \le \frac{8d}{\sigma N}\|\theta - \theta^*\|_\infty^2\, |\mathcal{R}_i| \cdot 64C^3\Big(\log \frac{N}{\delta}\Big)^{3/2} \le \frac{C'}{\sigma}\|\theta - \theta^*\|_\infty^2 \Big(\log \frac{N}{\delta}\Big)^{3/2},$$

where the second inequality follows from Lemma 3. This holds for any $i \in [d]$, so the proof is complete. ∎

With these results in hand, we are now ready to perform a sharp analysis of $\|\hat{\theta} - \theta^*\|_2$.

Proposition 20 *There exist absolute constants $C, C' > 0$ such that the following holds for any $\delta \in (0, 0.1)$ and any integer $L \geq 2$. Fix $\theta^{(0)} \in \mathcal{H}$ and let $\Delta^{(0)} := \max\{4L\|\theta^{(0)} - \theta^*\|_\infty, 6\sqrt{L}\,\sigma\}$. Also, let $\Delta := 2C\sigma\sqrt{\log \frac{N}{\delta}}$. Suppose that*

$$N \geq C \max\left\{L^2 d^{\frac{L}{L-1}} \log \frac{dL}{\delta}, L^3 d \log \frac{L}{\delta}\right\}, \qquad N \gg L^4\, d \left(\log \frac{N}{\delta}\right)^5, \tag{47a}$$

$$\max_{i\in[d]} |S_i(\Delta^{(0)})| \leq \frac{d}{CL^2}, \qquad \max_{i\in[d]} |S_i(\Delta)| = o\left(\frac{d}{(L\log(N/\delta))^{3/2}}\right), \tag{47b}$$

where $S_i(\cdot)$ is defined in (13a). Let $\hat{\theta}$ be the limit of the EM sequence $\{\theta^{(t)}\}_{t\geq 0}$ defined in (5) as guaranteed by Proposition 17. Then we have

$$\|\hat{\theta} - \theta^*\|_2^2 \leq (1 + o(1))\,\sigma^2 \frac{d-1}{2N}\,\mathsf{tr}(\hat{\Sigma}^\dagger) + C'\sigma^2\left(\frac{d^{3/2}}{N}\sqrt{\log\frac{1}{\delta}} + \frac{d}{N}\log\frac{1}{\delta}\right).$$

Proof There is an event $\mathcal{E}$ of probability at least $1 - \delta$ on which

- $|y_r - x_r^\top \theta^*| \leq C\sigma\sqrt{\log \frac{N}{\delta}}$ for all $r \in [N]$;
- the bounds in Lemma 3 hold;
- $\|\hat{\Sigma}\|_{op} \leq 3$, $\|\hat{\Sigma}^\dagger\|_{op} \leq 5$, and $\mathsf{tr}(\hat{\Sigma}^\dagger) \geq (d-1)/3$ by Proposition 4 and Lemma 5;
- $\|\hat{\Sigma}^\dagger\|_\infty \leq 2L$ by Proposition 11;
- $\|\hat{Q}(\theta^*) - \theta^*\|_2^2 \leq \sigma^2 \frac{d-1}{2N}\,\mathsf{tr}(\hat{\Sigma}^\dagger) + C\sigma^2\left(\frac{d^{3/2}}{N}\sqrt{\log\frac{1}{\delta}} + \frac{d}{N}\log\frac{1}{\delta}\right)$ by Lemmas 14 and 15;
- $\|\hat{\theta} - \theta^*\|_\infty \leq C\sigma\sqrt{\frac{Ld}{N}\log\frac{d}{\delta}}$ by Proposition 17;
- $\|A\|_\infty \leq C'\left(\max_{i\in[d]} \frac{|S_i(\Delta)|}{d} + \frac{d}{N}\log\frac{d}{\delta}\right)\log\frac{N}{\delta} \leq o\left(\frac{1}{L^{3/2}\sqrt{\log(d/\delta)}}\right)$ by Lemma 18 and the assumptions (47a) and (47b);
- $\|\omega(\hat{\theta}, t)\|_\infty \leq \frac{C'}{\sigma}\|\hat{\theta} - \theta^*\|_\infty^2 \left(\log\frac{N}{\delta}\right)^{3/2} \leq C'C^2\sigma\frac{Ld}{N}\left(\log\frac{N}{\delta}\right)^{5/2} \leq o\left(\frac{\sigma}{L}\sqrt{\frac{d}{N}}\right)$ by Lemma 19, the above bound on $\|\hat{\theta} - \theta^*\|_\infty$, and (47a).

Let us condition on the event $\mathcal{E}$ in the rest of the proof.

Since $\hat{\Sigma}\hat{\theta} = \hat{\Sigma}\hat{Q}(\hat{\theta}) = \bar{Q}(\hat{\theta})$, we have

$$\begin{aligned}\hat{\Sigma}(\hat{\theta} - \theta^*) &= \bar{Q}(\hat{\theta}) - \bar{Q}(\theta^*) + \bar{Q}(\theta^*) - \hat{\Sigma}\theta^* \\ &= \frac{d-1}{2N}\sum_{r=1}^N \left(\tanh\left(\frac{y_r x_r^\top \hat{\theta}}{\sigma^2}\right) - \tanh\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right)\right) y_r x_r + \bar{Q}(\theta^*) - \hat{\Sigma}\theta^*.\end{aligned}$$

By Taylor's theorem, it holds that

$$\tanh\left(\frac{y_r x_r^\top \hat{\theta}}{\sigma^2}\right) - \tanh\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right) = \tanh'\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right)\frac{y_r x_r^\top(\hat{\theta} - \theta^*)}{\sigma^2} + \omega(\hat{\theta}, r, t_r),$$

where the error term $\omega(\hat{\theta}, r, t_r)$ is defined in (45a) for some $t_r \in [0,1]$. With this in hand, we have

$$\begin{aligned}\hat{\Sigma}(\hat{\theta}-\theta^*) &= \frac{d-1}{2N}\sum_{r=1}^{N}\tanh'\left(\frac{y_r x_r^\top \theta^*}{\sigma^2}\right)\frac{y_r x_r^\top(\hat{\theta}-\theta^*)}{\sigma^2}y_r x_r + \omega(\hat{\theta},t) + \bar{Q}(\theta^*) - \hat{\Sigma}\theta^* \\ &= A(\hat{\theta}-\theta^*) + \bar{Q}(\theta^*) - \hat{\Sigma}\theta^* + \omega(\hat{\theta},t),\end{aligned}$$

where $\omega(\hat{\theta}, t)$ is defined in (45b) for $t = (t_1, \ldots, t_d)$, and A is defined in (43). It follows that

$$\hat{\theta} - \theta^* = \hat{Q}(\theta^*) - \theta^* + \hat{\Sigma}^\dagger A(\hat{\theta}-\theta^*) + \hat{\Sigma}^\dagger\omega(\hat{\theta},t). \tag{48}$$

On the event $\mathcal{E}$, the main term $\hat{Q}(\theta^*) - \theta^*$ in the above formula satisfies

$$\|\hat{Q}(\theta^*) - \theta^*\|_2^2 \le \sigma^2\frac{d-1}{2N}\,\mathsf{tr}(\hat{\Sigma}^\dagger) + C\sigma^2\left(\frac{d^{3/2}}{N}\sqrt{\log\frac{1}{\delta}} + \frac{d}{N}\log\frac{1}{\delta}\right). \tag{49}$$

The error term satisfies

$$\|\hat{\Sigma}^\dagger A(\hat{\theta}-\theta^*) + \hat{\Sigma}^\dagger\omega(\hat{\theta},t)\|_\infty \le \|\hat{\Sigma}^\dagger\|_\infty\|A\|_\infty\|\hat{\theta}-\theta^*\|_\infty + \|\hat{\Sigma}^\dagger\|_\infty\|\omega(\hat{\theta},t)\|_\infty \le o\left(\sigma\sqrt{d/N}\right)$$

in view of the bounds $\|\hat{\Sigma}^\dagger\|_\infty \le 2L$, $\|\hat{\theta}-\theta^*\|_\infty \le C\sigma\sqrt{\frac{Ld}{N}\log\frac{d}{\delta}}$, $\|A\|_\infty \le o\left(\frac{1}{L^{3/2}\sqrt{\log(d/\delta)}}\right)$, and $\|\omega(\hat{\theta},t)\|_\infty \le o\left(\frac{\sigma}{L}\sqrt{\frac{d}{N}}\right)$ on the event $\mathcal{E}$. It follows that

$$\|\hat{\Sigma}^\dagger A(\hat{\theta}-\theta^*) + \hat{\Sigma}^\dagger\omega(\hat{\theta},t)\|_2^2 \le o\left(\sigma^2 d^2/N\right). \tag{50}$$

Finally, since $\mathsf{tr}(\hat{\Sigma}^\dagger) \ge (d-1)/3$ on the event $\mathcal{E}$, the first term of the bound in (49) dominates the bound in (50). Combining (49) and (50) with (48) then completes the proof. ■

A.6. Proof of main results

Proof [Proof of Theorem 1] The theorem follows from Proposition 17 together with the assumption $\theta^* \in \Theta(\beta)$. We first check that the conditions in Theorem 1 imply the conditions in (36).
Checking (36a)**:** On the one hand, and as claimed in the statement of the theorem,

$$N \ge \max\{d^{1+\rho}, N_0\}. \tag{51}$$

On the other hand, choosing $L = 1 + \lceil 2/\rho \rceil$ and $\delta = N^{-D}$ in Proposition 17, we obtain the sample size requirement

$$N \ge C\max\left\{L^2 d^{\frac{L}{L-1}}\log\frac{dL}{\delta}, L^3 d\log\frac{L}{\delta}\right\}. \tag{52}$$

It can be easily verified that for a sufficiently large constant $N_0 > 0$, (51) implies (52).
Checking (36b)**:** Furthermore, since $\theta^* \in \Theta(\beta)$, if $|\theta_i^* - \theta_j^*| \le \Delta$, then $|i-j| \le \Delta d/\beta$. Therefore, we have

$$S_i(\Delta) = \{j \in [d]\setminus\{i\} : |\theta_i^* - \theta_j^*| \le \Delta\} \subset \{j \in [d]\setminus\{i\} : |i-j| \le \Delta d/\beta\},$$

so

$$|S_i(\Delta)| \le 2\Delta d/\beta.$$

As a result, for $\Delta^{(0)} = \max\{4L\|\theta^{(0)} - \theta^*\|_\infty, 6\sqrt{L}\,\sigma\}$, it holds that

$$\max_{i\in[d]} |S_i(\Delta^{(0)})| \le \max\{8L\|\theta^{(0)} - \theta^*\|_\infty d/\beta, 12\sqrt{L}\,\sigma d/\beta\} \le \frac{d}{CL^2} \tag{53}$$

in view of the assumptions $\|\theta^{(0)} - \theta^*\|_\infty \le c_1\beta$ and $\sigma \le c_1\beta$.

It remains to compare τ in Theorem 1 and Proposition 17: once we set $\delta = N^{-D}$, the two are the same up to multiplicative factors depending solely on the pair (ρ, D). ■

Proof [Proof of Theorem 2] The theorem follows from Proposition 20. Continuing from the proof of Theorem 1, we check the two additional conditions in (47) compared to (36).
Checking (47a)**:** Choosing $L = 1 + \lceil 2/\rho \rceil$ and $\delta = N^{-D}$ again, we can use the assumption $N \ge \max\{d^{1+\rho}, N_0\}$ to verify that

$$N \gg L^4\, d \left(\log \frac{N}{\delta}\right)^5.$$

Checking (47b)**:** We again have $|S_i(\Delta)| \le 2\Delta d/\beta$, so for $\Delta = 2C\sigma\sqrt{\log \frac{N}{\delta}}$, it holds

$$\max_{i\in[d]} |S_i(\Delta)| \le \frac{4C\sigma d}{\beta}\sqrt{\log \frac{N}{\delta}} = o\left(\frac{d}{(L\log(N/\delta))^{3/2}}\right)$$

in view of the assumption $\sigma = o\left(\frac{\beta}{(\log N)^2}\right)$.
Implication: Proposition 20 then gives

$$\|\hat\theta - \theta^*\|_2^2 \le (1 + o(1))\,\sigma^2 \frac{d-1}{2N}\,\mathsf{tr}(\hat\Sigma^\dagger) + C'\sigma^2\left(\frac{d^{3/2}}{N}\sqrt{\log N} + \frac{d}{N}\log N\right).$$

Finally, $\mathsf{tr}(\hat\Sigma^\dagger) \ge (d-1)/3$ with high probability by Lemma 5, so the first term above dominates the other terms when $d \gg \log N$. Therefore, Theorem 2 follows. ■

Appendix B. Probability tools

Lemma 21 (Gaussian concentration, Theorem 5.2.2 of Vershynin (2018)) *Consider a random vector $X \sim \mathcal{N}(0, I_n)$ and an L-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$. Then we have*

$$\Pr\{|f(X) - \mathbb{E}[f(X)]| \ge t\} \le 2\exp\left(-ct^2/L^2\right)$$

for an absolute constant $c > 0$.

Definition 22 (Convex concentration property) *Let X be a random vector in $\mathbb{R}^n$. We say that X has the convex concentration property with constant K if for every 1-Lipschitz convex function $\phi : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}[|\phi(X)|] < \infty$, and for every $t > 0$,*

$$\Pr\{|\phi(X) - \mathbb{E}[\phi(X)]| \ge t\} \le 2\exp\left(-t^2/K^2\right).$$

Lemma 23 (Theorem 2.3 of Adamczak (2015)) *Let X be a mean zero random vector in $\mathbb{R}^n$. If X has the convex concentration property with constant K, then for any $A \in \mathbb{R}^{n\times n}$ and any $t > 0$,*

$$\Pr\{|X^\top AX - \mathbb{E}[X^\top AX]| \geq t\} \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_{op}}\right)\right),$$

for an absolute constant $C > 0$.

Lemma 24 (Matrix Bernstein, Theorem 1.6.2 of Tropp (2015)) *Consider a finite sequence $\{X_k\}_{k=1}^N$ of independent, random, Hermitian matrices in $\mathbb{R}^{d\times d}$. Assume that*

$$\mathbb{E}[X_k] = 0 \quad \textit{and} \quad \|X_k\|_{op} \leq L \quad \textit{for each } k \in [N].$$

Introduce the random matrix

$$Y = \sum_{k=1}^N X_k.$$

Let $v(Y)$ be the matrix variance statistic of the sum:

$$v(Y) = \|\mathbb{E}[Y^2]\| = \left\|\sum_{k=1}^N \mathbb{E}[X_k^2]\right\|.$$

Then

$$\mathbb{E}[\|Y\|] \leq \sqrt{2v(Y)\log(2d)} + \frac{1}{3}L\log(2d).$$

Furthermore, for all $t > 0$,

$$\Pr\{\|Y\| \geq t\} \leq 2d\exp\left(-\frac{t^2/2}{v(Y) + Lt/3}\right).$$

Lemma 25 (Lemma 2.27 of Wainwright (2019)) *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a differentiable function. Then for every convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have*

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

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

Appendix C. Improved sample complexity

In Theorem 1, we assume the conditions

- (sample size) $N \geq d^{1+\rho}$;
- (noise) $\sigma \leq c_1\beta$;
- (initialization) $\|\theta^{(0)} - \theta^*\|_\infty \leq c_1\beta$.

It is possible to improve the sample complexity at the cost of strengthening the other two conditions. Namely, we may assume the alternative conditions

- (sample size) $N \geq d\,(\log d)^{C_1}$;
- (noise) $\sigma \leq \frac{c_2\,\beta}{(\log d)^{5/2}}$;
- (initialization) $\|\theta^{(0)} - \theta^*\|_\infty \leq \frac{c_2\,\beta}{(\log d)^3}$.

To see the sufficiency of this set of conditions, we apply Proposition 17 with $L = \lceil \frac{\log d}{\log\log d} \rceil$ and $\delta = N^{-D}$ for a constant $D > 0$. It suffices to check the two conditions in (36). We have

$$d^{\frac{1}{L-1}} \leq d^{\frac{2\log\log d}{\log d}} = (\log d)^2.$$

Therefore, if $N \geq d\,(\log d)^{C_1}$ for a sufficiently large constant $C_1 = C_1(D) > 0$, then

$$N \geq C \max\Big\{L^2 d^{\frac{L}{L-1}} \log\frac{dL}{\delta}, L^3 d \log\frac{L}{\delta}\Big\}.$$

Further, as in (53), we have

$$\max_{i\in[d]} |S_i(\Delta^{(0)})| \leq \max\{8L\|\theta^{(0)} - \theta^*\|_\infty d/\beta, 12\sqrt{L}\,\sigma d/\beta\} \leq \frac{d}{CL^2}$$

provided that $\|\theta^{(0)} - \theta^*\|_\infty \leq \frac{\beta}{8CL^3}$ and $\sigma \leq \frac{\beta}{12CL^{5/2}}$. These are satisfied once we choose $c_2 > 0$ to be sufficiently small.

Appendix D. Minimax lower bounds with a fixed design

If the signs z_r are given in the model (1) and the covariates x_r are fixed, then the problem reduces to linear regression with a fixed design. It is an elementary fact that the least squares estimator $\hat{\theta}^{\mathsf{LS}}$ achieves the risk

$$\mathbb{E}\,\|\hat{\theta}^{\mathsf{LS}} - \theta^*\|_2^2 = \sigma^2\,\mathsf{tr}\left(\Big(\sum_{r=1}^N x_r x_r^\top\Big)^\dagger\right).$$

It is well-known that this risk is minimax-optimal. To show that $\sigma^2\,\mathsf{tr}\left(\big(\sum_{r=1}^N x_r x_r^\top\big)^\dagger\right)$ is a lower bound on the minimax risk, the standard proof is via the Bayes risk assuming that θ^* has the prior distribution $\mathcal{N}\Big(0, \tau^2\big(\sum_{r=1}^N x_r x_r^\top\big)^\dagger\Big)$ for a parameter τ. One can compute the Bayes-optimal estimator which is the posterior mean of θ^*. Then the Bayes risk achieved by the posterior mean evaluates to $\frac{\tau^2}{1+\tau^2}\sigma^2\,\mathsf{tr}\left(\big(\sum_{r=1}^N x_r x_r^\top\big)^\dagger\right)$. Since the Bayes risk is a lower bound on the minimax risk, letting $\tau \to \infty$ proves the desired lower bound.

If the signs z_r are unknown in the mixture model (1) and the covariates x_r are fixed, then $\sigma^2\,\mathsf{tr}\left(\big(\sum_{r=1}^N x_r x_r^\top\big)^\dagger\right)$ is still a lower bound on the minimax risk for estimating θ^*. Therefore, conditional on a typical realization of $(x_r)_{r=1}^N$, the proof of Theorem 2 in fact shows that the EM fixed point $\hat{\theta}$ achieves the minimax rate with the sharp constant (see Theorem 2 and Eq. (7)). Since this work assumes a random design of the covariates, to keep our terminology consistent, we do not further formalize a minimax result in the fixed-design setting.