Beyond EM Algorithm on Over-specified Two-Component Location-Scale Gaussian Mixtures

Tongzheng Ren$^{\star,\diamond,\ddagger}$ Fuheng Cui$^{\star,\flat}$ Sujay Sanghavi† Nhat Ho$^{\flat,\ddagger}$

Department of Computer Science, University of Texas at Austin$^{\diamond}$,
Department of Statistics and Data Sciences, University of Texas at Austin$^{\flat}$
Department of Electrical and Computer Engineering, University of Texas at Austin†,
May 23, 2022

Abstract

The Expectation-Maximization (EM) algorithm has been predominantly used to approximate the maximum likelihood estimation of the location-scale Gaussian mixtures. However, when the models are over-specified, namely, the chosen number of components to fit the data is larger than the unknown true number of components, EM needs a polynomial number of iterations in terms of the sample size to reach the final statistical radius; this is computationally expensive in practice. The slow convergence of EM is due to the missing of the locally strong convexity with respect to the location parameter on the negative population log-likelihood function, i.e., the limit of the negative sample log-likelihood function when the sample size goes to infinity. To efficiently explore the curvature of the negative log-likelihood functions, by specifically considering two-component location-scale Gaussian mixtures, we develop the *Exponential Location Update* (ELU) algorithm. The idea of the ELU algorithm is that we first obtain the exact optimal solution for the scale parameter and then perform an exponential step-size gradient descent for the location parameter. We demonstrate theoretically and empirically that the ELU iterates converge to the final statistical radius of the models after a logarithmic number of iterations. To the best of our knowledge, it resolves the long-standing open question in the literature about developing an optimization algorithm that has optimal statistical and computational complexities for solving parameter estimation even under some specific settings of the over-specified Gaussian mixture models.

1   Introduction

Location-scale Gaussian mixture models are mixture models in which both the means and the covariances of the Gaussian components are unknown (and need to be estimated from data). Such models have been widely used to model the heterogeneity of the data [21, 23], and approximating the unknown distribution of data with the density function of the Gaussian mixture models [9]. In Location-scale Gaussian mixture models, the parameters are used to capture the distinct behaviors of each sub-population in the data. A popular approach to obtain parameter estimation in Gaussian mixture models is by using the maximum likelihood estimation (MLE). The convergence rates of MLE had been studied extensively in the literature. When the Gaussian mixture models are exactly-specified, namely, the number of components is known, Ho et al. [15] demonstrated that the convergence rates of MLE are parametric under Wasserstein metric, namely, they are at the order of $\mathcal{O}(n^{-1/2})$. However, the number of components in Gaussian mixture models is rarely known in practice. There

---

* Tongzheng Ren and Fuheng Cui contributed equally to this work.

‡ Correspondence to: Tongzheng Ren (tongzheng@utexas.edu) and Nhat Ho (minhnhat@utexas.edu).

are two popular line of directions to account for the unknown number of components. The first line of directions consist of approaches that penalize the sample log-likelihood function of the Gaussian mixture models via the number of parameters [25, 24] or via the separation of the parameters [22]. While these methods can guarantee the consistency of estimating the true number of components when the sample size is sufficiently large, they tend to be computationally expensive, especially when the true number of components is quite large. To overcome the high computation of the first line of directions, the second line of directions include approaches that over-specify the number of the components in Gaussian mixture models, namely, we choose a large number of components based on our domain knowledge of the data and fit the model using that number of components. These approaches along this direction are called over-specified location-scale Gaussian mixture models [4, 26, 12, 11]. Under the over-specified settings, the convergence rates of the MLE are determined by the amount of extra components [14]. In general, the rates of the MLE (or equivalently statistical rates) are very slow when the extra number of components is large.

Furthermore, due to the complicated sample log-likelihood function of the location-scale Gaussian mixtures [17], the MLE does not have closed-form expression. The Expectation-Maximization (EM) algorithm [6] has been widely used to approximate the optimal solution of the sample log-likelihood function. Under the exact-specified settings of the location-scale Gaussian mixtures when the true location and scale parameters have sufficiently large separation, the EM iterates have been shown to converge to a radius of convergence $\mathcal{O}((d/n)^{1/2})$ within the true parameter after a logarithmic number of iterations $\mathcal{O}(\log(n/d))$ [1, 2, 31, 5, 29, 28, 30]. However, for the over-specified settings of these models, the EM iterates can take polynomial number of iterations $\mathcal{O}(n^\tau)$ for some $\tau > 0$ to reach the final statistical radius. In particular, Dwivedi et al. [7] considered the isotropic symmetric two-component location-scale Gaussian mixture models and demonstrated that when the model is over-specified and dimension $d = 1$, the EM iterates for the location and scale parameters respectively reach the final statistical radii $\mathcal{O}(n^{-1/8})$ and $\mathcal{O}(n^{-1/4})$ within the true location and scale parameters after $\mathcal{O}(n^{3/4})$ number of iterations. When the dimension $d \geq 2$, the statistical rates of the EM updates for location and scale parameters are $\mathcal{O}((d/n)^{1/4})$ and $\mathcal{O}((nd)^{-1/2})$ and these rates are achieved after $\mathcal{O}((n/d)^{1/2})$ number of iterations. An important insight for that slow convergence of the EM iterates to the final statistical radii is that the population log-likelihood function, namely, the limit of the sample log-likelihood function when the sample size goes to infinity, is locally strongly convex with respect to the scale parameter but only locally convex with respect to the location parameter. The lack of the strong convexity with respect to the location parameter indicates that the EM iterates have sub-linear convergence rate to the true location parameter, which leads to the polynomial number of iterations for the EM iterates to reach the statistical radius of convergence.

In practice, the polynomial number of iterations of the EM algorithm for reaching the final statistical radius can be computationally expensive and undesirable when the sample size n is sufficiently large. It leads to the following long-standing open question that we aim to address in the paper:

"*Is it possible to develop an optimization algorithm that converges to the final statistical radius within the true parameters after a logarithmic number of iterations in terms of the sample size and dimension even under some specific settings of the over-specified Gaussian mixture models?*"

Contribution. In this work, we provide an affirmative answer to this question. We specifically consider the over-specified settings of the symmetric two-component location-scale Gaus-

sian mixture models, which had been used extensively in the literature to study the non-asymptotic behaviors of the EM algorithm [1, 32, 8, 7, 5, 29]. In particular, we use the symmetric two-component location-scale Gaussian mixture model $\frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d)$ to fit the data that are i.i.d. from a multivariate normal distribution $\mathcal{N}(\theta^*, (\sigma^*)^2 I_d)$ where $\theta^* = 0$. We develop the *Exponential Location Update* (ELU) algorithm for solving parameter estimation in these models. The high level idea of ELU is that, the optimal scale parameter can be written as a function of the optimal location parameter, and we can plug-in that scale parameter to the negative sample log-likelihood function and utilize an exponential rate schedule for gradient descent to update the location parameter. The exponential rate schedule for gradient descent is used to efficiently explore the flat curvature of the negative sample log-likelihood function along the direction of location parameter. After we obtain an approximation of the location parameter, we can approximate the scale parameter with the relation between the location parameter and the scale parameter.

When dimension $d = 1$, we demonstrate that ELU iterates for the location and scale parameters respectively reach the final statistical radii $\mathcal{O}(n^{-1/8})$ and $\mathcal{O}(n^{-1/4})$ within the true location parameter θ^* and the true scale parameter σ^* after $\mathcal{O}(\log(n))$ number of iterations. When $d \geq 2$, these iterates converge to the neighborhoods of radii $\mathcal{O}((d/n)^{1/4})$ and $\mathcal{O}((nd)^{-1/2})$ within θ^* and σ^* after logarithmic number of iterations $\mathcal{O}(\log(n/d))$. These results are in stark difference from the polynomial number of iterations of the EM algorithm for solving parameter estimation [7], which are $\mathcal{O}(n^{3/4})$ when $d = 1$ and $\mathcal{O}((n/d)^{1/2})$ when $d \geq 2$. As a consequence, for fixed dimension d, given that per iteration cost of the ELU algorithm is $\mathcal{O}(nd)$, the ELU algorithm has optimal computational complexity $\mathcal{O}(nd\log(n/d))$ for reaching the final statistical radii in all of these models.

Beyond symmetric two-component Gaussian mixtures: In Appendix B, we provide a discussion showing that the ELU algorithm can still be useful in more general settings than the over-specified symmetric two-component location-scale Gaussian mixtures (1). We specifically consider two settings: (i) beyond the isotropic covariance matrix in Appendix B.1; (ii) beyond the symmetric location parameters in Appendix B.2. To the best of our knowledge, the theoretical analysis of optimization algorithms for these settings has not been established before in the literature. In these Appendices, we provide the insight into the fast convergence of the ELU algorithms (via both empirical and theoretical results) for solving parameter estimation of these models.

Organization. The paper is organized as follows. We first provide background for the over-specified settings of the symmetric two-component location-scale Gaussian mixtures in Section 2.1. We then interpret the EM algorithm for solving these models as coordinate descent algorithm and demonstrate in high level why the EM iterates of the location parameter have slow convergence towards to the true location parameter. To overcome the slow convergence of the EM, we develop the Exponential Location Update (ELU) algorithm in Section 2.2 and demonstrate that the iterates of ELU converge to the final statistical radii after a logarithmic number of iterations. We then provide the proof sketch for the convergence of the ELU iterates in Section 2.3 while concluding the paper in Section 3. Finally, the proofs of all the results in the main text as well as the remaining materials are deferred to the supplementary materials.

Notation. For any matrix $A \in \mathbb{R}^{d \times d}$, we denote by $\lambda_{\max}(A)$ the maximum eigenvalue of the matrix A. For any $x \in \mathbb{R}^d$, $\|x\|$ denotes the ℓ_2 norm of x. For any two sequences $\{a_n\}_{n \geq 1}, \{b_n\}_{n \geq 1}$, we denote $a_n = \mathcal{O}(b_n)$ to mean that $a_n \leq C b_n$ for all $n \geq 1$ where C is some

universal constant. Furthermore, we denote $a_n = \Theta(b_n)$ to indicate that $C_1 b_n \leq a_n \leq C_2 b_n$ for any $n \geq 1$ where C_1, C_2 are some universal constants.

2 Symmetric Two-Component Location-Scale Gaussian Mixtures

In this section, we consider the symmetric two-component location-scale Gaussian mixture with isotropic covariance matrices. Although this setting can be a little bit simplistic, it had been used extensively in the recent literature to shed light into the non-asymptotic behaviors of the EM algorithm [1, 31, 8, 7, 5, 28, 29, 19]. Furthermore, to the best of our knowledge, it is also the only settings that the non-asymptotic behaviors of the EM algorithm were established. Therefore, we will use these settings to gain insight into developing an optimal optimization algorithm that outperforms the EM algorithm under the over-specified regime of these settings.

2.1 Problem Settings

We assume that $X_1, \cdots, X_n$ are i.i.d. samples from the symmetric two-component location-scale Gaussian mixture

$$\frac{1}{2}\mathcal{N}(-\theta^*, (\sigma^*)^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^*, (\sigma^*)^2 I_d) \tag{1}$$

where θ^* and $\sigma^* = 1$ are true but unknown parameters. Note that, the assumption $\sigma^* = 1$ is just for the simplicity of the proof presentation; the results in this section can be generalized to any unknown value of σ^*. When $\|\theta^*\|/\sigma^* \geq C$, this setting is referred as high signal-to-noise regime of model (1). It is also widely regarded as strong separated (or equivalently exactly-specified) setting of location-scale Gaussian mixtures. When $\theta^* = 0$, this setting corresponds to the over-specified setting (or equivalently low signal-to-noise regime) of model (1).

To estimate θ^* and σ^*, we consider the symmetric two-component location-scale Gaussian mixture with isotropic covariance matrix, namely, we fit the following model to the data:

$$\frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d). \tag{2}$$

The popular way to obtain estimates for θ^* and σ^* is to solve for the MLE of model (2), which is given by:

$$-\operatorname*{arg\,min}_{\theta,\sigma} \mathcal{L}_n(\theta, \sigma) := -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{2}\phi(X_i; \theta, \sigma^2 I_d) + \frac{1}{2}\phi(X_i; -\theta, \sigma^2 I_d)\right), \tag{3}$$

where $\phi(\cdot; \theta, \sigma^2 I_d)$ is the density function of multivariate Gaussian distribution with mean θ and covariance matrix $\sigma^2 I_d$. Unfortunately, the MLE does not have closed-form expression; hence, we need to leverage the optimization algorithms to approximate the solution of MLE.

Insight into the EM algorithm: The EM algorithm has been widely used to solve the MLE approximately [6]. Under the strong signal-to-noise regime, namely, $\|\theta^*\|/\sigma^* \geq C$ for some universal constant C, the EM iterates reach the final statistical radii $\mathcal{O}((d/n)^{1/2})$ within the true parameters θ^* and σ^* after $\mathcal{O}(\log(n/d))$ number of iterations [2]. However, when $\theta^* = 0$, i.e., the over-specified settings of model 2, the EM iterates have slow convergence to

4

the true parameters [7]. To gain an insight into the slow convergence of EM, we compute the updates of the EM algorithm for location and scale parameters, which are given by:

$$\theta_{n,\text{EM}}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} X_i \tanh\left(\frac{X_i^{\top} \theta_{n,\text{EM}}^t}{\left(\sigma_{n,\text{EM}}^{t+1}\right)^2}\right), \tag{4}$$

$$\sigma_{n,\text{EM}}^{t+1} = \frac{1}{nd} \sum_{i=1}^{n} \|X_i\|^2 - \frac{\|\theta_{n,\text{EM}}^{t+1}\|^2}{d}. \tag{5}$$

With some computations, these EM updates can be rewritten in the following way:

$$\theta_{n,\text{EM}}^{t+1} = \theta_{n,\text{EM}}^t - (\sigma_{n,\text{EM}}^t)^2 \nabla_\theta \mathcal{L}_n(\theta_{n,\text{EM}}^t, \sigma_{n,\text{EM}}^t), \tag{6}$$

$$\sigma_{n,\text{EM}}^{t+1} = \min_\sigma \nabla_\theta \mathcal{L}_n(\theta_{n,\text{EM}}^{t+1}, \sigma). \tag{7}$$

Therefore, at a high level, the EM performs the *coordinate descent* updates, namely, the EM update of the scale parameter at step $t+1$ performs an exact minimization problem of the negative log-likelihood function $\mathcal{L}_n$ with respect to the scale parameter at $\theta_{n,\text{EM}}^{t+1}$. Meanwhile, the EM update of the location parameter at step $t+1$ performs adaptive gradient descent update of the negative log-likelihood function $\mathcal{L}_n$ with respect to the location parameter where the step-size is $(\sigma_{n,\text{EM}}^t)^2$. As $\sum_{i=1}^{n} \|X_i\|^2/(nd)$ converges to 1 when n goes to infinity and $\theta_{n,\text{EM}}^{t+1}$ converges to a neighborhood close to $\theta^* = 0$ when t approaches infinity, the adaptive step-size of the GD algorithm for the location parameter (4) eventually approaches a constant step-size when n and t go to infinity. However, when n goes to infinity and $\theta_{n,\text{EM}}^t$ approaches θ^*, we have $\|\nabla_\theta \mathcal{L}_n(\theta_{n,\text{EM}}^t, \sigma_{n,\text{EM}}^t)\| \approx \|\theta_{n,\text{EM}}^t\|^7$ when $d = 1$ and $\|\mathcal{L}_n(\theta_{n,\text{EM}}^t, \sigma_{n,\text{EM}}^t)\| \approx \|\theta_{n,\text{EM}}^t\|^3$ when $d \geq 2$ (cf. Lemma 1 and Lemma 3 in [7]). It indicates that when n goes to infinity, the negative log-likelihood function $\mathcal{L}_n$ with respect to the location parameter is flat around the optimal parameter and the contraction coefficient in the convergence of $\theta_{n,\text{EM}}^t$ to θ^* goes to 1, which leads to the slow convergence of the EM update for the location parameter.

2.2 Exponential Location Update (ELU) algorithm

As we have seen, the slow convergence of the EM update for the location parameter (6) to the true location parameter is due to the asymptotically constant step-size schedule and the flat curvature of the negative log-likelihood function $\mathcal{L}_n$ when the sample size n goes to infinity. It suggests that to improve the convergence of the EM iterates for the location parameter, we need to utilize an increasing step-size schedule that takes into account the flatness of the function $\mathcal{L}_n$ with respect to the location parameter.

In this section, we develop a new optimization algorithm based on that spirit. With simple computation, we first observe that the optimal scale parameter and the optimal location parameter should satisfy Equation (5). Hence, we can directly optimize the location parameter with respect to the function $f_n(\theta) := \mathcal{L}_n(\theta, \frac{1}{nd} \sum_{i=1}^{n} \|X_i\|^2 - \frac{\|\theta\|^2}{d})$ and performs an exponential step-size schedule for the gradient descent of the function f_n to update the location parameter [20, 16]. The new algorithm (See Algorithm 1), named *Exponential Location Update* (ELU) algorithm, admits the following updates for the location and scale parameters:

$$\theta_n^{t+1} = \theta_n^t - \frac{\eta}{\beta^t} \nabla f_n(\theta_n^t), \tag{8}$$

$$(\sigma_n^{t+1})^2 = \frac{\sum_{i=1}^{n} \|X_i\|^2}{nd} - \frac{\|\theta_n^{t+1}\|^2}{d}, \tag{9}$$

Algorithm 1: Exponential Location Update (ELU)

> **Input:** The step size η, and the scaling parameter $\beta \in (0,1)$
> **Output:** The updates θ_n^T, σ_n^T for the location and scale parameters
> Initialize θ_n^0 and σ_n^0
> **for** $t = 1$ **to** $T-1$ **do**
>   Update location parameter: $\theta_n^{t+1} = \theta_n^t - \frac{\eta}{\beta^t} \nabla f_n(\theta_n^t)$,
>   Update scale parameter: $(\sigma_n^{t+1})^2 = \frac{\sum_{i=1}^n \|X_i\|^2}{nd} - \frac{\|\theta_n^{t+1}\|^2}{d}$
> **end for**
> Return θ_n^T, σ_n^T

where (θ_n^0, σ_n^0) are initialization of the ELU algorithm. Here, $\eta > 0$ is the step-size and $\beta \in (0,1]$ is a given parameter. When $\beta = 1$, the update in equation (8) becomes the standard gradient descent update on the function f_n.

Remarks on the ELU algorithm in Algorithm 1: First, the intuition behind scaling the step size as β^{-t} is that, such step size schedule prevents the contraction coefficient of the location updates in equation (8) to go to 1 when the location updates approach the optimal parameter, which enables the linear convergence of ELU. To be more concrete, consider the case when $d = 1$, where $\nabla f_n(\theta_n^t) \approx (\theta_n^t)^7$ when n goes to infinity (see Lemma 1). If we want $\theta_n^t = \Theta(\kappa^t)$, we need to scale the step-size as κ^{-6t}, which is exponential in t. Second, the update of ELU for the location parameter in equation 8 indicates that we take the whole gradient of the function f_n to update the location parameter, which also needs to propagate the gradient from the scale parameter in $\mathcal{L}_n$. It is different from the EM update of the location parameter in equation (6) where we first compute the gradient of the negative log-likelihood function $\mathcal{L}_n$, then substitute the most recent scale parameter into that gradient and update the location parameter. Such design is necessary to apply the population to sample analysis [1] if the update does not have certain close-form.

Optimality of the ELU algorithm under the over-specified settings: We now provide theoretical analysis for the statistical behaviors and computational complexity of the ELU iterates $(\theta_n^t, \sigma_n^t)_{t \geq 0}$ under the over-specified setting of model (2), namely, $\theta^* = 0$.

Theorem 1. *Given the over-specified settings of the symmetric two-component location-scale Gaussian mixture* (2), *assume that the step size η and the scaling parameter β of the ELU algorithm in Algorithm 1 are chosen such that $\frac{\eta C_1 \|\theta_n^0 - \theta^*\|^\alpha \beta}{(\alpha+1)(\alpha+2)} + \beta^{2/\alpha} \geq 1$ and $\eta C_1 (\alpha+2) 3^\alpha \|\theta_n^0 - \theta^*\|^\alpha \leq \alpha+1$ where $\alpha = 6$ when $d = 1$ and $\alpha = 2$ when $d \geq 2$ and C_1 is some universal constant. Then, as long as the sample size n is large enough such that $n \geq cd \log(d/\delta)$ for some universal constant c, with probability $1 - \delta$ for any fixed $\delta \in (0,1)$ there exist universal constants $\{c_i\}_{i=1}^4$ and $\{c_i'\}_{i=1}^2$ such that the ELU iterates $\{\theta_n^t\}$ and $\{\sigma_n^t\}$ in equations* (8) *and* (9) *satisfy the following bounds:*

(a) (Univariate setting) When $d = 1$, as long as $t \geq c_1' \log \left(\frac{n}{d \log(1/\delta)} \right)$ we find that

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^*\| \leq \frac{c_1 \log^{10}(5n/\delta)}{n^{1/8}}, \qquad \min_{1 \leq k \leq t} |(\sigma_n^k)^2 - (\sigma^*)^2| \leq \frac{c_2 \log^{10}(5n/\delta)}{n^{1/4}}.$$
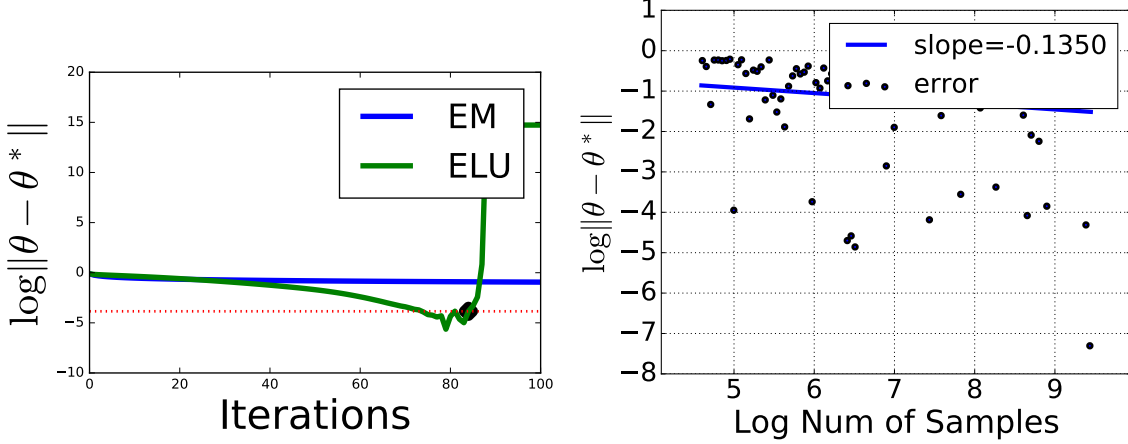
6

Figure 1. Illustrations for the case with $d = 1$. **Left:** The optimization rate of EM and ELU. The black diamond shows the ELU iterates with minimum validation error. ELU can converge to the statistical radius with a linear rate then diverge, while EM converge to the statistical radius with a sub-linear rate. **Right:** ELU can find a solution of θ within the statistical radius $\mathcal{O}(n^{-1/8})$. Given the update of σ in ELU algorithm, its statistical radius is directly $\mathcal{O}(n^{-1/4})$.

(b)] *(Multivariate setting) When $d \geq 2$, as long as $t \geq c'_2 \log\left(\frac{n}{d\log(1/\delta)}\right)$ we find that*

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^*\| \leq c_3 \left(\frac{d + \log(1/\delta)}{n}\right)^{1/4}, \quad \min_{1 \leq k \leq t} |(\sigma_n^k)^2 - (\sigma^*)^2| \leq \frac{c_4 \log(1/\delta)}{(n(d + \log(1/\delta)))^{1/2}}.$$

The proof of Theorem 1 is in Appendix A. We provide road map of the proof of Theorem 1 in Section 2.3. A few comments with Theorem 1 are in order.

Comparing to the EM algorithm: When $d = 1$, the result of part (a) of Theorem 1 indicates that the ELU iterates for the location and scale parameters respectively reach the final statistical radii $\mathcal{O}(n^{-1/8})$ and $\mathcal{O}(n^{-1/4})$ within θ^* and σ^* after a logarithmic number of iterations $\mathcal{O}(\log(n))$. On the other hand, the result of Theorem 1 of [7] indicates that the EM updates for the location and scale parameters also reach the similar final statistical radii as those of the ELU iterates; however, the EM algorithm requires $\mathcal{O}(n^{3/4})$ number of iterations to get to these radii, which is computationally much more expensive than the logarithmic number of iterations of the ELU algorithm.

When $d \geq 2$, part (b) of Theorem 1 proves that the final statistical radii of the ELU iterates for the location and scale parameters are respectively $\mathcal{O}((d/n)^{1/4})$ and $\mathcal{O}((nd)^{-1/2})$ and these radii are achieved after $\mathcal{O}(\log(n/d))$ number of iterations. For the EM iterates, Theorem 2 of [7] shows that their statistical radii are similar to those of the ELU iterates and these radii are reached after $\mathcal{O}((n/d)^{1/2})$ number of iterations. Therefore, the ELU algorithm also requires exponentially less iterations than the EM algorithm to reach these statistical radii.

On the assumptions of η and β: As β approaches 1, the ratio $(1 - \beta^{2/\alpha})/\beta$ goes to 0. Therefore, as long as we choose β sufficiently close to 1, the first condition $\frac{\eta C_1 \|\theta_n^0 - \theta^*\|^\alpha \beta}{(\alpha+1)(\alpha+2)} + \beta^{2/\alpha} \geq 1$ is satisfied. For the second assumption $\eta C_1(\alpha + 2)3^\alpha \|\theta_n^0 - \theta^*\|^\alpha \leq \alpha + 1$, it can be simply satisfied as long as we choose the step size η to be sufficiently small.

On the minimum over the iterations: As the results of Theorem 1 indicate, the statistical rates of the ELU iterates are only hold for some $1 \leq k \leq t$ where $t \geq C \log(n/(d\log(1/\delta)))$
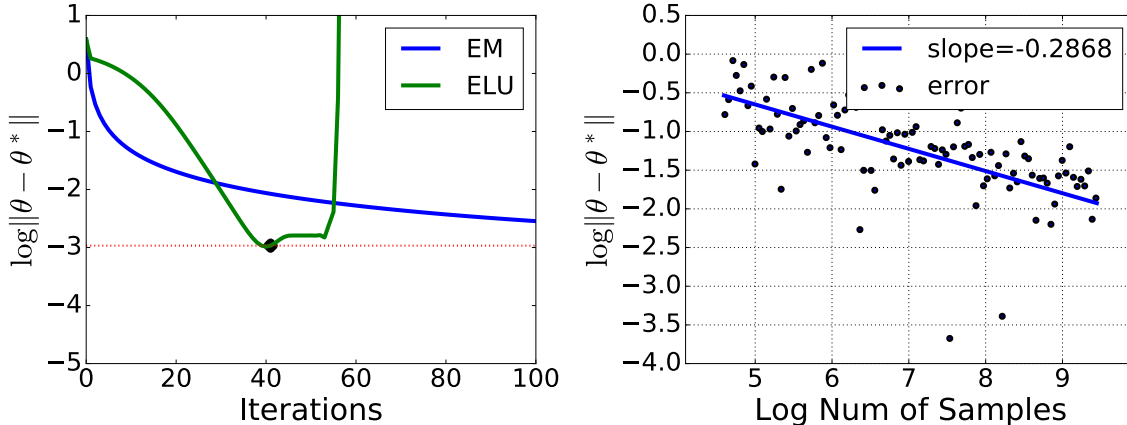
Figure 2. Illustrations for the case with $d = 4$. **Left:** The optimization rates of the EM and ELU. The black diamond shows the ELU iterates with minimum validation error. ELU can converge to the statistical radius with a linear rate then diverge, while EM converge to the statistical radius with a sub-linear rate. **Right:** ELU can find a solution of θ within the statistical radius $\mathcal{O}(n^{-1/4})$. Given the update of σ in ELU algorithm, its statistical radius is directly $\mathcal{O}(n^{-1/2})$.

where C is some universal constant. It means that after reaching the final statistical radii under both the univariate and multivariate settings, the ELU iterates can diverge. This phenomenon is unavoidable in practice due to the nature of the exponentially learning rate of the gradient descent [16]. While it may sound negative, we would like to remark that we can apply early stopping for the ELU iterates via cross-validation with an extra computation of $\mathcal{O}(nd)$ and it does not affect the final computational complexity of the ELU iterates to reach the final statistical radii. See the black diamonds in Figures 1 and 2 for a demonstration of that early stopping.

Minimax optimality of the statistical radii of the ELU iterates: As being demonstrated in Proposition 1 in Appendix B of [7], the statistical radii of $\mathcal{O}(n^{-1/8})$ and $\mathcal{O}(n^{-1/4})$ for the location and scale parameters when $d = 1$ and $\mathcal{O}((d/n)^{1/4})$ and $\mathcal{O}((nd)^{-1/2})$ for the location and scale parameters when $d \geq 2$ are minimax optimal. Therefore, the ELU iterates have both optimal computational complexity $\mathcal{O}(nd\log(n/d))$ and minimax optimal statistical radii.

Experiments: To illustrate the performance of the ELU, we compare the ELU with the EM on both univariate setting (shown in Figure 1) and multivariate setting (shown in Figure 2). We use $\eta = 0.01$, $\beta = 0.8$, and $n = 10^6$ for both cases. 90% of data are used for training and the remaining 10% of data are used for cross-validation. For both cases, the iterates of ELU converge linearly to the statistical radius then start to diverge, while the iterates of EM converge to the statistical radius sub-linearly. For the univariate setting, ELU can find a solution of θ via cross-validation within the statistical radius $\mathcal{O}(n^{-1/8})$, while for the multivariate setting, ELU finds the solution θ within the statistical radius $\mathcal{O}(n^{-1/4})$, which demonstrates the effectiveness of ELU.

Practical implications: An important practical scenario is that the value of θ^* in the symmetric two-component location-scale Gaussian mixtures is generally unknown, namely, we do not know whether we are under the exactly-specified and over-specified settings of these models. However, the faster convergence of the ELU algorithm over the EM algorithm

in the over-specified settings indicates that in practice, we can simultaneously run both the EM and ELU algorithms for solving parameter estimation of these models when we do not know about θ^*. If we observe that the EM iterates converge geometrically fast, it indicates that we are in the exactly-specified settings. On the other hand, if we observe that the EM iterates converge slowly, it suggests that we are in the over-specified settings and we can use the ELU algorithm for solving the parameter estimation under these settings. Since the ELU and EM algorithms have similar per iteration cost, running simultaneously the EM and ELU algorithms only slightly increases the computational complexity comparing to when we run individual algorithm.

Beyond symmetric two-component Gaussian mixtures: In Appendix B, we provide a discussion showing that the ELU algorithm can still be useful in more general settings than the over-specified symmetric two-component location-scale Gaussian mixtures (1). We specifically consider two settings: (i) beyond the isotropic covariance matrix in Appendix B.1; (ii) beyond the symmetric location parameters in Appendix B.2. To the best of our knowledge, the theoretical analysis of optimization algorithms for these settings has not been established before in the literature. In these Appendices, we provide the insight into the fast convergence of the ELU algorithms (via both empirical and theoretical results) for solving parameter estimation of these models.

2.3 Proof sketch of Theorem 1

We now provide a roadmap for the proof of the statistical and computational complexities of the ELU algorithm in Theorem 1. According to the updates in equations (8) and (9), since can view σ_n^t as a function of θ_n^t, it is sufficient to analyze the statistical behaviors of θ_n^t from equation (8) to obtain the statistical behaviors of the update for the scale parameter. Our analysis is based on an application of the general theory built in Theorem 1 of the recent work of Ho et al. [16]. In particular, that theory requires the following two conditions.

The first condition is the *homogeneous condition* of the population version of the function f_n, namely the limit of f_n when n goes to infinity, which is given by:

$$f(\theta) := \mathcal{L}\left(\theta, 1 - \frac{\|\theta\|^2}{d}\right), \tag{10}$$

where $\mathcal{L}(\theta, \sigma) := -\mathbb{E}\left[\log\left(\frac{1}{2}\phi(X; \theta, \sigma^2 I_d) + \frac{1}{2}\phi(X; -\theta, \sigma^2 I_d)\right)\right]$ is the negative population log-likelihood function of model (2). Here, the outer expectation is taken with respect to $X \sim \mathcal{N}(\theta^*, (\sigma^*)^2 I_d)$ and $\theta^* = 0$ and $\sigma^* = 1$.

The second condition is *stability condition*, which had been used extensively in the literature to establish the statistical and computational complexities of optimization algorithms for solving parameter estimation in statistical models [1, 10, 18, 3, 8, 7, 13]. The stability condition is about the uniform concentration bound of $\nabla f_n(\theta)$ around $\nabla f(\theta)$ when $\|\theta - \theta^*\| \leq r$ for any radius $r > 0$.

Statistical rate of ELU iterates based on the homogeneous and stability conditions: To ease our ensuing discussion, we first define formally the homogeneous and stability conditions below.

Definition 1. *(Homogeneous Condition) The function f is homogeneous with constant $\alpha > 0$ in $\mathbb{B}(\theta^*, \rho)$ for some radius ρ if f is locally convex in $\mathbb{B}(\theta^*, \rho)$ and the following conditions hold:*

$$\lambda_{\max}(\nabla^2 f(\theta)) \leq c_1 \|\theta - \theta^*\|^\alpha, \tag{11}$$

$$\|\nabla f(\theta)\| \geq c_2 (f(\theta) - f(\theta^*))^{1 - \frac{1}{\alpha+2}}, \tag{12}$$

for any $\theta \in \mathbb{B}(\theta^*, \rho)$ where c_1 and c_2 are some universal constants.

The specific values of α for the function f are in Lemma 1. The inequality (11) characterizes the growth condition of the maximum eigenvalue of the Hessian matrix of the population function f at θ when θ approaches θ^*. It is different from the traditional smoothness condition of locally strongly convex function when $\alpha = 0$. The inequality (12) is a generalized PL condition and characterizes the local growth condition of the gradient of the population function f in terms of the polynomial functions.

Definition 2. *(Stability Condition) The function f_n is stable with a constant $\gamma \geq 0$ around the function f if there exist a noise function $\varepsilon : \mathbb{N} \times (0, 1] \rightarrow \mathbb{R}^+$ and universal constants $c_3, \rho > 0$ such that*

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla f_n(\theta) - \nabla f(\theta)\| \leq c_3 r^\gamma \varepsilon(n, \delta),$$

for all $r \in (0, \rho)$ with probability $1 - \delta$.

The values of γ and the noise function $\varepsilon(n, \delta)$ for the stability of the function f_n around f are in Lemma 2. The key insight of the uniform concentration bounds in Lemma 2 is that when θ approaches θ^*, the gradient of f_n approaches that of f. Furthermore, they provide tight polynomial growths on the difference between ∇f_n and ∇f. .

Based on the homogeneous condition in Definition 1 and the stability condition in Definition 2, an application of Theorem 2 from [16] to the ELU iterates leads to the following result.

Proposition 1. *Assume that the function f is homogeneous with constant α and the function f_n is stable with constant γ around the function f. Furthermore, the step size η and the scaling parameter β of the ELU algorithm in Algorithm 1 are chosen such that $\frac{\eta c_1 \|\theta_n^0 - \theta^*\|^\alpha \beta}{(\alpha+1)(\alpha+2)} + \beta^{2/\alpha} \geq 1$ and $\eta c_1 (\alpha+2) 3^\alpha \|\theta_n^0 - \theta^*\|^\alpha \leq \alpha + 1$ where α and c_1 are constants in Definition 1. Then, we can find universal constants C_1 and C_2 such that when the sample size n is large enough such that $t \geq C_1 \log(1/\varepsilon(n, \delta))$, with probability $1 - \delta$ the ELU iterates $\{\theta_n^k\}_{k \geq 0}$ for the location parameter have the following statistical bound:*

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^*\| \leq C_2 \left(\varepsilon(n, \delta)\right)^{\frac{1}{\alpha+1-\gamma}}.$$

Given the result of Proposition 1, as long as we can identify the constants α, γ and the noise function $\varepsilon(n, \delta)$ in the homogeneous and stability conditions, we obtain the statistical rates of the ELU iterates in Theorem 1.

Verifying homogeneous condition: We first start with the homogenous condition. We summarize the homogeneous condition of the function f in the following lemma.

Lemma 1. *There exist universal constants $\{C_i\}_{i=1}^4$ and universal constants ρ_1 and ρ_2 such that the following holds:*

(a) When $d = 1$, the function f is locally convex in $\mathbb{B}(\theta^*, \rho_1)$ and for all $\theta \in \mathbb{B}(\theta^*, \rho_1)$ we find that

$$\nabla_\theta^2 f(\theta) \leq C_1 |\theta - \theta^*|^6, \tag{13}$$

$$|\nabla f(\theta)| \geq C_2 (f(\theta) - f(\theta^*))^{7/8}, \tag{14}$$

(b) When $d \geq 2$, the function f is locally convex in $\mathbb{B}(\theta^*, \rho_2)$ and for all $\theta \in \mathbb{B}(\theta^*, \rho_2)$, we obtain

$$\lambda_{\max}(\nabla_\theta^2 f(\theta)) \leq C_1 \|\theta - \theta^*\|^2, \tag{15}$$

$$\|\nabla f(\theta)\| \geq C_2 (f(\theta) - f(\theta^*))^{3/4}, \tag{16}$$

Proof of Lemma 1 is in Appendix A.1. The results of Lemma 1 indicate that the population loss function f is locally convex but not strongly convex around θ^* for all dimension $d \geq 1$. Furthermore, the function f is flatter in one dimension than in $d \geq 2$ dimension. Finally, the results of Lemma 1 indicate that the function f is homogeneous with constant $\alpha = 6$ when $d = 1$ and with constant $\alpha = 2$ when $d \geq 2$.

Verifying stability condition: We now move to the stability condition of the function f_n around the function f, which is summarized in the following lemma.

Lemma 2. *(a) When $d = 1$, there exists universal constants $\{C_i\}_{i=1}^2$ and ρ_1' such that as long as $n \geq C_1 \log(1/\delta)$ with probability $1 - \delta$ for any $r \in (0, \rho_1')$ we find that*

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla f_n(\theta) - \nabla f(\theta)\| \leq C_2 r \sqrt{\frac{\log(1/\delta)}{n}}, \tag{17}$$

Furthermore, if we have $r \leq C_3 n^{-1/16}$ for some universal constant C_3, then there exists the universal constant C_4 such that the concentration bound (17) can be improved as follows:

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla f_n(\theta) - \nabla f(\theta)\| \leq C_4 r^3 \frac{\log^{10}(5n/\delta)}{\sqrt{n}}. \tag{18}$$

(b) When $d \geq 2$, there exists universal constants $\{C_i'\}_{i=1}^2$ and ρ_2' such that as long as $n \geq C_1' d \log(1/\delta)$ with probability $1 - \delta$ for any $r \in (0, \rho_2')$ we find that

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla f_n(\theta) - \nabla f(\theta)\| \leq C_2' r \sqrt{\frac{d + \log(1/\delta)}{n}}. \tag{19}$$

Proof of Lemma 2 is in Appendix A.2. The results of Lemma 2 indicate that the function f_n is stable with constant $\gamma = 1$ and the noise function $\varepsilon(n, \delta) = \sqrt{(d + \log(1/\delta))/n}$ when $d \geq 2$. When $d = 1$, there are two phases of the stability condition. In Phase 1, when the radius $r > C_3 n^{-1/16}$ where C_3 is constant in Lemma 2, the function f_n is stable with constant $\gamma = 1$ and the noise function $\varepsilon(n, \delta) = \sqrt{\log(1/\delta)/n}$. In Phase 2, when the radius $r \leq C_3 n^{-1/16}$, the function f_n is stable with constant $\gamma = 3$ and the noise function $\varepsilon(n, \delta) = \log^{10}(5n/\delta)/\sqrt{n}$. Given these two phases of stability condition, an application of Proposition 1 indicates that the ELU iterates $\{\theta_n^k\}_{k \geq 0}$ first reach the statistical radius $n^{-1/12}$ in Phase 1 after a logarithmic number of iterations $\mathcal{O}(\log(n))$. Then in Phase 2, these iterates continue reaching the final statistical radius $n^{-1/8}$ after $\mathcal{O}(\log(n))$ iterations. With the relation that $(\sigma_n^k)^2 = \frac{\sum_{i=1}^n \|X_i\|^2}{nd} - \frac{\|\theta_n^k\|^2}{d}$, we can conclude the proof for Theorem 1.

3 Conclusion

In the paper, we propose the Exponential Location Update (ELU) algorithm for solving parameter estimation of the over-specified settings of the symmetric two-component Gaussian mixtures. When $d = 1$, we demonstrate that the ELU iterates for solving the location

and scale parameters respectively reach the minimax optimal statistical radii $\mathcal{O}(n^{-1/8})$ and $\mathcal{O}(n^{-1/4})$ within the true location and scale parameters after a logarithmic number of iterations $\mathcal{O}(\log(n))$. Notably, it is significantly cheaper than the polynomial number of iterations $\mathcal{O}(n^{-3/4})$ of the EM algorithm to reach the similar final statistical radii. When $d = 2$, we prove that the statistical radii of the ELU iterates for the location and scale parameters are $\mathcal{O}((d/n)^{1/4})$ and $\mathcal{O}((nd)^{-1/2})$ and these radii are achieved after $\log(n/d)$ number of iterations. It is also significantly cheaper than $\mathcal{O}(\sqrt{n/d})$ number of iterations of the EM algorithm. As a consequence, the ELU algorithm has an optimal computational complexity for solving parameter estimation in the over-specified settings of the symmetric two-component Gaussian mixtures.

4 Acknowledgment

Supplement for "Beyond EM Algorithm on Over-Specified Symmetric Two-Component Location-Scale Gaussian Mixtures"

In this supplementary material, we present proofs of Theorem 1 in Appendix A. Then, in Appendix B, we provide discussion showing that the ELU algorithm can still be useful in more general settings than the over-specified symmetric two-component location-scale Gaussian mixtures. Finally, we provide proofs for the remaining results in the paper in Appendix C.

A   Proof of Theorem 1

We first provide proofs for Lemma 1 regarding the homogeneous condition of the population loss function f in Appendix A.1. Then, we establish the stability condition of the function f_n around f in Appendix A.2. To ease the ensuing discussion, we recall the definitions of the function f and f_n as follows:

$$f_n(\theta) = \mathcal{L}_n(\theta, \frac{1}{nd} \sum_{i=1}^{n} \|X_i\|^2 - \frac{\|\theta\|^2}{d}),$$

$$f(\theta) = \mathcal{L}\left(\theta, 1 - \frac{\|\theta\|^2}{d}\right),$$

where $\mathcal{L}_n$ is the negative sample log-likelihood function of the symmetric two-component location-scale Gaussian mixtures in equation (3) and

$$\mathcal{L}(\theta, \sigma) = -\mathbb{E}\left[\log\left(\frac{1}{2}\phi(X; \theta, \sigma^2 I_d) + \frac{1}{2}\phi(X; -\theta, \sigma^2 I_d)\right)\right]$$

is the negative population log-likelihood function of model (2). Here, the outer expectation is taken with respect to $X \sim \mathcal{N}(\theta^*, (\sigma^*)^2 I_d)$ and $\theta^* = 0$ and $\sigma^* = 1$.

A.1 Proof of Lemma 1

Direct calculation yields that

$$f(\theta) = \log 2 + \frac{d}{2}\log(2\pi) + \frac{d}{2}\log\left(1 - \frac{\|\theta\|^2}{d}\right) + \frac{d + \|\theta\|^2}{2(1 - \frac{\|\theta\|^2}{d})}$$
$$- \mathbb{E}\left[\log\left(\exp\left(-\frac{X^\top\theta}{1 - \frac{\|\theta\|^2}{d}}\right) + \exp\left(\frac{X^\top\theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right)\right].$$

To simplify the calculation of $f(\theta)$, we perform a change of coordinates via an orthogonal matrix R such that $R\theta = \|\theta\|e_1$ where e_1 denotes the first canonical basis in dimension d. By denoting $V = RX$, since $X \sim \mathcal{N}(\theta^*, I_d)$ where $\theta^* = 0$ we have $V = (V_1, \ldots, V_d) \sim \mathcal{N}(0, I_d)$. Therefore, we can rewrite the function f as follows:

$$f(\theta) = \log 2 + \frac{d}{2}\log(2\pi) + \frac{d}{2}\log\left(1 - \frac{\|\theta\|^2}{d}\right) + \frac{d + \|\theta\|^2}{2\left(1 - \frac{\|\theta\|^2}{d}\right)}$$
$$- \mathbb{E}\left[\log\left(\exp\left(-\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right) + \exp\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right)\right],$$

where the outer expectation is taken with respect to $V_1 \sim \mathcal{N}(0, 1)$. Similarly, direct calculation of the gradient of the function f yields that

$$\nabla_\theta f(\theta) = \frac{\theta\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2} - \mathbb{E}\left[\frac{V_1\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\frac{\theta}{\|\theta\|}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]$$
$$= \frac{\theta\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\left(1 - \mathbb{E}\left[\frac{V_1}{\|\theta\|}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]\right),$$

and

$$\nabla_\theta^2 f(\theta) = \frac{\left(\left(1 + \frac{\|\theta\|^2}{d}\right)I + 2\frac{\theta\theta^\top}{d}\right)\left(1 - \frac{\|\theta\|^2}{d}\right) + \frac{4}{d}\left(1 + \frac{\|\theta\|^2}{d}\right)\theta\theta^\top}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3}\left(1 - \mathbb{E}\left[\frac{V_1}{\|\theta\|}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]\right)$$
$$- \frac{\theta\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\mathbb{E}\left[-\frac{V_1\theta^\top}{\|\theta\|^3}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right) + \frac{V_1^2\left(1 + \frac{\|\theta\|^2}{d}\right)\theta^\top}{\|\theta\|^2\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\text{sech}^2\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right].$$

We now divide our proof of this lemma into two settings: univariate setting ($d = 1$) and multivariate setting ($d \geq 2$).

A.1.1 Univariate setting

Proof for the convexity of the function f: We first show that the function f is convex w.r.t θ when $|\theta| \leq 0.5$. As we perform the change of the coordinate, we assume without loss of generality that $\theta \geq 0$ throughout. Note that, $\nabla f(\theta) = c(\theta)\theta$ where $c(\theta)$ is the following scalar function:

$$c(\theta) = \frac{1 + \theta^2}{(1 - \theta^2)^2}\left(1 - \mathbb{E}\left[\frac{V_1}{\theta}\tanh\left(\frac{V_1\theta}{1 - \theta^2}\right)\right]\right).$$

If for all $0 \leq \theta \leq 0.5$, $c(\theta) > 0$, then we directly obtain that $f(\theta)$ is convex w.r.t θ. Indeed, using the fact that $x \tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}$ for all $x \in \mathbb{R}$, we have the following inequality:

$$1 - \mathbb{E}\left[\frac{V_1}{\theta} \tanh\left(\frac{V_1\theta}{1-\theta^2}\right)\right]$$
$$\leq 1 - \frac{1-\theta^2}{\theta^2}\left(\frac{\theta^2}{(1-\theta^2)^2} - \frac{\theta^4}{(1-\theta^2)^4} + \frac{2\theta^6}{(1-\theta^2)^6} - \frac{17\theta^8}{3(1-\theta^2)^8} + \frac{3\theta^{10}}{(1-\theta^2)^{10}}\right)$$
$$= \frac{\theta^6(12 - 13\theta^2 - 30\theta^4 + 48\theta^6 - 27\theta^8 - 8\theta^{10} + \theta^{12})}{(1-x^2)^9}.$$

When $0 \leq \theta \leq 0.5$, we can see $(12 - 13\theta^2 - 30\theta^4 + 48\theta^6 - 27\theta^8 - 8\theta^{10} + \theta^{12}) \geq 0$, which means $1 - \mathbb{E}\left[\frac{V_1}{\theta} \tanh\left(\frac{V_1\theta}{1-\theta^2}\right)\right] \geq 0$. Therefore, $c(\theta) \geq 0$, which leads to the convexity of the function f when $|\theta| \leq 0.5$.

Proof for inequality (13): Now, we provide the proof for inequality (13). Note that $x \tanh(x) \geq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315}$ and $\operatorname{sech}^2(x) \geq 1 - x^2 + \frac{2x^4}{3} - \frac{17x^6}{45}$ for all $x \in \mathbb{R}$. Hence, we find that

$$\frac{2\theta(\theta^2+3)}{(1-\theta^2)^3}\mathbb{E}\left[X \tanh\left(\frac{X\theta}{1-\theta^2}\right)\right] \geq \frac{2(\theta^2+3)}{(1-\theta^2)^2}\left[\frac{\theta^2}{(1-\theta^2)^2} - \frac{\theta^4}{(1-\theta^2)^4} + \frac{2\theta^6}{(1-\theta^2)^6} - \frac{17\theta^8}{3(1-\theta^2)^8}\right],$$
$$\frac{(1+\theta^2)^2}{(1-\theta^2)^4}\mathbb{E}\left[X^2\operatorname{sech}^2\left(\frac{X\theta}{1-\theta^2}\right)\right] \geq \frac{(1+\theta^2)^2}{(1-\theta^2)^4}\left[1 - \frac{3\theta^2}{(1-\theta^2)^2} + \frac{10\theta^4}{(1-\theta^2)^4} - \frac{119\theta^6}{3(1-\theta^2)^6}\right].$$

As a result, we have that

$$\frac{2\theta(\theta^2+3)}{(1-\theta^2)^3}\mathbb{E}\left[X \tanh\left(\frac{X\theta}{1-\theta^2}\right)\right] + \frac{(1+\theta^2)^2}{(1-\theta^2)^4}\mathbb{E}\left[X^2\operatorname{sech}^2\left(\frac{X\theta}{1-\theta^2}\right)\right]$$
$$\geq \frac{3\theta^4 + 8\theta^2 + 1}{(1-\theta^2)^4} - \frac{5\theta^6 + 12\theta^4 + 3\theta^2}{(1-\theta^2)^6} + \frac{14\theta^8 + 32\theta^6 + 10\theta^4}{(1-\theta^2)^8} - \frac{153\theta^{10} + 340\theta^8 + 119\theta^6}{3(1-\theta^2)^{10}},$$

and

$$\nabla_\theta^2 f(\theta) \leq \frac{\theta^6(14 + 418\theta^2 + 396\theta^4 - 312\theta^6 + 105\theta^8 - 6\theta^{10} - 3\theta^{12})}{(1-\theta^2)^{10}}.$$

With the above expression, we know for any $\rho^2 < 1$, there exists absolute constant C_1 that only depends on ρ, such that for all $\theta^2 \leq \rho^2$, we always have $\nabla_\theta^2 f(\theta) \leq C_1\theta^6$, which concludes the proof of the inequality (13).

Proof for inequality (14): Recall when $|\theta| \leq 0.5$, we have that $1 - \mathbb{E}\left[\frac{V_1}{\theta} \tanh\left(\frac{V_1\theta}{1-\theta^2}\right)\right] \geq C\theta^6$ where C is some absolute constant. Hence, we know $|\nabla f(\theta)| \geq L_1|\theta|^7$ when $|\theta| \leq 0.5$ where L_1 is some absolute constant.

Meanwhile, recall when $d = 1$, we have

$$f(\theta) - f(\theta^*) = \frac{1}{2}\log(1-\theta^2) + \frac{\theta^2}{1-\theta^2} - \mathbb{E}\left[\log\left(\frac{1}{2}\left(\exp\left(-\frac{X\theta}{1-\theta^2}\right) + \exp\left(\frac{X\theta}{1-\theta^2}\right)\right)\right)\right].$$

Use the fact that $\log(\frac{1}{2}(\exp(x) + \exp(-x))) \geq \frac{x^2}{2} - \frac{x^4}{12} + \frac{x^6}{45} - \frac{17x^8}{2520}$ for all $x \in \mathbb{R}$, we find that

$$\mathbb{E}\left[\log\left(\frac{1}{2}\left(\exp\left(-\frac{X\theta}{1-\theta^2}\right) + \exp\left(\frac{X\theta}{1-\theta^2}\right)\right)\right)\right]$$

14

$$\geq \frac{\theta^2}{2(1-\theta^2)^2} - \frac{\theta^4}{4(1-\theta^2)^4} + \frac{\theta^6}{3(1-\theta^2)^6} - \frac{17\theta^8}{24(1-\theta^2)^8}.$$

Meanwhile, note that

$$\log(1+x) \geq x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4}, \quad \forall x \geq 0.$$

Take $x = \frac{\theta^2}{1-\theta^2}$, we have

$$\log(1-\theta^2) \leq -\frac{\theta^2}{1-\theta^2} + \frac{\theta^4}{2(1-\theta^2)^2} - \frac{\theta^6}{3(1-\theta^2)^3} + \frac{\theta^8}{4(1-\theta^2)^4}$$

Combined these inequalities, we have that

$$f(\theta) - f(\theta^*) \leq \frac{\theta^8(\theta^4 + 6\theta^2 + 2)}{42(1-\theta^2)^6}.$$

Hence, we know for any $\rho^2 < 1$, there exists absolute constant L_2 only depending on ρ, such that for all $\theta^2 \leq \rho^2$, $f(\theta) - f(\theta^*) \leq L_2\theta^8$, which concludes the proof of the inequality (14) with the bound $|\nabla f(\theta)| \geq L_1\theta^7$.

A.1.2 Multivariate setting

Under the multivariate setting, we first show $f(\theta)$ is convex w.r.t θ when $\|\theta\|^2 < d/3$. Note that $\nabla_\theta f(\theta) = c(\theta)\theta$ where $c(\theta)$ is a scalar function depends on θ. Hence, if $c(\theta) > 0, \forall 0 < \|\theta\| < d/3$, we can directly obtain that $f(\theta)$ is convex w.r.t θ. It is equivalent to show that

$$\mathbb{E}\left[\frac{V_1}{\|\theta\|} \tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \leq 1. \tag{20}$$

From the inequality $x\tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15}$ for all $x \in \mathbb{R}$, we obtain that

$$1 - \mathbb{E}\left[\frac{V_1}{\|\theta\|} \tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]$$

$$\geq \|\theta\|^2 \left(-\frac{1}{d\left(1 - \frac{\|\theta\|^2}{d}\right)} + \frac{1}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3} - \frac{2\|\theta\|^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^5}\right)$$

$$= \|\theta\|^2 \frac{(d-1)\left(1 - \frac{3\|\theta\|^2}{d}\right) - \frac{3\|\theta\|^4}{d^2} + \frac{\|\theta\|^6}{d^3}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^5}.$$

Hence, as long as $\|\theta\|^2 < d/3$, we have $c(\theta) > 0$, which shows $f(\theta)$ is convex w.r.t θ.

Proof of the inequality (15): We denote

$$\nabla_\theta^2 f(\theta) = \lambda_0 I + \lambda_1 \theta\theta^\top,$$

15

where

$$\lambda_0 = \left(1 - \mathbb{E}\left[\frac{V_1}{\|\theta\|}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]\right)\frac{1 + \frac{\|\theta\|^2}{d}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2},$$

$$\lambda_1 = \frac{\frac{6}{d} + \frac{2\|\theta\|^2}{d^2}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3} - \frac{\frac{6\|\theta\|^2}{d} + \frac{3\|\theta\|^4}{d^2} - 1}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3\|\theta\|^3}\mathbb{E}\left[V_1\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]$$

$$- \frac{\left(1 + \frac{\|\theta\|^2}{d}\right)^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^4\|\theta\|^2}\mathbb{E}\left[V_1^2\mathrm{sech}^2\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right].$$

As a result, $\nabla_\theta^2 f(\theta)$ only has two different eigenvalues λ_0 and $\lambda_0 + \lambda_1\|\theta\|^2$. We then consider λ_0 and λ_1 separately. For λ_0, note that $x\tanh(x) \geq x^2 - \frac{x^4}{3}$, thus

$$\mathbb{E}\left[\frac{V_1}{\|\theta\|}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \geq \frac{1 - \frac{\|\theta\|^2}{d}}{\|\theta\|^2}\left(\frac{\|\theta\|^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2} - \frac{\|\theta\|^4}{\left(1 - \frac{\theta^2}{d}\right)^4}\right) = \frac{1}{1 - \frac{\|\theta\|^2}{d}} - \frac{\|\theta\|^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3},$$

and

$$1 - \mathbb{E}\left[\frac{V_1}{\|\theta\|}\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \leq \|\theta\|^2\frac{1 - \frac{1}{d}\left(1 - \frac{\|\theta\|^2}{d}\right)^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3}$$

Hence, for all $\rho^2 < d$, there exists absolute constant L_1, such that for all $\|\theta\|^2 \leq \rho^2$, $\lambda_0 \leq L_1\|\theta\|^2$.

As f is convex, if $\lambda_1 \leq 0$, then the inequality (15) naturally holds. Otherwise, use the fact that $\mathrm{sech}^2(x) \geq 1 - x^2$, we have that

$$\mathbb{E}\left[V_1^2\mathrm{sech}^2\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \geq 1 - \frac{3\|\theta\|^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}.$$

Hence,

$$\frac{\frac{6\|\theta\|^2}{d} + \frac{3\|\theta\|^4}{d^2} - 1}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3\|\theta\|^3}\mathbb{E}\left[V_1\tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right] + \frac{\left(1 + \frac{\|\theta\|^2}{d}\right)^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^4\|\theta\|^2}\mathbb{E}\left[V_1^2\mathrm{sech}^2\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right]$$

$$\geq \frac{\frac{6\|\theta\|^2}{d} + \frac{3\|\theta\|^4}{d^2} - 1}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3\|\theta\|^2}\left(\frac{1}{1 - \frac{\|\theta\|^2}{d}} - \frac{\|\theta\|^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^3}\right) + \frac{\left(1 + \frac{\|\theta\|^2}{d}\right)^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^4\|\theta\|^2}\left(1 - \frac{3\|\theta\|^2}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\right)$$

$$\geq \frac{\frac{8}{d} + \frac{4\|\theta\|^2}{d^2}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^4} - \frac{2 + \frac{12\|\theta\|^2}{d} + \frac{6\|\theta\|^4}{d}}{\left(1 - \frac{\|\theta\|^2}{d^2}\right)^6}.$$

and

$$\lambda_1 \leq -\frac{\frac{2}{d} + \frac{8\|\theta\|^2}{d^2} + \frac{2\|\theta\|^4}{d^3}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^4} + \frac{2 + \frac{12\|\theta\|^2}{d} + \frac{6\|\theta\|^4}{d^2}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^6}.$$

16

Hence, there exists absolute constant L_2, such that for all $\|\theta\|^2 \le \rho^2$, when $\lambda_1 \ge 0$, $\lambda_1 \le L_2$, which concludes the proof of inequality (15).

Proof of the inequality (16): Recall that, as long as $\|\theta\|^2 \le d/3$, there exists absolute constant C, such that

$$1 - \mathbb{E}\left[\frac{V_1}{\|\theta\|} \tanh\left(\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \ge C\|\theta\|^2.$$

Hence, for all $\rho^2 < d/3$, there exists absolute constant L_1 depends on ρ, such that for all $\theta^2 \le \rho^2$, we have $\|\nabla_\theta f(\theta)\| \ge L_1\|\theta\|^3$. Meanwhile, when $d \ge 2$, we have

$$f(\theta) - f(\theta^*) = \frac{d}{2} \log\left(1 - \frac{\|\theta\|^2}{d}\right) + \frac{d + \|\theta\|^2}{2\left(1 - \frac{\|\theta\|^2}{d}\right)} - \frac{d}{2}$$

$$- \mathbb{E}\left[\log\left(\frac{1}{2}\left(\exp\left(-\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right) + \exp\left(-\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right)\right)\right].$$

Use the fact that $\log\left(\frac{1}{2}(\exp(x) + \exp(-x))\right) \ge \frac{x^2}{2} - \frac{x^4}{12}$ for all $x \in \mathbb{R}$, we have that

$$\mathbb{E}\left[\log\left(\frac{1}{2}\left(\exp\left(-\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right) + \exp\left(-\frac{V_1\|\theta\|}{1 - \frac{\|\theta\|^2}{d}}\right)\right)\right)\right] \ge \frac{\|\theta\|^2}{2\left(1 - \frac{\|\theta\|^2}{d}\right)^2} - \frac{\|\theta\|^4}{4\left(1 - \frac{\|\theta\|^2}{d}\right)^4}.$$

Meanwhile, note that

$$\log(1 + x) \ge x - \frac{x^2}{2}, \qquad \forall x \ge 0.$$

Take $x = \frac{\|\theta\|^2}{d\left(1 - \frac{\|\theta\|^2}{d}\right)}$, we have

$$\log\left(1 - \frac{\|\theta\|^2}{d}\right) \le -\frac{\|\theta\|^2}{d\left(1 - \frac{\|\theta\|^2}{d}\right)} + \frac{\|\theta\|^4}{d^2\left(1 - \frac{\|\theta\|^2}{d}\right)^2}.$$

Combine the inequalities, we have that

$$f(\theta) - f(\theta)^* \le \frac{\|\theta\|^2}{2\left(1 - \frac{\|\theta\|^2}{d}\right)} + \frac{\|\theta\|^4}{2d\left(1 - \frac{\|\theta\|^2}{d}\right)^2} - \frac{\|\theta\|^2}{2\left(1 - \frac{\|\theta\|^2}{d}\right)^2} + \frac{\|\theta\|^4}{4\left(1 - \frac{\|\theta\|^2}{d}\right)^4}$$

$$\le \frac{\|\theta\|^4}{d\left(1 - \frac{\|\theta\|^2}{d}\right)^2} + \frac{\|\theta\|^4}{4\left(1 - \frac{\|\theta\|^2}{d}\right)^4}.$$

Hence, for all $\rho^2 < d$, there exists absolute constant L_2 depends on ρ, such that $\forall \theta^2 \le \rho^2$, we have $f(\theta) - f(\theta)^* \le L_2\|\theta\|^4$. Combined with the inequality that $\|\nabla_\theta f(\theta)\| \ge L_1\|\theta\|^3$, we conclude the proof for inequality (16).

A.2   Proof of Lemma 2

In this appendix, we first provide proofs for uniform concentration bounds in equations (17) and (19) in Appendix A.2.1. Then, we provide proof for the improved concentration bound (18) in Appendix A.2.2. In this proof, the values of universal constants can change from line to line.

A.2.1 Proofs of equations (17) and (19)

From the definition of $\mathcal{L}_n(\theta, \sigma)$, simple calculation yields that

$$\mathcal{L}_n(\theta, \sigma) = \log(2(\sqrt{2\pi})^d) + d\log(\sigma) + \frac{1}{n}\sum_{i=1}^{n}\frac{\|X_i\|^2 + \|\theta\|^2}{2\sigma^2}$$
$$-\frac{1}{n}\sum_{i=1}^{n}\log\left(\exp\left(\frac{X_i^\top\theta}{\sigma^2}\right) + \exp\left(-\frac{X_i^\top\theta}{\sigma^2}\right)\right).$$

Therefore, we obtain that

$$\nabla_\theta f_n(\theta) = \frac{\theta\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 + \frac{\|\theta\|^2}{d}\right)}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2}$$
$$-\frac{1}{n}\sum_{i=1}^{n}\frac{X_i\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right) + \frac{2X_i^\top\theta\theta}{d}}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2}\tanh\left(\frac{X_i^\top\theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right).$$

Recall that, we have

$$\nabla_\theta f(\theta) = \frac{\theta\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2} - \mathbb{E}\left[\frac{X(1 - \frac{\|\theta\|^2}{d}) + 2\frac{X^\top\theta\theta}{d}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\tanh\left(\frac{X^\top\theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right].$$

An application of the triangle inequality leads to

$$\sup_{\theta\in\mathbb{B}(\theta^*,r)}\|\nabla_\theta f_n(\theta) - \nabla_\theta f(\theta)\| \leq T_1 + T_2 + T_3,$$

where we define the terms T_1, T_2, T_3 as follows:

$$T_1 := \sup_{\theta\in\mathbb{B}(\theta^*,r)}\|\theta\|\left|\frac{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 + \frac{\|\theta\|^2}{d}\right)}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2} - \frac{1 + \frac{\|\theta\|^2}{d}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2}\right|,$$

$$T_2 := \sup_{\theta\in\mathbb{B}(\theta^*,r)}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top\theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right)\right.$$
$$\left.-\mathbb{E}\left[\frac{X}{1 - \frac{\|\theta\|^2}{d}}\tanh\left(\frac{X^\top\theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right]\right\|,$$

$$T_3 := \sup_{\theta\in\mathbb{B}(\theta^*,r)}2\|\theta\|\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top\theta}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top\theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right)\right.$$
$$\left.-\mathbb{E}\left[\frac{X^\top\theta}{d\left(1 - \frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X^\top\theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right]\right|.$$

From the standard chi-square concentration [27], we know

$$\mathbb{P}\left(\left|\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - 1\right| \geq \sqrt{\frac{8\log(1/\delta)}{nd}}\right) \leq \delta.$$

18

To ease the ensuing proof argument, from now, we always condition on the events $\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 \in \left[1 - \sqrt{\frac{8\log(1/\delta)}{nd}}, 1 + \sqrt{\frac{8\log(1/\delta)}{nd}}\right]$ in the following analysis. Furthermore, we assume $r^2 < d$, which is required by the condition $\sigma^2 > 0$ in the population update.

Bound for T_1: For the term T_1, when $\|\theta - \theta^*\| \le r$ we have that

$$
\left| \frac{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 + \frac{\|\theta\|^2}{d}\right)}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2} - \frac{1 + \frac{\|\theta\|^2}{d}}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2} \right|
$$

$$
= \left| \frac{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 + \frac{\|\theta\|^2}{d}\right)\left(1 - \frac{\|\theta\|^2}{d}\right)^2 - \left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2\left(1 - \frac{\|\theta\|^2}{d}\right)^2} \right|
$$

$$
\le \left| \frac{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - 1\right)\left(1 + \frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{2\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2} \right| \left(1 + \frac{\|\theta\|^2}{d}\right) + \left| \frac{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - 1}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2} \right|
$$

$$
\le \frac{\sqrt{\frac{8\log(1/\delta)}{nd}}\left(2 + \sqrt{\frac{8\log 1/\delta}{nd}} - \frac{2\|\theta\|^2}{d}\right)\left(1 + \frac{\|\theta\|^2}{d}\right)}{\left(1 - \frac{\|\theta\|^2}{d}\right)^2} + \frac{\sqrt{\frac{8\log(1/\delta)}{nd}}}{\left(1 - \sqrt{\frac{8\log(1/\delta)}{nd}} - \frac{\|\theta\|^2}{d}\right)^2}
$$

$$
\le C\sqrt{\frac{\log(1/\delta)}{nd}},
$$

where C is some universal constant. Therefore, we obtain that

$$
T_1 \le Cr\sqrt{\frac{\log(1/\delta)}{nd}}. \tag{21}
$$

Bound for T_2: For the term T_2, the variational characterization of vector norm shows

$$
T_2 = \max_{\|u\|=1} \sup_{\theta \in \mathbb{B}(\theta^*,r)} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X^\top \theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right) \right.
$$
$$
\left. - \mathbb{E}\left[\frac{X^\top u}{1 - \frac{\|\theta\|^2}{d}}\tanh\left(\frac{X^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \right|.
$$

With a standard discretization arguments (e.g. [27, Chapter 6]), assume U is a $1/8$ cover of the $\mathcal{S}^{d-1}$ whose cardinality is at most 17^d, we know

$$
T_2 \le 2\sup_{u \in U} \sup_{\theta \in \mathbb{B}(\theta^*,r)} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X^\top \theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right) \right.
$$
$$
\left. - \mathbb{E}\left[\frac{X^\top u}{1 - \frac{\|\theta\|^2}{d}}\tanh\left(\frac{X^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \right|.
$$

We then bound the RHS with the sum of the following terms:

$$
T_{21} := \sup_{u \in U} \sup_{\theta \in \mathbb{B}(\theta^*,r)} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X^\top \theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right) \right.
$$

19

$$-\sum_{i=1}^{n} \frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X_i^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right)\Bigg|,$$

$$T_{22} := \sup_{u\in U}\sup_{\theta\in\mathbb{B}(\theta^*,r)} \left|\sum_{i=1}^{n}\left[\frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right) - \frac{X_i^\top u}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right)\right]\right|.$$

$$T_{23} := \sup_{u\in U}\sup_{\theta\in\mathbb{B}(\theta^*,r)} \left|\sum_{i=1}^{n}\frac{X_i^\top u}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right) - \mathbb{E}\left[\frac{X^\top u}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right)\right]\right|.$$

For the term T_{21}, note that $\tanh(x)$ is 1-Lipschitz, we have

$$T_{21} \leq \sup_{u\in U}\sup_{\theta\in\mathbb{B}(\theta^*,r)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)}\right|\left|\frac{X_i^\top \theta}{\sum_{i=1}^{n}\frac{1}{nd}\|X_i\|^2 - \frac{\|\theta\|^2}{d}} - \frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right|$$

$$\leq \frac{C\sqrt{\frac{\log(1/\delta)}{nd}}}{n}\sup_{u\in U}\sup_{\theta\in\mathbb{B}(\theta^*,r)}\sum_{i=1}^{n}\frac{\left|(X_i^\top u)(X_i^\top \theta)\right|}{\left|\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)^2\left(1-\frac{\|\theta\|^2}{d}\right)\right|}$$

$$\leq Cr\sqrt{\frac{\log(1/\delta)}{nd}}\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\right)$$

$$\leq Cr\sqrt{\frac{d\log(1/\delta)}{n}}.$$

For the term T_{22}, note that $|\tanh(x)| \leq |x|$, hence

$$T_{22} \leq \sup_{u\in U}\sup_{\theta\in\mathbb{B}(\theta^*,r)} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i^\top u}{\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} - \frac{X_i^\top u}{1-\frac{\|\theta\|^2}{d}}\right|\left|\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right|$$

$$\leq C\sqrt{\frac{\log(1/\delta)}{nd}}\sup_{u\in U}\sup_{\theta\in\mathbb{B}(\theta^*,r)}\sum_{i=1}^{n}\frac{\left|(X_i^\top u)(X_i^\top \theta)\right|}{\left|\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)\left(1-\frac{\|\theta\|^2}{d}\right)^2\right|}$$

$$\leq Cr\sqrt{\frac{\log(1/\delta)}{nd}}\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\right)$$

$$\leq Cr\sqrt{\frac{d\log(1/\delta)}{n}}.$$

For the term of T_{23}, with a symmetrization arguments [27], we find that

$$\mathbb{E}\left[\exp\left(\lambda\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top u}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right) - \mathbb{E}\left[\frac{X^\top u}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right)\right]\right|\right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(\lambda\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2}{n}\sum_{i=1}^{n}\frac{\varepsilon_i X_i^\top u}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right)\right|\right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i X_i^\top u\tanh\left(\frac{X_i^\top \theta}{1-\frac{\|\theta\|^2}{d}}\right)\right|\right)\right],$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. As $\tanh(x)$ is 1-Lipschitz with $\tanh(0) = 0$, with Ledoux-Talagrand contraction inequality [27], we obtain that

$$\mathbb{E}\left[\exp\left(C\lambda \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i X_i^\top u \tanh\left(\frac{X_i^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right) \right| \right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i \frac{X_i^\top u X_i^\top \theta}{1 - \frac{\|\theta\|^2}{d}} \right| \right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i u^\top X_i X_i^\top \theta \right| \right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda r \left\| \frac{\varepsilon_i X_i X_i^\top}{n} \right\|_{\mathrm{op}}\right)\right].$$

Use the method identical to the proof of the bound (63b) in [8], we obtain that

$$\mathbb{E}\left[\exp\left(C\lambda r \left\| \frac{\varepsilon_i X_i X_i^\top}{n} \right\|_{\mathrm{op}}\right)\right] \leq 2 \cdot 17^d \exp\left(\frac{C\lambda^2 r^2}{n}\right), \quad \forall |\lambda| \leq \frac{n}{Cr}.$$

Hence, we find that

$$\mathbb{E}\left[\exp\left(\lambda \sup_{u \in U} \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i \frac{X_i^\top u}{1 - \frac{\|\theta\|^2}{d}} \tanh\left(\frac{X_i^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right) - \mathbb{E}\left[\frac{X^\top}{1 - \frac{\|\theta\|^2}{d}} \tanh\left(\frac{X^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \right| \right)\right]$$

$$\leq 2 \cdot 17^{2d} \exp\left(\frac{C\lambda^2 r^2}{n}\right), \quad \forall |\lambda| \leq \frac{n}{Cr}.$$

With the Chernoff method, we can obtain that $T_{23} \leq Cr\sqrt{\frac{d \log 1/\delta}{n}}$ as long as $n \geq Cd\log(1/\delta)$. Putting these results together leads to the following bound for T_2:

$$T_2 \leq Cr\sqrt{\frac{d \log 1/\delta}{n}}, \tag{22}$$

as long as $n \geq Cd\log(1/\delta)$.

Bound for T_3: For the term T_3, it is sufficient to consider the term

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \left| \frac{1}{n} \sum_{i=1}^n \frac{X_i^\top \theta}{d\left(\frac{1}{nd}\sum_{i=1}^n \|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X_i^\top \theta}{\frac{1}{nd}\sum_{i=1}^n \|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right) \right.$$

$$\left. - \mathbb{E}\left[\frac{X^\top \theta}{d\left(1 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X^\top \theta}{1 - \frac{\|\theta\|^2}{d}}\right)\right] \right|.$$

We upper bound this term with the summation of the following terms:

$$T_{31} := \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left| \frac{1}{n} \sum_{i=1}^n \frac{X_i^\top \theta}{d\left(\frac{1}{nd}\sum_{i=1}^n \|X_i\|^2 - \frac{\|\theta\|^2}{d}\right)} \tanh\left(\frac{X_i^\top \theta}{\frac{1}{nd}\sum_{i=1}^n \|X_i\|^2 - \frac{\|\theta\|^2}{d}}\right) \right.$$

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top\theta}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\Bigg|,$$

$$T_{32}:=\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top\theta}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)-\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top\theta}{d\left(1-\frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\right|,$$

$$T_{33}:=\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top\theta}{d\left(1-\frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)-\mathbb{E}\left[\frac{X^\top\theta}{d\left(1-\frac{\|\theta\|^2}{d}\right)}\tanh\left(\frac{X^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\right]\right|.$$

For the term T_{31}, use the fact that $\tanh(x)$ is 1-Lipschitz, we have that

$$T_{31}\leq\sup_{\theta\in\mathbb{B}(\theta^*,r)}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i^\top\theta}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}\right)}\right|\left|\frac{X_i^\top\theta}{\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}}-\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right|$$

$$\leq\frac{C\sqrt{\frac{\log(1/\delta)}{nd}}}{n}\sup_{\theta\in\mathbb{B}(\theta^*,r)}\sum_{i=1}^{n}\frac{(X_i^\top\theta)^2}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}\right)^2\left(1-\frac{\|\theta\|^2}{d}\right)}$$

$$\leq\frac{Cr^2\sqrt{\frac{\log(1/\delta)}{nd}}}{d}\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\right)$$

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

where for the last inequality we use the fact that $r^2\leq d$. For the term T_{32}, use the fact that $|\tanh(x)|\leq|x|$, we have that

$$T_{32}\leq\sup_{\theta\in\mathbb{B}(\theta^*,r)}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i^\top\theta}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}\right)}-\frac{X_i^\top\theta}{d\left(1-\frac{\|\theta\|^2}{d}\right)}\right|\left|\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right|$$

$$\leq\frac{C\sqrt{\frac{\log(1/\delta)}{nd}}}{n}\sup_{\theta\in\mathbb{B}(\theta^*,r)}\sum_{i=1}^{n}\frac{(X_i^\top\theta)^2}{d\left(\frac{1}{nd}\sum_{i=1}^{n}\|X_i\|^2-\frac{\|\theta\|^2}{d}\right)\left(1-\frac{\|\theta\|^2}{d}\right)^2}$$

$$\leq\frac{Cr^2\sqrt{\frac{\log(1/\delta)}{nd}}}{d}\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\right)$$

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

where for the last inequality we still use the fact that $r^2\leq d$. For the term T_{33}, a standard symmetrization argument shows that

$$\mathbb{E}\left[\exp\left(\lambda\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)-\mathbb{E}\left[\frac{X^\top\theta}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\right]\right|\right)\right]$$

$$\leq\mathbb{E}\left[\exp\left(\lambda\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2}{n}\sum_{i=1}^{n}\frac{\varepsilon_i X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\right|\right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda \sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i X_i^\top\theta\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\right|\right)\right],$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d Rademacher sequence. As $\tanh(x)$ is 1-Lipschitz with $\tanh(0)=0$, we use Ledoux-Talagrand contraction inequality, which shows

$$\mathbb{E}\left[\exp\left(C\lambda \sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i X_i^\top\theta\tanh\left(\frac{X_i^\top\theta}{1-\frac{\|\theta\|^2}{d}}\right)\right|\right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda \sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2}{n}\sum_{i=1}^{n}\frac{\varepsilon_i\left(X_i^\top\theta\right)^2}{1-\frac{\|\theta\|^2}{d}}\right|\right)\right]$$

$$\leq \mathbb{E}\left[\exp\left(C\lambda\|\theta\|^2\left\|\frac{\varepsilon_i X_i X_i^\top}{n}\right\|_{\mathrm{op}}\right)\right].$$

Similarly, we have that

$$\mathbb{E}\left[\exp\left(C\lambda\|\theta\|^2\left\|\frac{\varepsilon_i X_i X_i^\top}{n}\right\|_{\mathrm{op}}\right)\right]\leq 2\cdot 17^d\left(\frac{C\lambda^2 r^4}{n}\right),\quad \forall|\lambda|\leq\frac{n}{Cr^2}.$$

With the Chernoff method, we can obtain that $T_{33}\leq\frac{Cr^2}{d}\sqrt{\frac{d\log 1/\delta}{n}}\leq C\sqrt{\frac{d\log 1/\delta}{n}}$, as long as $n\geq Cd\log(1/\delta)$. Combined with the upper bound on T_{31} and T_{32}, we know $T_3\leq Cr\sqrt{\frac{d\log 1/\delta}{n}}$ when $n\geq Cd\log(1/\delta)$, which finishes the proof.

A.2.2 Proof of equation (18)

Now, we provide proof for the improved concentration bound (18) when $|\theta-\theta^*|\leq Cn^{-1/16}$ for some universal constant C. Indeed, an application of triangle inequality leads to

$$|\nabla_\theta f_n(\theta)-\nabla_\theta f(\theta)|\leq|A_n-A||B_n|+|A||B_n-B|,\tag{23}$$

where the terms A_n, B_n, A, B are defined as follows:

$$A_n=\theta\left(a_n+\theta^2\right)-\frac{1}{n}\sum_{i=1}^{n}X_i(a_n+\theta^2)\tanh\left(\frac{X_i\theta}{a_n-\theta^2}\right),\qquad B_n=(a_n-\theta^2)^{-2},$$

$$A=\theta(1+\theta^2)-\mathbb{E}\left[X(1+\theta^2)\tanh\left(\frac{X\theta}{1-\theta^2}\right)\right],\qquad B=(1-\theta^2)^{-2},$$

where $a_n=\frac{1}{n}\sum_{i=1}^{n}X_i^2$. From the standard chi-square concentration [27], we have

$$\mathbb{P}\left(|a_n-1|\geq\sqrt{\frac{8\log(1/\delta)}{n}}\right)\leq\delta.$$

Therefore, we obtain that

$$|B_n-B|=\frac{|a_n-1|(a_n+1-2\theta^2)}{(a_n-\theta^2)^2(1-\theta^2)^2}\leq C_1\sqrt{\frac{\log(1/\delta)}{n}},\tag{24}$$

with probability $1-\delta$ for some universal constant C_1.

To bound A, an application of the Stein's lemma indicates that

$$A = \left| \theta(1 + \theta^2) - \frac{\theta(1 + \theta^2)}{1 - \theta^2} \mathbb{E}\left[\operatorname{sech}^2\left(\frac{X\theta}{1 - \theta^2}\right)\right]\right|$$

$$= \frac{|\theta|(1 + \theta^2)}{1 - \theta^2}\left| 1 - \theta^2 - \mathbb{E}\left[\operatorname{sech}^2\left(\frac{X\theta}{1 - \theta^2}\right)\right]\right|.$$

By means of the inequality $1 - x^2 \le \operatorname{sech}^2(x) \le 1 - x^2 + \frac{2x^4}{3}$, we obtain that

$$1 - \frac{\theta^2}{(1 - \theta^2)^2} \le \mathbb{E}\left[\operatorname{sech}^2\left(\frac{X\theta}{1 - \theta^2}\right)\right] \le 1 - \frac{\theta^2}{(1 - \theta^2)^2} + \frac{2\theta^4}{(1 - \theta^2)^4}.$$

As long as $|\theta| \le Cn^{-1/16} < \rho$ for some constant ρ, we have

$$\left| 1 - \theta^2 - \mathbb{E}\left[\operatorname{sech}^2\left(\frac{X\theta}{1 - \theta^2}\right)\right]\right| \le C_2|\theta|^4,$$

where C_2 is some universal constant. Putting the above results together, we find that

$$A \le c|\theta|^5, \tag{25}$$

where c is some universal constant.

Now, we move to upper bound $|A_n - A|$. Direct application of the triangle inequality leads to

$$|A_n - A| \le |a_n - 1|\left| \theta - \frac{1}{n}\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{a_n - \theta^2}\right)\right|$$

$$+ (1 + \theta^2)\left|\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{a_n - \theta^2}\right) - \mathbb{E}\left[X\tanh\left(\frac{X\theta}{1 - \theta^2}\right)\right]\right|.$$

From the result of Lemma 1 in [7], as long as $|\theta - \theta^*| \le Cn^{-1/16}$ we have

$$\left|\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{a_n - \theta^2}\right) - \mathbb{E}\left[X\tanh\left(\frac{X\theta}{1 - \theta^2}\right)\right]\right| \le c'|\theta|^3 \sqrt{\frac{\log^{10}(5n/\delta)}{n}},$$

with probability $1 - \delta$ for some universal constant c'. Given the bound on A and the above concentration bound, we find that

$$\left| \theta - \frac{1}{n}\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{a_n - \theta^2}\right)\right| \le \frac{|A|}{1 + \theta^2} + \left|\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{a_n - \theta^2}\right) - \mathbb{E}\left[X\tanh\left(\frac{X\theta}{1 - \theta^2}\right)\right]\right|$$

$$\le c|\theta|^5 + c'|\theta|^3|\theta|^3 \sqrt{\frac{\log^{10}(5n/\delta)}{n}},$$

with probability $1 - \delta$. Therefore, we obtain that

$$|A_n - A| \le c''|\theta|^3 \sqrt{\frac{\log^{10}(5n/\delta)}{n}}, \tag{26}$$

where c'' is some universal constant. By plugging the results from equations (24), (25), and (26) to equation (23), we obtain the conclusion of the improved concentration bound (18).

24

B   Discussion: Beyond Symmetric Settings

In this section, we provide discussion showing that the ELU algorithm can still be useful in more general settings than the over-specified symmetric two-component location-scale Gaussian mixtures (1). We specifically consider two settings: (i) Beyond the isotropic covariance matrix in Section B.1; (ii) Beyond the symmetric location parameters in Section B.2. To the best of our knowledge, the theoretical analysis of optimization algorithms for these settings has not been established before in the literature. We aim to provide the insight into the behaviors of the ELU and EM algorithms for solving parameter estimation of these models.

B.1   Beyond Isotropic Covariance Matrix

We first consider the over-specified settings of the symmetric two-component location-scale Gaussian mixtures with diagonal covariance matrix (or in short, over-specified diagonal symmetric two-component location-scale Gaussian mixtures). In particular, we assume that $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(\theta^*, \mathrm{diag}((\sigma_1^*)^2, \ldots, (\sigma_d^*)^2))$ where θ^* and $\sigma_1^*, \ldots, \sigma_d^*$ are true but unknown parameters. Similar to Section 2, for the ease of argument, we assume that $\sigma_1^* = \ldots \sigma_d^* = 1$ (The results in this section still hold for general unknown values of these scale parameters by scaling each dimension by its corresponding scale value). To estimate θ^* and σ_i^* for $i \in [d]$, we also consider fitting the diagonal symmetric two-component location-scale Gaussian mixtures to the data, which is given by:

$$\frac{1}{2}\mathcal{N}(-\theta, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)) + \frac{1}{2}\mathcal{N}(\theta, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)). \tag{27}$$

When $d = 1$, the diagonal model (27) is identical to the isotropic models (1) and (2). When $d \geq 2$, that diagonal model is more general than the isotropic models. To the best of our knowledge, there has not been a theoretical analysis for any optimization algorithms under the over-specified diagonal symmetric two-component location-scale Gaussian mixtures.

EM algorithm: We now derive the EM algorithm for solving parameter estimation of the diagonal model (27). We first describe the latent variable representation of that model. In particular, assume that the latent variable $Z \in \{0, 1\}$ is such that $\mathbb{P}(Z = 0) = \mathbb{P}(Z = 1) = \frac{1}{2}$. Then, we define the following conditional distributions:

$$(X|Z = 0) \sim \mathcal{N}(-\theta, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)), \qquad (X|Z = 1) \sim \mathcal{N}(\theta, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)).$$

For the E-step of the EM algorithm, we first compute the conditional distribution of Z given X, namely, by denoting $\omega_{\theta, \sigma_1, \ldots, \sigma_d}(x) = \mathbb{P}(Z = 1|X = x)$, we have

$$\omega_{\theta, \boldsymbol{\sigma}}(x) = \frac{\exp\left(-\sum_{i=1}^d \frac{(x_i - \theta_i)^2}{2\sigma_i^2}\right)}{\exp\left(-\sum_{i=1}^d \frac{(x_i - \theta_i)^2}{2\sigma_i^2}\right) + \exp\left(-\sum_{i=1}^d \frac{(x_i + \theta_i)^2}{2\sigma_i^2}\right)}$$

$$= \frac{\exp\left(\sum_{i=1}^d \frac{x_i \theta_i}{\sigma_i^2}\right)}{\exp\left(\sum_{i=1}^d \frac{x_i \theta_i}{\sigma_i^2}\right) + \exp\left(-\sum_{i=1}^d \frac{x_i \theta_i}{\sigma_i^2}\right)},$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$. Then, given the location θ and the scale parameters $\sigma_1, \ldots, \sigma_d$, the M-step involves computing the minorization function $(\theta', \boldsymbol{\sigma}') \rightarrow Q(\theta', \boldsymbol{\sigma}'; \theta, \boldsymbol{\sigma})$ where $\boldsymbol{\sigma}' =$

$(\sigma'_1, \ldots, \sigma'_d)$, which is given by:

$$Q(\theta', \boldsymbol{\sigma}'; \theta, \boldsymbol{\sigma}) = \frac{1}{n} \sum_{i=1}^{n} \Bigg(\omega_{\theta,\boldsymbol{\sigma}}(X_i) \log \big(\phi(X_i | \theta', \mathrm{diag}((\sigma'_1)^2, \ldots, (\sigma'_d)^2))) \big)$$

$$+ (1 - \omega_{\theta,\boldsymbol{\sigma}}(X_i)) \log \big(\phi(X_i | -\theta', \mathrm{diag}((\sigma'_1)^2, \ldots, (\sigma'_d)^2))) \big) \Bigg).$$

$$= -\log((\sqrt{2\pi})^d \prod_{j=1}^{d} \sigma'_j) - \frac{1}{n} \sum_{i=1}^{n} \Bigg(\omega_{\theta,\boldsymbol{\sigma}}(X_i) \left(\sum_{j=1}^{d} \frac{(X_{ij} - \theta'_j)^2}{2(\sigma'_j)^2} \right)$$

$$+ (1 - \omega_{\theta,\boldsymbol{\sigma}}(X_i)) \left(\sum_{j=1}^{d} \frac{(X_{ij} + \theta'_j)^2}{2(\sigma'_j)^2} \right) \Bigg).$$

To obtain the EM updates for the location and scale parameters, we maximize the function Q with respect to θ' and $\boldsymbol{\sigma}'$, which leads to the following updates:

$$(\bar{\theta}_{n,\mathrm{EM}}^{t+1})_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} \tanh \left(\sum_{j'=1}^{d} \frac{X_{ij'}(\bar{\theta}_{n,\mathrm{EM}}^{t})_{j'}}{(\bar{\sigma}_{n,\mathrm{EM}}^{t})_{j'}^2} \right), \tag{28}$$

$$(\bar{\sigma}_{n,\mathrm{EM}}^{t+1})_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 - (\bar{\theta}_{n,\mathrm{EM}}^{t+1})_j^2, \tag{29}$$

where $j \in [d]$. Here, we denote $(\bar{\theta}_{n,\mathrm{EM}}^{t+1})_j$ as the j-th element of $(\bar{\theta}_{n,\mathrm{EM}}^{t+1})$ for all $j \in [d]$.

ELU algorithm: Now, we derive the ELU algorithm for solving parameter estimation of the diagonal model (27). The idea is that we first obtain an exact minimization of the scale parameter by solving the negative sample log-likelihood function of model (27), namely, we solve

$$\widehat{\boldsymbol{\sigma}}_n \in -\arg\min_{\sigma} \bar{\mathcal{L}}_n(\theta, \sigma_1, \ldots, \sigma_d)$$

$$:= -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1}{2} \phi(X_i; \theta, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)) + \frac{1}{2} \phi(X_i; -\theta, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)) \right)$$

where $\widehat{\boldsymbol{\sigma}}_n = ((\widehat{\sigma}_n)_1, \ldots, (\widehat{\sigma}_n)_d)$ and obtain the closed-form expression:

$$(\widehat{\sigma}_n)_j^2 = \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 - \theta_j^2 \qquad \forall \, j \in [d]. \tag{30}$$

Given the closed-form expression (30) for the scale parameters, we can utilize the exponential step size gradient descent for the following function

$$\bar{f}_n(\theta) := \bar{\mathcal{L}}_n(\theta, \frac{1}{n} \sum_{i=1}^{n} X_{i1}^2 - \theta_1^2, \ldots, \frac{1}{n} \sum_{i=1}^{n} X_{id}^2 - \theta_d^2). \tag{31}$$

Therefore, we update the location and scale parameters of the ELU algorithm as follows:

$$\bar{\theta}_n^{t+1} = \bar{\theta}_n^t - \frac{\eta}{\beta^t} \nabla_\theta \bar{f}_n(\bar{\theta}_n^t), \tag{32}$$

Algorithm 2: Exponential Location Update (ELU) for Diagonal Model (27)

Input: The step size η, and the scaling parameter $\beta \in (0, 1)$

Output: The updates $\bar{\theta}_n^T, (\bar{\sigma}_n^T)_1, \ldots, (\bar{\sigma}_n^T)_d$ for the location and scale parameters

Initialize $\bar{\theta}_n^0$ and $(\bar{\sigma}_n^0)_1, \ldots, (\bar{\sigma}_n^0)_d$

for $t = 1$ **to** $T - 1$ **do**

    Update location parameter: $\bar{\theta}_n^{t+1} = \bar{\theta}_n^t - \frac{\eta}{\beta^t} \nabla_\theta \bar{f}_n(\bar{\theta}_n^t)$ where the function $\bar{f}_n$ is given in equation (31),

    Update scale parameters: $(\bar{\sigma}_n^{t+1})_j^2 = \frac{1}{n} \sum_{i=1}^n X_{ij}^2 - (\bar{\theta}_n^{t+1})_j^2$

end for

Return $\bar{\theta}_n^T, (\bar{\sigma}_n^T)_1, \ldots, (\bar{\sigma}_n^T)_d$
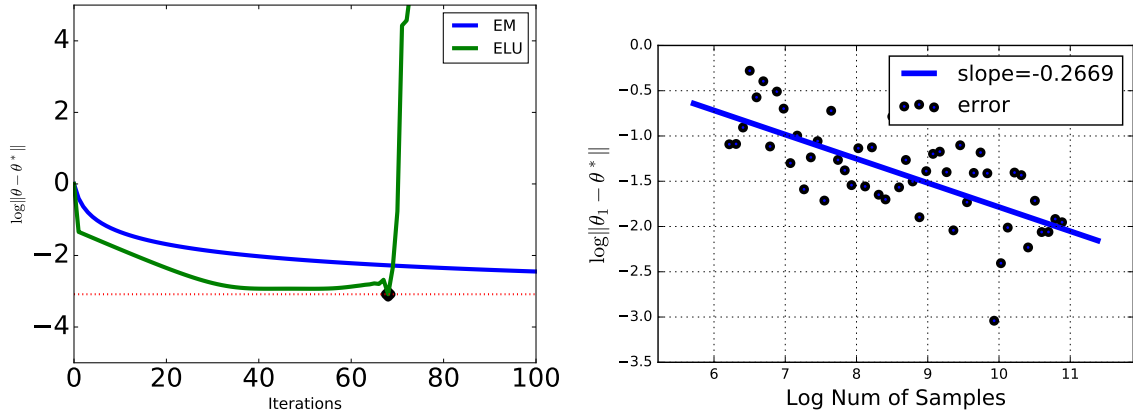

Figure 3. Illustrations for the diagonal model with $d = 4$. **Left:** The optimization rates of the EM and ELU. The black diamond shows the ELU iterates with minimum validation error. ELU can converge to the statistical radius with a linear rate then diverge, while EM converge to the statistical radius with a sub-linear rate. **Right:** ELU can find a solution of θ within the statistical radius $\mathcal{O}(n^{-1/4})$. Given the update of $\sigma_1, \ldots, \sigma_d$ in Algorithm 2, their statistical radii are directly $\mathcal{O}(n^{-1/2})$.

$$(\bar{\sigma}_n^{t+1})_j^2 = \frac{1}{n} \sum_{i=1}^n X_{ij}^2 - (\bar{\theta}_n^{t+1})_j^2, \tag{33}$$

where $(\bar{\theta}_n^{t+1})_j$ is j-th element of $\bar{\theta}_n^{t+1}$. We summarize the ELU algorithm for solving parameter estimation of the diagonal model (27) in Algorithm 2.

Experiments: We now compare the performance of the EM and ELU algorithms for solving parameter estimation of the diagonal model (27). For the settings of our experiments, we use $d = 4$, $\eta = 1$ and $\beta = 0.9$. To compare the optimization rates of ELU and EM, we use $n = 10^6$ samples. The result is shown in the left part of Figure 3, in which ELU iterates converge to the statistical radius linearly then diverge, and EM iterates converge to the statistical radius sub-linearly. In the right part of Figure 3, we show that the statistical radius of θ is $\mathcal{O}(n^{-1/4})$. With the relationship $(\sigma_n)_j^2 = \frac{1}{n} \sum_{i=1}^n X_{ij}^2 - (\theta_n)_j^2$, we directly obtain the statistical radii of $\sigma_1, \sigma_2, \ldots, \sigma_d$ to be $\mathcal{O}(n^{-1/2})$.

27

B.1.1 Insight into the ELU algorithm 2

As we have seen from Figure 3, the ELU updates converge to the statistical radii of the true location and scale parameters geometrically fast. We now aim to provide insight into these behaviors of the ELU updates.

When $d = 1$: The ELU algorithm for diagonal model in Algorithm 2 is similar to the ELU algorithm for the isotropic setting (1) in Algorithm 1 when the dimension $d = 1$. It indicates that the statistical rate of the ELU updates for the location and scale parameters converge to the statistical radii $\mathcal{O}(n^{-1/8})$ and $\mathcal{O}(n^{-1/4})$ within the true location and scale parameters after $\mathcal{O}(\log n)$ number of iterations.

When $d \geq 2$: Given the updates of the location and scale parameters in the ELU algorithm 2, it is sufficient to analyze the convergence of the ELU updates $\{\bar{\theta}_n^t\}_{t \geq 0}$ for the location parameter. As highlighted in the Proof of Theorem 1, a key component to the analysis for the convergence rate of these ELU updates is to study the homogeneity of the population version of the function $\bar{f}_n$ in equation 31, which is given by:

$$\bar{f}(\theta) = \bar{\mathcal{L}}(\theta, 1 - \theta_1^2, \ldots, 1 - \theta_d^2), \tag{34}$$

where $\bar{\mathcal{L}}(\theta, \sigma_1, \ldots, \sigma_d) := -\mathbb{E}\left[\log\left(\frac{1}{2}\phi(X; \theta, \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)) + \frac{1}{2}\phi(X; -\theta, \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)))\right]\right]$. Here, the outer expectation is taken with respect to $X \sim \mathcal{N}(\theta^*, \text{diag}((\sigma_1^*)^2, \ldots, (\sigma_d^*)^2))$ where $\theta^* = 0$ and $\sigma_1^* = \ldots = \sigma_d^* = 1$.

Homogeneity of the function $\bar{f}$ when $d \geq 2$: Given the formulation of the function $\bar{f}$ in equation (34), $\theta^* = 0$ is the global minima. Note that, $\bar{f}(\theta)$ is not locally convex at θ^*. Fortunately, we have the following pseudo-convexity property of the function f at θ^*, which is sufficient for analyzing the optimization aspect of the exponential step-size gradient descent for the location parameter (32) of the ELU algorithm.

Lemma 3. *There exists universal constants ρ, such that for all $\theta \in \mathbb{B}(\theta^*, \rho)$, we have*

$$\bar{f}(\theta) - \bar{f}(\theta^*) \leq \langle \nabla f(\theta), \theta - \theta^* \rangle.$$

Proof of Lemma 3 is in Appendix C.2. Note that, the convexity property requires the inequality in Lemma 3 to hold for any (θ_1, θ_2). However, for the convergence proof for exponential step-size gradient descent, we only requires that $\bar{f}$ has unique global minima and Lemma 3 holds. Furthermore, the landscape of $\bar{f}$ can be described with the following lemma:

Lemma 4. *There exist universal constants $\{\bar{C}_i\}_{i=1}^4$ and universal constants ρ_1 and ρ_2 such that the following holds:*

(a) *When θ has at least two non-zero entries, for all $\theta \in \mathbb{B}(\theta^*, \rho_1)$, we have that*

$$\|\nabla_\theta^2 \bar{f}(\theta)\|_{\text{op}} \leq \bar{C}_1 \|\theta - \theta^*\|^2, \tag{35}$$

$$\|\nabla f(\theta)\| \geq \bar{C}_2 (\bar{f}(\theta) - \bar{f}(\theta)^*)^{3/4}. \tag{36}$$

(b) *When θ has exactly one non-zero entries, without loss of generality, we assume the first entry is non-zero, then we have $(\nabla \bar{f}(\theta))_i = 0$, for all $i \neq 1$. Meanwhile, for all $\theta_1 \in \mathbb{B}(0, \rho_2)$, we have that*

$$\nabla_{\theta_1}^2 \bar{f}(\theta) \leq \bar{C}_3 (\theta_1)^6, \tag{37}$$

$$\nabla_{\theta_1} f(\theta) \geq \bar{C}_4 (\bar{f}(\theta) - \bar{f}(\theta^*))^{7/8}. \tag{38}$$

Proof of Lemma 4 is in Appendix C.1. Note that, part (b) in Lemma 4 corresponds to the case $d = 1$ for the isotropic case which we have discussed in Lemma 1.

Behaviors of the EGD updates when the sample size is infinite: Hence, if we consider the setting when the sample size is infinite, the ELU updates for the location parameter on the population objective $\bar{f}(\theta)$ can consist of two phases. In the first phase where θ has at least two non-zero entries, the EGD iterates converge with rate $\mathcal{O}(\beta^{t/2})$, until the iterates find the global minima, or find some θ with only one non-zero entry. If the EGD iterates find some θ with only one non-zero entry, then we enter the second phase, in which we only update on the non-zero entry, that converges with rate $\mathcal{O}(\beta^{t/6})$ until converge to the exact minima θ^*. As both phases converge linearly, the ELU updates converge linearly on $\bar{f}$ when the sample size is infinite.

Remark on behaviors of the EGD updates when the sample size is finite: In order to establish the statistical behaviors of the EGD iterates for the location when the sample size is finite, as Proposition 1 indicates, we need to establish the stability condition of the function $\bar{f}_n$ around the function $\bar{f}$. However, given the two-phase behaviors of the EGD updates from Lemma 4, it indicates that we need to take into account these two regimes in the stability of $\bar{f}_n$ around $\bar{f}$, which is non-trivial. We conjecture that the empirical iterates will not enter the second phase with high probability, which eventually means that we only need to consider the stability of $\bar{f}_n$ around $\bar{f}$ for the first regime and it directly provides a $\mathcal{O}(n^{1/4})$ statistical radius for the ELU updates for the location parameter after $\mathcal{O}(\log(n/d))$ number of iterations. We leave a formal statement of these results as a future work.

B.2 Two-Component Gaussian Mixtures with General Locations

We now demonstrate empirically that the ELU algorithm is also useful for parameter estimation under the over-specified settings of the two-component location-scale Gaussian mixture with general means. In particular, we assume that the data $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(\theta^*, (\sigma^*)^2 I_d)$ and we fit the following model to the data:

$$\frac{1}{2}\mathcal{N}(\theta_1, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta_2, \sigma^2 I_d). \tag{39}$$

We first derive the updates of the EM and ELU algorithms and then provide experiments to show that the ELU updates converge geometrically fast to the radius of convergence while the EM updates have sub-linear convergence.

EM algorithm: We now derive the EM updates for solving parameter estimation of model (39). Similar to the diagonal model (27), We first describe the latent variable representation of the general mean model. In particular, assume that the latent variable $Z \in \{0, 1\}$ is such that $\mathbb{P}(Z = 0) = \mathbb{P}(Z = 1) = \frac{1}{2}$. Then, we define the following conditional distributions:

$$(X|Z = 0) \sim \mathcal{N}(\theta_1, \sigma^2 I_d), \qquad (X|Z = 1) \sim \mathcal{N}(\theta_2, \sigma^2 I_d).$$

For the E-step of the EM algorithm, we first compute the conditional distribution of Z given X, namely, by denoting $\bar{\omega}_{\theta_1, \theta_2, \sigma}(x) = \mathbb{P}(Z = 1 | X = x)$, we have

$$\bar{\omega}_{\theta_1, \theta_2, \sigma}(x) = \frac{\exp\left(-\frac{\|x - \theta_2\|^2}{2\sigma^2}\right)}{\exp\left(-\frac{\|x - \theta_2\|^2}{2\sigma^2}\right) + \exp\left(-\frac{\|x - \theta_1\|^2}{2\sigma^2}\right)}.$$

Algorithm 3: Exponential Location Update (ELU) for General Mean Model (39)

Input: The step size η, and the scaling parameter $\beta \in (0,1)$
Output: The updates $(\widetilde{\theta}_n^T)_1, (\widetilde{\theta}_n^T)_2, \widetilde{\sigma}_n^T$ for the location and scale parameters
Initialize $(\widetilde{\theta}_n^0)_1, (\widetilde{\theta}_n^0)_2$, and $\widetilde{\sigma}_n^0$
for $t = 1$ **to** $T - 1$ **do**
  Update location parameters: $(\widetilde{\theta}_n^{t+1})_1 = (\widetilde{\theta}_n^t)_1 - \frac{\eta}{\beta^t}\nabla_{\theta_1}\widetilde{f}_n((\widetilde{\theta}_n^t)_1, (\widetilde{\theta}_n^t)_2)$ and
  $(\widetilde{\theta}_n^{t+1})_2 = (\widetilde{\theta}_n^t)_2 - \frac{\eta}{\beta^t}\nabla_{\theta_2}\widetilde{f}_n((\widetilde{\theta}_n^t)_1, (\widetilde{\theta}_n^t)_2)$ where the function $\widetilde{f}_n$ is given in equation (44),
  Update scale parameter:
  $(\widetilde{\sigma}_n^{t+1})^2 = \frac{1}{nd}\sum_{i=1}^n \left\| X_i - \frac{1}{2}\left((\widetilde{\theta}_n^{t+1})_1 + (\widetilde{\theta}_n^{t+1})_2\right)\right\|^2 - \frac{\|(\widetilde{\theta}_n^{t+1})_1 - (\widetilde{\theta}_n^{t+1})_2\|^2}{4d}$
end for
Return $(\widetilde{\theta}_n^T)_1, (\widetilde{\theta}_n^T)_2, \widetilde{\sigma}_n^T$

Then, given the locations θ_1, θ_2, and the scale parameter σ, the M-step involves computing the minorization function $(\theta_1', \theta_2', \sigma') \to \bar{Q}(\theta_1', \theta_2', \sigma'; \theta_1, \theta_2, \sigma)$, which is given by:

$$\bar{Q}(\theta_1', \theta_2', \sigma'; \theta_1, \theta_2, \sigma) = \frac{1}{n}\sum_{i=1}^n \bigg(\bar{\omega}_{\theta_1,\theta_2,\sigma}(X_i)\log\big(\phi(X_i|\theta_2', (\sigma')^2 I_d)$$

$$+ (1 - \bar{\omega}_{\theta_1,\theta_2,\sigma}(X_i))\log\big(\phi(X_i|\theta_1', (\sigma')^2 I_d)\big)\bigg).$$

$$= -d\log(\sqrt{2\pi}\sigma') - \frac{1}{n}\sum_{i=1}^n \bigg(\bar{\omega}_{\theta_1,\theta_2,\sigma}(X_i)\frac{\|X_i - \theta_2'\|^2}{2(\sigma')^2}$$

$$+ (1 - \bar{\omega}_{\theta_1,\theta_2,\sigma}(X_i))\frac{\|X_i - \theta_1'\|^2}{2(\sigma')^2}\bigg).$$

To obtain the EM updates for the location and scale parameters, we maximize the function $\bar{Q}$ with respect to θ_1', θ_2', and σ', which leads to the following updates:

$$(\widetilde{\theta}_{n,\text{EM}}^{t+1})_1 = \frac{\sum_{i=1}^n X_i(1 - \bar{\omega}_{(\widetilde{\theta}_{n,\text{EM}}^t)_1, (\widetilde{\theta}_{n,\text{EM}}^t)_2, \widetilde{\sigma}_{n,\text{EM}}^t}(X_i))}{\sum_{i=1}^n (1 - \bar{\omega}_{(\widetilde{\theta}_{n,\text{EM}}^t)_1, (\widetilde{\theta}_{n,\text{EM}}^t)_2, \widetilde{\sigma}_{n,\text{EM}}^t}(X_i))}, \tag{40}$$

$$(\widetilde{\theta}_{n,\text{EM}}^{t+1})_2 = \frac{\sum_{i=1}^n X_i\bar{\omega}_{(\widetilde{\theta}_{n,\text{EM}}^t)_1, (\widetilde{\theta}_{n,\text{EM}}^t)_2, \widetilde{\sigma}_{n,\text{EM}}^t}(X_i)}{\sum_{i=1}^n \bar{\omega}_{(\widetilde{\theta}_{n,\text{EM}}^t)_1, (\widetilde{\theta}_{n,\text{EM}}^t)_2, \widetilde{\sigma}_{n,\text{EM}}^t}(X_i)}, \tag{41}$$

$$(\widetilde{\sigma}_{n,\text{EM}}^{t+1})^2 = \frac{1}{nd}\sum_{i=1}^d \|X_i\|^2 - \frac{\|(\widetilde{\theta}_{n,\text{EM}}^{t+1})_1\|^2}{nd}\sum_{i=1}^n (1 - \bar{\omega}_{(\widetilde{\theta}_{n,\text{EM}}^t)_1, (\widetilde{\theta}_{n,\text{EM}}^t)_2, \widetilde{\sigma}_{n,\text{EM}}^t}(X_i))$$

$$- \frac{\|(\widetilde{\theta}_{n,\text{EM}}^{t+1})_2\|^2}{nd}\sum_{i=1}^n \bar{\omega}_{(\widetilde{\theta}_{n,\text{EM}}^t)_1, (\widetilde{\theta}_{n,\text{EM}}^t)_2, \widetilde{\sigma}_{n,\text{EM}}^t}(X_i). \tag{42}$$

ELU algorithm: Now, we derive the ELU algorithm for solving parameter estimation of model (39). Note that, the update of scale parameter can be achieved by an exact minimization of the negative sample log-likelihood function of the general mean setting (39), which can be written equivalently as:

$$(\widetilde{\sigma}_n^t)^2 = \frac{1}{nd}\sum_{i=1}^n \left\| X_i - \frac{1}{2}\left((\theta_n^t)_1 + (\theta_n^t)_2\right)\right\|^2 - \frac{\|(\theta_n^t)_1 - (\theta_n^t)_2\|^2}{4d} \tag{43}$$
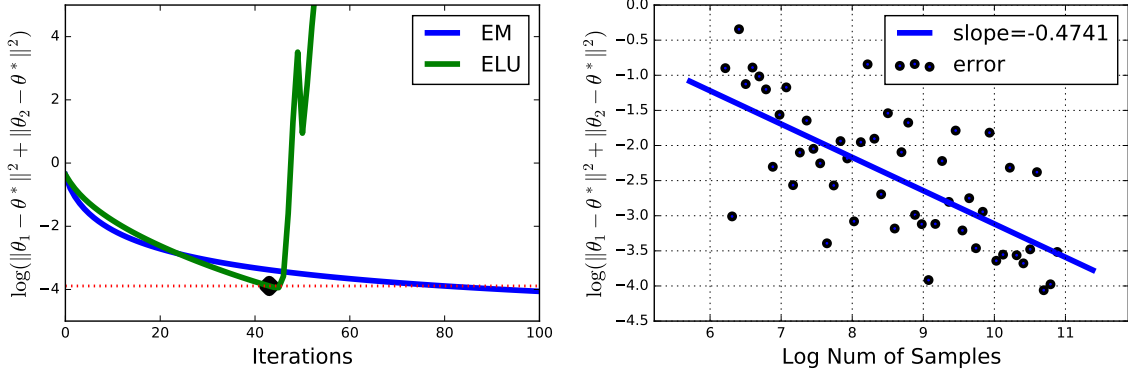
Figure 4. Illustrations for the general location model with $d = 4$. **Left:** The optimization rates of the EM and ELU under Wasserstein metric [14], which had been used to establish convergence rate of parameter estimation in Gaussian mixtures. The black diamond shows the ELU iterates with minimum validation error. ELU can converge to the statistical radius with a linear rate then diverge, while EM converge to the statistical radius with a sub-linear rate. **Right:** ELU can find a solution of θ_1 and θ_2 within the statistical radius $\mathcal{O}(n^{-1/4})$. Given the update of σ in Algorithm 3, its statistical radius is directly $\mathcal{O}(n^{-1/2})$.

Given the closed form expression (43) for the scale parameters, we can utilize the exponential step size gradient descent for the following function:

$$\widetilde{f}_n(\theta_1, \theta_2) = \widetilde{\mathcal{L}}_n \left(\theta_1, \theta_2, \frac{1}{nd} \sum_{i=1}^{n} \left\| X_i - \frac{1}{2} (\theta_1 + \theta_2) \right\|^2 - \frac{\|\theta_1 - \theta_2\|^2}{4d} \right) \tag{44}$$

where the negative sample log-likelihood function $\widetilde{\mathcal{L}}_n$ for model (39) is defined as :

$$\widetilde{\mathcal{L}}_n(\theta_1, \theta_2, \sigma) = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1}{2} \phi(X_i; \theta_1, \sigma^2 I_d) + \frac{1}{2} \phi(X_i; \theta_2, \sigma^2 I_d) \right). \tag{45}$$

Therefore, we update the location and scale parameters of the ELU algorithm as follows:

$$(\widetilde{\theta}_n^{t+1})_1 = (\widetilde{\theta}_n^t)_1 - \frac{\eta}{\beta^t} \nabla_{\theta_1} \widetilde{f}_n((\widetilde{\theta}_n^t)_1, (\widetilde{\theta}_n^t)_2), \tag{46}$$

$$(\widetilde{\theta}_n^{t+1})_2 = (\widetilde{\theta}_n^t)_2 - \frac{\eta}{\beta^t} \nabla_{\theta_2} \widetilde{f}_n((\widetilde{\theta}_n^t)_1, (\widetilde{\theta}_n^t)_2), \tag{47}$$

$$(\widetilde{\sigma}_n^{t+1})^2 = \frac{1}{nd} \sum_{i=1}^{n} \left\| X_i - \frac{1}{2} \left((\widetilde{\theta}_n^{t+1})_1 + (\widetilde{\theta}_n^{t+1})_2 \right) \right\|^2 - \frac{\|(\widetilde{\theta}_n^{t+1})_1 - (\widetilde{\theta}_n^{t+1})_2\|^2}{4d}. \tag{48}$$

We summarize the details of these updates in Algorithm 3.

Experiments: For experiments, we use $d = 4$, $\eta = 1$ and $\beta = 0.9$. To compare the optimization rates of ELU and EM, we use $n = 10^6$ samples. Furthermore, we use Wasserstein metric to measure the differences of the EM and ELU updates to the true parameter, which had been used in previous work to establish the convergence rate of parameter estimation in Gaussian mixture model [14]. The result is shown in the left part of Figure 4, in which ELU iterates converge to the statistical radius linearly then diverge, and EM iterates converge to the statistical radius sub-linearly. In the right part of Figure 4, the statistical radii of θ_1 and θ_2 are $\mathcal{O}(n^{-1/4})$ while that of σ is $\mathcal{O}(n^{-1/2})$.

31

C Proof For Diagonal Model

By direct computation, the population negative log-likelihood function of the diagonal model (27) is given by:

$$\overline{\mathcal{L}}(\theta, \sigma_1, \ldots, \sigma_d) := -\mathbb{E}\left[\log\left(\frac{1}{2}\phi(X; \theta, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2)) + \frac{1}{2}\phi(X; -\theta, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d))\right)\right]$$

$$= \log 2 + \frac{d}{2}\log(2\pi) + \sum_{j=1}^{d}\log(\sigma_j) + \sum_{j=1}^{d}\frac{1 + \theta_j^2}{2\sigma_j^2}$$

$$- \mathbb{E}\left[\log\left(\exp\left(-\sum_{j=1}^{d}\frac{X_j\theta_j}{\sigma_j^2}\right) + \exp\left(\sum_{j=1}^{d}\frac{X_j\theta_j}{\sigma_j^2}\right)\right)\right],$$

where the outer expectation is taken with respect to $X = (X_1, \ldots, X_d)$. Given the definition of the function $\bar{f}$ in equation 34, we have

$$\bar{f}(\theta) = \overline{\mathcal{L}}(\theta, 1 - \theta_1^2, \ldots, 1 - \theta_d^2)$$

$$= \log 2 + \frac{d}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{d}\log(1 - \theta_j^2) + \sum_{j=1}^{d}\frac{1 + \theta_j^2}{2(1 - \theta_j^2)}$$

$$- \mathbb{E}\left[\log\left(\exp\left(-\sum_{j=1}^{d}\frac{X_j\theta_j}{1 - \theta_j^2}\right) + \exp\left(\sum_{j=1}^{d}\frac{X_j\theta_j}{1 - \theta_j^2}\right)\right)\right].$$

With a slightly abuse of notation, for the function $g : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla_i g(\theta) = [\nabla g(\theta)]_i$ and $\nabla_{ij}^2 g(\theta) = [\nabla^2 g(\theta)]_{i,j}$. Then for any $j \in [d]$, direct calculation yields that

$$\nabla_j \bar{f}(\theta) = \frac{\theta_j(1 + \theta_j^2)}{(1 - \theta_j^2)^2} - \frac{1 + \theta_j^2}{(1 - \theta_j^2)^2}\mathbb{E}\left[X_j \tanh\left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1 - \theta_l^2}\right)\right],$$

where $\theta = (\theta_1, \ldots, \theta_d)$. Furthermore, we have

$$\nabla_{jj}^2 \bar{f}(\theta) = \frac{\theta_j^4 + 6\theta_j^2 + 1}{(1 - \theta_j^2)^3} - \frac{2\theta_j(\theta_j^2 + 3)}{(1 - \theta_j^2)^3}\mathbb{E}\left[X_j \tanh\left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1 - \theta_l^2}\right)\right]$$

$$- \frac{(1 + \theta_j^2)^2}{(1 - \theta_j^2)^4}\mathbb{E}\left[X_j^2 \operatorname{sech}^2\left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1 - \theta_l^2}\right)\right],$$

and when $j_1 \neq j_2$ we have

$$\nabla_{j_1 j_2}^2 \bar{f}(\theta) = -\frac{1 + \theta_{j_1}^2}{(1 - \theta_{j_1}^2)^2}\frac{1 + \theta_{j_2}^2}{(1 - \theta_{j_2}^2)^2}\mathbb{E}\left[X_{j_1}X_{j_2}\operatorname{sech}^2\left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1 - \theta_l^2}\right)\right].$$

C.1 Proof of Lemma 4

We first provide proof of part (a) and then proof of part (b).

C.1.1 Proof of Part (a)

Now we focus on the proof of the inequalities (35) and (36).

Proof of inequality (35): With Stein's Lemma, we know

$$\nabla_j \bar{f}(\theta) = \frac{\theta_j(1+\theta_j^2)}{(1-\theta_j^2)^2} - \frac{1+\theta_j^2}{(1-\theta_j^2)^2}\mathbb{E}\left[X_j \tanh\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\right]$$

$$= \frac{\theta_j(1+\theta_j^2)}{(1-\theta_j^2)^2} - \frac{\theta_j\left(1+\theta_j^2\right)}{(1-\theta_j^2)^3}\mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\right].$$

We define

$$C(\theta) := \mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\right],$$

with the gradients of the function C can be computed as follows:

$$\nabla_j C(\theta) = -2\frac{1+\theta_j^2}{(1-\theta_j^2)^2}\mathbb{E}\left[X_j \tanh\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\text{sech}^2\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\right]$$

$$= -2\frac{\theta_j(1+\theta_j^2)}{(1-\theta_j^2)^3}\mathbb{E}\left[\text{sech}^4\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right) - 2\tanh^2\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\text{sech}^2\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\right]$$

$$= -2\frac{\theta_j(1+\theta_j^2)}{(1-\theta_j^2)^3}\mathbb{E}\left[3\text{sech}^4\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right) - 2\text{sech}^2\left(\sum_{l=1}^d \frac{X_l\theta_l}{1-\theta_l^2}\right)\right]$$

$$=: -2g(\theta_j)C'(\theta),$$

where the second equality is due to Stein's lemma. We can rewrite $\nabla_{j_1 j_2}^2 f(\theta)$ for $j_1 \neq j_2$ as

$$\nabla_{j_1 j_2}^2 \bar{f}(\theta) = 2C'(\theta)g(\theta_{j_1})g(\theta_{j_2}).$$

So $\nabla_\theta^2 \bar{f}(\theta)$ can be written as

$$\nabla_\theta^2 \bar{f}(\theta) = D + 2C'(\theta)G(\theta)G(\theta)^\top,$$

where D is a diagonal matrix with

$$D_{jj} = \frac{\theta_j^4 + 6\theta_j^2 + 1}{(1-\theta_j^2)^3} - \frac{3\theta_j^4 + 8\theta_j^2 + 1}{(1-\theta_j^2)^4}C(\theta),$$

and $G(\theta) = [g(\theta_l)]_{l\in[d]}^\top$ for all $\theta \in \mathbb{R}^d$.

As $\text{sech}^2(x) \geq 1 - x^2$, we know as long as $\|\theta\|_2 \leq \rho < 1$, we have

$$D_{jj} \leq -\frac{\theta_j^2(\theta_j^4 + 8\theta_j^2 + 3)}{(1-\theta_j^2)^4} + \frac{3\theta_j^4 + 8\theta_j^2 + 1}{(1-\theta_j^2)^4}\sum_{l=1}^d \frac{\theta_l^2}{(1-\theta_l^2)^2} \leq \frac{3\theta_j^4 + 8\theta_j^2 + 1}{(1-\theta_j^2)^4}\sum_{l=1}^d \frac{\theta_l^2}{(1-\theta_l^2)^2} \leq c_1\|\theta\|^2,$$

where c_1 is some positive absolute constant that only depends on ρ. Meanwhile, use the fact that $\text{sech}^2(x) \leq 1$, we know as long as $\|\theta\|_2 \leq \rho < 1$, we have

$$D_{jj} \geq -\frac{\theta_j^2(\theta_j^4 + 8\theta_j^2 + 3)}{(1-\theta_j^2)^4} \geq c_2\|\theta\|_2^2,$$

33

where c_2 is some negative absolute constant that only depends on ρ.

Meanwhile, use the fact that $3\text{sech}^4(x) - 2\text{sech}^2(x) \le 1$, we know that, as long as $\|\theta\|_2 \le \rho < 1$, we have

$$2C'(\theta)G(\theta)^\top G(\theta) \le \sum_{l=1}^{d} \frac{\theta_l^2 (1 + \theta_l^2)^2}{(1 - \theta_l^2)^6} \le c_3 \|\theta\|^2,$$

where c_3 is another universal constant that only depends on ρ. Hence, with Weyl's inequality, we conclude the inequality (35).

Proof of inequality (36): Recall that

$$
\begin{aligned}
\nabla_j \bar{f}(\theta) =& \frac{\theta_j (1 + \theta_j^2)}{(1 - \theta_j^2)^2} - \frac{\theta_j (1 + \theta_j^2)}{(1 - \theta_j^2)^3} \mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^{d} \frac{X_l \theta_l}{1 - \theta_l^2}\right)\right] \\
=& \frac{\theta_j (1 + \theta_j^2)}{(1 - \theta_j)^3}\left(1 - \theta_j^2 - \mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^{d} \frac{X_l \theta_l}{1 - \theta_l^2}\right)\right]\right).
\end{aligned}
$$

Use the fact that $1 - x^2 + \frac{2x^4}{3} - \frac{17x^6}{45} \le \text{sech}^2(x) \le 1 - x^2 + \frac{2x^4}{3}$, we know as long as $\|\theta\|^2 \le \rho$ for some $\rho > 0$, we have

$$
\begin{aligned}
& \left(1 - \theta_j^2 - \mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^{d} \frac{X_l \theta_l}{1 - \theta_l^2}\right)\right]\right)^2 \\
=& (1 - \theta_j^2)^2 - 2(1 - \theta_j^2)\mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^{d} \frac{X_l \theta_l}{1 - \theta_l^2}\right)\right] + \left(\mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^{d} \frac{X_l \theta_l}{1 - \theta_l^2}\right)\right]\right)^2 \\
\ge& (1 - \theta_j^2)^2 - 2(1 - \theta_j^2) + 2(1 - \theta_j^2)\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2} - 4(1 - \theta_j^2)\left(\sum_{l=1}^{d} \frac{\theta_l^4}{(1 - \theta_l^2)^2}\right)^2 \\
& + 1 - 2\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2} + 5\left(\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2}\right)^2 - \frac{56}{3}\left(\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2}\right)^3 \\
=& \left(\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2} - \theta_j^2\right)^2 - \frac{56}{3}\left(\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2}\right)^3 \\
\ge& c_1 \|\theta\|^2
\end{aligned}
$$

where c_1 is some universal constant. Hence, we have that $\|\nabla \bar{f}(\theta)\|^2 \ge C_1 \|\theta\|^3$. Meanwhile, note that

$$
\begin{aligned}
\bar{f}(\theta) - \bar{f}(\theta^*) =& \frac{1}{2}\sum_{j=1}^{d} \log(1 - \theta_j^2) + \sum_{j=1}^{d} \frac{1 + \theta_j^2}{2(1 - \theta_j^2)} - \frac{d}{2} \\
& - \mathbb{E}\left[\log\left(\frac{1}{2}\left(\exp\left(-\sum_{j=1}^{d} \frac{X_j \theta_j}{1 - \theta_j^2}\right) + \exp\left(\sum_{j=1}^{d} \frac{X_j \theta_j}{1 - \theta_j^2}\right)\right)\right)\right].
\end{aligned}
$$

34

Use the fact that $\log(\frac{1}{2}(\exp(x) + \exp(-x))) \geq \frac{x^2}{2} - \frac{x^4}{12}$, we have that

$$\mathbb{E}\left[\log\left(\frac{1}{2}\left(\exp\left(-\sum_{j=1}^{d}\frac{X_j\theta_j}{1-\theta_j^2}\right) + \exp\left(\sum_{j=1}^{d}\frac{X_j\theta_j}{1-\theta_j^2}\right)\right)\right)\right]$$

$$\geq \frac{1}{2}\sum_{j=1}^{d}\frac{\theta_j^2}{(1-\theta_j^2)^2} - \frac{1}{4}\left(\sum_{j=1}^{d}\frac{\theta_j^2}{(1-\theta_j^2)^2}\right)^2.$$

Meanwhile, note that

$$\log(1+x) \geq x - \frac{x^2}{2}, \forall x \geq 0.$$

Take $x = \frac{\theta_j^2}{1-\theta_j^2}$, we have

$$\log(1-\theta_j^2) \leq -\frac{\theta_j^2}{1-\theta_j^2} + \frac{\theta_j^4}{2(1-\theta_j^2)^2}.$$

Hence,

$$\bar{f}(\theta) - \bar{f}(\theta^*) \leq \sum_{j=1}^{d}\left(-\frac{\theta_j^4}{4(1-\theta_j^2)^2} + \frac{\theta_j^4}{4(1-\theta_j^2)^4}\right) + \frac{1}{2}\sum_{u\neq v}\frac{\theta_u^2\theta_v^2}{(1-\theta_u^2)^2(1-\theta_v^2)^2}$$

$$\leq \sum_{j=1}^{d}\frac{2\theta_j^6 - \theta_j^8}{4(1-\theta_j^2)^4} + \frac{1}{2}\sum_{u\neq v}\frac{\theta_u^2\theta_v^2}{(1-\theta_u^2)^2(1-\theta_v^2)^2} \leq c_2\|\theta\|^4$$

where c_2 is some universal constant as θ have at least two non-zero entries. Combined with the fact that $\|\nabla f(\theta)\| \geq c_1\|\theta\|^3$, we conclude the inequality (36).

C.1.2 Proof of Part (b)

The proof of part (b) follows directly from the proof of Lemma 1 when the dimension $d = 1$. As a consequence, we obtain the conclusion of part (b).

C.2 Proof of Lemma 3

We now provide the proof for Lemma 3. Note that, we have

$$\nabla\bar{f}(\theta)^\top\theta = \sum_{j=1}^{d}\frac{\theta_j^2(1+\theta_j^2)}{(1-\theta_j^2)^3}\left(1-\theta_j^2 - \mathbb{E}\left[\text{sech}^2\left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1-\theta_l^2}\right)\right]\right)$$

$$\geq \sum_{j=1}^{d}\frac{\theta_j^2(1+\theta_j^2)}{(1-\theta_j^2)^3}\left(1-\theta_j^2 - \left(\mathbb{E}\left[1 - \left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1-\theta_l^2}\right)^2 + \frac{2}{3}\left(\sum_{l=1}^{d}\frac{X_l\theta_l}{1-\theta_l^2}\right)^4\right]\right)\right)$$

$$= \sum_{j=1}^{d}\frac{\theta_j^2(1+\theta_j^2)}{(1-\theta_j^2)^3}\left(-\theta_j^2 + \sum_{l=1}^{d}\frac{\theta_l^2}{(1-\theta_l^2)^2} - 2\left(\sum_{l=1}^{d}\frac{\theta_l^2}{(1-\theta_l^2)^2}\right)^2\right)$$

$$\geq 2 \sum_{u \neq v} \frac{\theta_u^2 \theta_v^2}{(1 - \theta_u^2)^2 (1 - \theta_v^2)^2} - 2 \left(\sum_{j=1}^{d} \frac{\theta_j^2 (1 + \theta_j^2)}{(1 - \theta_j^2)^3} \right) \left(\sum_{l=1}^{d} \frac{\theta_l^2}{(1 - \theta_l^2)^2} \right)^2.$$

Meanwhile, we have

$$\bar{f}(\theta) - \bar{f}(\theta^*) \leq \sum_{j=1}^{d} \left(-\frac{\theta_j^4}{4(1 - \theta_j^2)^2} + \frac{\theta_j^4}{4(1 - \theta_j^2)^4} \right) + \frac{1}{2} \sum_{u \neq v} \frac{\theta_u^2 \theta_v^2}{(1 - \theta_u^2)^2 (1 - \theta_v^2)^2}$$

$$\leq \sum_{j=1}^{d} \frac{2\theta_j^6 - \theta_j^8}{4(1 - \theta_j^2)^4} + \frac{1}{2} \sum_{u \neq v} \frac{\theta_u^2 \theta_v^2}{(1 - \theta_u^2)^2 (1 - \theta_v^2)^2}.$$

Hence, there exists some ρ, such that as long as $\|\theta\| \leq \rho$, we have $\nabla \bar{f}(\theta)^\top \theta \geq \bar{f}(\theta) - \bar{f}(\theta^*)$.

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017. (Cited on pages 2, 3, 4, 6, and 9.)

[2] T. T. Cai, J. Ma, and L. Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019. (Cited on pages 2 and 4.)

[3] Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754, 2018. (Cited on page 9.)

[4] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995. (Cited on page 2.)

[5] C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, 2017. (Cited on pages 2, 3, and 4.)

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–38, 1997. (Cited on pages 2 and 4.)

[7] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020. (Cited on pages 2, 3, 4, 5, 7, 8, 9, and 24.)

[8] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 44:2726–2755, 2020. (Cited on pages 3, 4, 9, and 21.)

[9] S. Ghosal and A. van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263, 2001. (Cited on page 1.)

[10] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on page 9.)

[11] Z. V. Havre, N. White, J. Rousseau, and K. Mengersen. Overfitting Bayesian mixture models with an unknown number of components. *PLOS One*, 10, 2015. (Cited on page 2.)

[12] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46:2844–2870, 2018. (Cited on page 2.)

[13] N. Ho, K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu. Instability, computational efficiency and statistical accuracy. *Arxiv Preprint Arxiv: 2005.11411*, 2020. (Cited on page 9.)

[14] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016. (Cited on pages 2 and 31.)

[15] N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307, 2016. (Cited on page 1.)

[16] N. Ho, T. Ren, S. Sanghavi, P. Sarkar, and R. Ward. An exponentially increasing step-size for parameter estimation in statistical models. *arXiv preprint arXiv: 2205.07999*, 2022. (Cited on pages 5, 8, 9, and 10.)

[17] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems 29*, 2016. (Cited on page 2.)

[18] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824, 2018. (Cited on page 9.)

[19] J. Kwon, W. Qian, C. Caramanis, Y. Chen, , and D. Damek. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory (COLT)*, 2019. (Cited on page 4.)

[20] Z. Li and S. Arora. An exponential learning rate schedule for deep learning. In *ICLR*, 2019. (Cited on page 5.)

[21] B. Lindsay. *Mixture Models: Theory, Geometry and Applications*. In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995. (Cited on page 1.)

[22] T. Manole and A. Khalili. Estimating the number of components in finite mixture models via the group-sort-fuse procedure. *Annals of Statistics*, 49:3043–3069, 2021. (Cited on page 2.)

[23] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs.* New York, 1988. (Cited on page 1.)

[24] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59:731–792, 1997. (Cited on page 2.)

[25] K. Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89:487–495, 1994. (Cited on page 2.)

[26] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710, 2011. (Cited on page 2.)

[27] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. (Cited on pages 18, 19, 20, 21, and 23.)

[28] Z. Wang, Q. Gu, Y. Ning, and H. Liu. High-dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems 28*, 2015. (Cited on pages 2 and 4.)

[29] J. Xu, D. Hsu, and A. Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*, 2016. (Cited on pages 2, 3, and 4.)

[30] B. Yan, M. Yin, and P. Sarkar. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems 30*, 2017. (Cited on page 2.)

[31] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems 28*, 2015. (Cited on pages 2 and 4.)

[32] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015. (Cited on page 3.)