Statistical Convergence Analysis of Gradient EM on General Gaussian Mixture Models

Bowei Yan[1], Mingzhang Yin[1], and Purnamrita Sarkar [1]

[1]University of Texas at Austin

Abstract

In this paper, we study convergence properties of the gradient Expectation-Maximization algorithm [9] for Gaussian Mixture Models for general number of clusters and mixing coefficients. We derive the convergence rate depending on the mixing coefficients, minimum and maximum pairwise distances between the true centers and dimensionality and number of components; and obtain a near-optimal local contraction radius. While there have been some recent notable works that derive local convergence rates for EM in the two equal mixture symmetric GMM, in the more general case, the derivations need structurally different and non-trivial arguments. We use recent tools from learning theory and empirical processes to achieve our theoretical results.

1 Introduction

Proposed by [7] in 1977, the Expectation-Maximization (EM) algorithm is a powerful tool for statistical inference in latent variable models. Typically, problems where EM is used contain either missing values, or the problem is conceptually simplified by assuming there are unobserved or latent variables. For example, for estimating a mixture of different distributions from a given parametric family, one can simply pretend that every datapoint has an unobserved label which indicates which mixture component it came from.

Typically the incomplete loglikelihood (which integrates the latent variables out) is hard to optimize, and hence EM iteratively optimizes a lower bound of it and obtains a sequence of estimators. EM is guaranteed to obtain a local optimum, since it always improves the objective function. While it is established in [4] that the true parameter vector is the global maximizer of the log-likelihood function, there has been much effort to understand the behavior of the local optima obtained via EM.

Gradient EM, as a popular variant of EM, is widely used when exact M-step is burdensome. [9] introduces a gradient algorithm using one iteration of Newton's method and shows the local properties of the gradient EM are almost identical with those of the EM. In [1], the convergence guarantees for gradient EM is studied in two balanced component Gaussian Mixture Model (GMM). [8] shows a counter example that gradient EM with bad initialization can converge to arbitrarily worse sub-optimal point.

Early literature [19, 21] mostly focuses on the convergence to the stationary points or local optima.In [19] it is proven that the convergence of the sequence of estimators in EM to stationary point when the lower bound function from E-step is continuous. In addition, some conditions are derived under which EM converges to local maxima instead of saddle points; but these are typically hard to check. A link between EM and gradient methods is forged in[21] via a projection matrix and the local convergence rate of EM is obtained. In particular, it is shown that for GMM with well-separated centers, the EM achieves faster convergence rates comparable to a quasi-Newton algorithm. While the convergence of EM deteriorates under worse separations, it is observed in [17] that the mixture density determined by estimator sequence of EM reflects the sample data well.

In recent years, there has been a renewed wave of interest in studying the behavior of EM especially in GMM. The global convergence of EM for a mixture of two equal proportion Gaussian distributions is fully characterized in [20]. For more than two clusters, a negative result on EM and gradient EM being trapped in local minima arbitrarily far away from the global optimum is shown in [8].

Lloyd's algorithm is another fast non-convex clustering algorithm, which has a similar flavor as EM. At each step, it recomputes the centroids of each cluster and updates the membership assignments alternatively. While EM does soft clustering at each step, Lloyd's algorithm obtains hard clustering. The clustering error of Lloyd's algorithm for arbitrary number of clusters is studied in [11]. The authors also show local convergence results where the contraction region is less restrictive than [1].

For high dimensional GMMs with M components, the parameters are learned via reducing the dimensionality via a random projection in [5]. In [6] the two-round method is proposed, where one first initializes with more than M points, then prune to get one point in every cluster. It is pointed out in this paper that in high dimensional space, when the clusters are well separated, the mixing weight will go to either 0 or 1 after one single update. [22] shows one can achieve exact recovery for high dimensional sub-gaussian mixtures by convex relaxations. [14] use semi definite programming relaxations to cluster subgaussian mixtures.

For the convergence rate of EM algorithm, it is observed in [16] that a very small mixing coefficient/proportion for one mixture component compared to others leads to slow convergence. [1] gives non-asymptotic convergence guarantees in isotropic, balanced, two-component GMM; their result proves the linear convergence of EM if the center is initialized in a small neighborhood of the true parameters. The local convergence result in this paper has a sub-optimal contraction region.

In this paper, we study the convergence rate and local contraction radius of gradient EM under isotropic GMM with arbitrary number of clusters and arbitrary mixing weights. We obtain a near-optimal condition on the contraction region in contrast to [1]'s contraction radius for the mixture of two equal weight Gaussians. We want to point out that, while the authors of [1] provide a general set of conditions to establish local convergence for a broad class of mixture models, the derivation of specific results and conditions on local convergence are tailored to the balanced and symmetric two component GMM.

We follow the same general route: first we obtain conditions for population gradient EM, where all sample averages are replaced by their expected versions. Then we translate the population version to a sample one. While the first part is conceptually similar, the general setting complicates it. The second step typically makes use of concepts from empirical processes, by pairing up Ledoux Talagrand contraction type arguments with well established symmetrization results. However, in our case, the function is not 1-Lipschitz like the symmetric two component case, since it involves the cluster estimates of all M components. Furthermore, the standard analysis of concentration inequalities by McDiarmid's inequality gets complicated because our subgaussian mixtures can be unbounded. We take advantage of recent tools in Rademacher averaging for vector valued function classes, and get a Rademacher complexity bound for the gradient of general finite sample from GMMs, which might be of independent interest.

The rest of the paper is organized as follows. In Section 2, we state the problem and the notations. In Section 3, we provide the main results in local convergence rate and region for both population and sample based gradient EM in GMM. Section 4 and 5 provide the proof sketches of population and sample based theoretical results, followed by the numerical result in Section 6. And we conclude the paper with some discussions.

2 Problem Setup and Notations

Consider a GMM with M clusters in d dimensional space, with weights $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_M)$. Let $\boldsymbol{\mu}_i \in \mathbb{R}^d$ be the mean of cluster i. Without loss of generality, we assume $\mathbb{E}X = \sum_i \pi_i \boldsymbol{\mu}_i = 0$. In this paper we focus on the isotropic case, for simplicity, we assume the covariance matrix for each cluster is I_d. Let $\boldsymbol{\mu} \in \mathbb{R}^{Md}$ be the vector stacking the $\boldsymbol{\mu}_i$s vertically. We represent the mixture as $X \sim \text{GMM}(\pi, \boldsymbol{\mu}, I_d)$, which has the density function $p(x|\boldsymbol{\mu}) = \sum_{i=1}^M \pi_i \phi_i(x|\boldsymbol{\mu}_i, I_d)$. where $\phi(x; \boldsymbol{\mu}, \Sigma)$ is the PDF of $N(\boldsymbol{\mu}, \Sigma)$. Then the population log-likelihood function as $\mathcal{L}(\boldsymbol{\mu}) = \mathbb{E}_X \log \left(\sum_{i=1}^M \pi_i \phi(X|\boldsymbol{\mu}_i, I_d) \right)$. The Maximum Likelihood

Estimator is then defined as $\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \arg\max p(X|\boldsymbol{\mu})$. EM algorithm is based on using an auxiliary function to lower bound the log likelihood. Define $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) = \mathbb{E}_X\left[\sum_z f(z|X;\boldsymbol{\mu}^t)\log\phi(X,z;\boldsymbol{\mu},I_d)\right]$. The standard EM update is $\boldsymbol{\mu}^{t+1} = \arg\max_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t)$. Define

$$w_i(X;\boldsymbol{\mu}) = \frac{\pi_i\phi(X|\boldsymbol{\mu}_i, I_d)}{\sum_{j=1}^{M}\pi_j\phi(X|\boldsymbol{\mu}_j, I_d)} \tag{1}$$

The update step for gradient EM is

$$\boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^t + s[\nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t)]_i = \boldsymbol{\mu}_i^t + s\mathbb{E}_X\left[\pi_i w_i(X;\boldsymbol{\mu}^t)(X-\boldsymbol{\mu}_i^t)\right]. \tag{2}$$

We assume we are given an initialization $\boldsymbol{\mu}_i^0$ and the true mixing weight π_i for each component.

2.1 Notations

Define $R_{\max}$ and $R_{\min}$ as the largest and smallest distance between cluster centers i.e., $R_{\max} = \max_{i\neq j}\|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|$, $R_{\min} = \min_{i\neq j}\|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|$. Let $\pi_{\max}$ and $\pi_{\min}$ be the maximal and minimal cluster weights, and define κ as $\kappa = \frac{\pi_{\max}}{\pi_{\min}}$. Standard complexity analysis notation $o(\cdot), O(\cdot), \Theta(\cdot), \Omega(\cdot)$ will be used. $f(n) = \tilde{\Omega}(g(n))$ is short for $\Omega(g(n))$ ignoring log factor, equivalent to $f(n) \geq Cg(n)\log^k(g(n))$, similar for others. We use $\otimes$ to represent the kronecker product.

3 Main Results

Despite being a non-convex problem, EM and gradient EM algorithms have been shown to exhibit good convergence behavior in practice, especially with good initializations. However, existing local convergence theory only applies for two-cluster equal-weight GMM. In this section, we present our main result in two parts. First we show the convergence rate and present a near-optimal radius for contraction region for population gradient EM. Then in the second part we connect the population version to finite sample results using concepts from empirical processes and learning theory.

3.1 Local contraction for population gradient EM

Intuitively, when $\boldsymbol{\mu}^t$ equals the ground truth $\boldsymbol{\mu}^*$, then the $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^*)$ function will be well-behaved. This function is a key ingredient in [1], where the curvature of the $Q(\cdot|\boldsymbol{\mu})$ function is shown to be close to the curvature of $Q(\cdot|\boldsymbol{\mu}^*)$ when the $\boldsymbol{\mu}$ is close to $\boldsymbol{\mu}^*$. This is a local property that only requires the gradient to be stable at one point.

Definition 1 (Gradient Stability)**.** *The Gradient Stability (GS) condition, denoted by GS(γ, a), is satisfied if there exists $\gamma > 0$, such that for $\boldsymbol{\mu}_i^t \in \mathbb{B}(\boldsymbol{\mu}_i^*, a)$ with some $a > 0$, for $\forall i \in [M]$.*

$$\|\nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^*) - \nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t)\| \leq \gamma\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$$

The GS condition is used to prove contraction of the sequence of estimators produced by population gradient EM. However, for most latent variable models, it is typically challenging to verify the GS condition and obtain a tight bound on the parameter γ. Even for two equal weight symmetric GMM, there is no derivation of the GS condition found in the literature to the best of our knowledge. We derive a stronger version of the GS condition (see Theorem 4 in Section 4), which has a uniform bound on the deviation of the partial gradient evaluated at $\boldsymbol{\mu}_i^t$. This immediately implies the global GS condition defined in 1. Equipped with this result, we achieve a nearly optimal local convergence radius for general GMMs in Theorem 1. The proof of this theorem can be found in Appendix B.2.

Theorem 1 (Convergence for Population gradient EM). *If $R_{\min} = \tilde{\Omega}(\sqrt{\min\{d, M\}})$, with initialization $\boldsymbol{\mu}^0$ satisfying, $\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\| \le a, \forall i \in [M]$, where*

$$a \le \frac{R_{\min}}{2} - \sqrt{\min\{d, M\}} O\left(\sqrt{\log\left(\max\left\{\frac{M^2\kappa}{\pi_{\min}}, R_{\max}, \min\{d, M\}\right\}\right)}\right)$$

then the Population EM converges:

$$\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\| \le \zeta^t \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}^*\|, \qquad \zeta = \frac{\pi_{\max} - \pi_{\min} + 2\gamma}{\pi_{\max} + \pi_{\min}} < 1$$

where $\gamma = M^2(2\kappa + 4)\left(2R_{\max} + \min\{d, M\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right) < \pi_{\min}$.

Remark 1. *The local contraction radius is largely improved compared to that in [1], which has $R_{\min}/8$ in the two equal sized symmetric GMM setting. It can be seen that in Theorem 1, $a/R_{\min}$ goes to $\frac{1}{2}$ as the signal to noise ratio goes to infinity. When a exceeds $R_{\min}/2$, typically label switching happens because of overlap in the initialization regions. We also show in simulations that when initialized from some point that lies $R_{\min}/2$ away from the true center, gradient EM only converges to a stationary point which is not a global optima. More discussion can be found in Section 6.*

3.2 Finite Sample bound for gradient EM

In the finite sample setting, as long as the deviation of the sample gradient from the population gradient is uniformly bounded, the convergence in the population setting implies the convergence in finite sample scenario. Typically, in empirical processes, one uses Ledoux-Talagrand type contraction results [10] coupled with argGradient EM, as a popular variant of EM, is widely used when exact M-step is burdensome. [9] introduces a gradient algorithm using one iteration of Newton's method and shows the local properties of the EM gradient algorithm are almost identical with those of the EM. In [1], the convergence guarantees for gradient EM is studied in two balanced Gaussian Mixture. And paper [8] shows an counter example that gradient EM with bad initialization can converge to arbitrarily worse sub-optimal point. uments that use covering to obtain uniform deviation bounds. In order to use the Ledoux-Talagrand contraction, one needs a 1-Lipschitz function, which we do not have, because our function involves $\boldsymbol{\mu}_i$, $i \in [M]$. Also, the weight functions w_i are not separable in terms of the $\boldsymbol{\mu}_i$'s.

Instead we take the route of achieving concentration bounds via obtaining the Rademacher complexity, coupled with McDiarmid's inequality [13]. Again, this requires the derivation of the Rademacher complexity for general GMMs and application of McDiarmid becomes problematic, because for GMMs, the bounded difference condition does not hold.

We will start with some notations. Recall the gradient of Q in Eq. (2). To show the sample gradient is close to its population counterpart, consider the following class of functions indexed by $\boldsymbol{\mu}$ and some unit vector on d dimensional sphere $u \in \mathcal{S}^{d-1}$:

$$\mathcal{F}_i^u = \{f^i : \mathcal{X} \to \mathbb{R} | f^i(X; \boldsymbol{\mu}, u) = w_i(X; \boldsymbol{\mu})\langle X - \boldsymbol{\mu}_i, u\rangle\} \tag{3}$$

We need to bound the M functions classes separately for each mixture. Given a finite n-sample $(X_1, \cdots, X_n)$, for each class, we define the Rademacher complexity as the expectation of empirical Rademacher complexity.

$$\hat{R}_n(\mathcal{F}_i^u) = \mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n}\sum_{j=1}^n \epsilon_i f^i(X_j; \boldsymbol{\mu}, u)\right]; \qquad R_n(\mathcal{F}_i^u) = \mathbb{E}_X \hat{R}_n(\mathcal{F}_i^u)$$

where ϵ_i's are the i.i.d. Rademacher random variables.

However, to bound the empirical difference by Rademacher complexity, we need to use tools in the martingale concentration. Unfortunately for the functions defined in Eq. (3), the martingale difference sequence is neither independent or bounded. Hence classical result such as Azuma-Hoeffding's inequality (for independent sequences) and McDiarmid's inequality (for sequences with bounded difference) do not apply. To resolve this, we instead use an extension of McDiarmid's inequality which relies on boundedness with high probability rather than almost sure [3].

To get the Rademacher complexity for the class defined in Eq. (3), note that the weight function w_i brings in the centers of other clusters $\boldsymbol{\mu}_j$, and f^i is not Lipschitz in terms of $\boldsymbol{\mu}$. Therefore the classical contraction lemma does not apply. In our analysis, we need to introduce a vector-valued function, with each element involving only one $\boldsymbol{\mu}_i$, and apply a recent result of vector-versioned contraction lemma [12]. With some careful treatment, we get the following.

Proposition 1. *Let $\mathcal{F}_i^u$ as defined in Eq. (3) for $\forall i \in [M]$, then for some universal constant c,*

$$R_n(\mathcal{F}_i^u) \le \frac{cM^{3/2}(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}}{\sqrt{n}}$$

Combined with the concentration of the empirical process, we have our second main result.

Theorem 2 (Convergence for sample based gradient EM)**.** *Let $\zeta := \frac{\pi_{\max}-\pi_{\min}+2\gamma}{\pi_{\max}+\pi_{\min}}$ be the contraction parameter in Theorem 3. If $\epsilon^{unif}(n) \le (1-\zeta)a$, then sample-based gradient EM satisfies*

$$\left\|\hat{\boldsymbol{\mu}}_i^t - \boldsymbol{\mu}_i^*\right\| \le \zeta^t \left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\|_2 + \frac{1}{1-\zeta}\epsilon^{unif}(n); \quad \forall i \in [M]$$

with probability at least $1-n^{-cd}$, where c is positive constant.

4 Local Convergence of Population Gradient EM

In this section we present the proof sketch for Theorem 1. The complete proofs in this section are deferred to Appendix B. To start with, we calculate the close-form characterization of the gradient and Hessian of $q(\boldsymbol{\mu})$ as stated in the following lemma.

Lemma 1. *Define $q(\boldsymbol{\mu}) = Q(\boldsymbol{\mu}|\boldsymbol{\mu}^*)$. The gradient of $q(\boldsymbol{\mu})$ is $\nabla q(\boldsymbol{\mu}) = (diag(\pi) \otimes I_d)(\boldsymbol{\mu}^* - \boldsymbol{\mu})$.*

If we know the parameter γ in the gradient stability condition, then the convergence rate depends only on the condition number of the Hessian of $q(\cdot)$ and γ.

Theorem 3 (Convergence rate for population gradient EM)**.** *If Q satisfies the GS condition with parameter $0 < \gamma < \pi_{\min}$, $d_t := \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|$, then with step size $s = \frac{2}{\pi_{\min}+\pi_{\max}}$ guarantees,*

$$d_{t+1} \le \left(\frac{\pi_{\max}-\pi_{\min}+2\gamma}{\pi_{\max}+\pi_{\min}}\right)^t d_0$$

The proof uses an approximation on gradient and standard techniques in analysis of gradient descent.

Remark 2. *It can be verified that the convergence rate is equivalent as that shown in [1] when applied to GMMs. The convergence slows down as the proportion imbalance $\kappa = \pi_{\max}/\pi_{\min}$ increases, which matches the observation in [16].*

Now to verify the GS condition, we have the following theorem.

Theorem 4 (GS condition for general GMM)**.** *If* $R_{\min} = \tilde{\Omega}(\sqrt{\min\{d, M\}})$, *and* $\boldsymbol{\mu}_i \in \mathbb{B}(\boldsymbol{\mu}_i^*, a), \forall i \in [M]$ *where* $a \leq \frac{R_{\min}}{2} - \sqrt{\min\{d, M\}} \max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3})$, *then* $\|\nabla_{\boldsymbol{\mu}_i} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla_{\boldsymbol{\mu}_i} q(\boldsymbol{\mu})\| \leq \frac{\gamma}{M} \sum_{i=1}^M \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\| \leq \frac{\gamma}{\sqrt{M}} \|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$,
where $\gamma = M^2(2\kappa + 4)\left(2R_{\max} + \min\{d, M\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)$.
Furthermore, $\|\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu})\| \leq \gamma \|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$.

Proof sketch of Theorem 4. W.l.o.g. we show the proof with the first cluster, consider the difference of the gradient corresponding to $\boldsymbol{\mu}_1$.

$$\nabla_{\boldsymbol{\mu}_1} Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu}^t) = \mathbb{E}(w_1(X; \boldsymbol{\mu}^t) - w_1(X; \boldsymbol{\mu}^*))(X - \boldsymbol{\mu}_1^t) \tag{4}$$

For any given X, consider the function $\boldsymbol{\mu} \to w_1(X; \boldsymbol{\mu})$, we have

$$\nabla_{\boldsymbol{\mu}} w_1(X; \boldsymbol{\mu}) = \begin{pmatrix} w_1(X; \boldsymbol{\mu})(1 - w_1(X; \boldsymbol{\mu}))(X - \boldsymbol{\mu}_1)^T \\ -w_1(X; \boldsymbol{\mu}) w_2(X; \boldsymbol{\mu})(X - \boldsymbol{\mu}_2)^T \\ \vdots \\ -w_1(X; \boldsymbol{\mu}) w_M(X; \boldsymbol{\mu})(X - \boldsymbol{\mu}_M)^T \end{pmatrix} \tag{5}$$

Let $\boldsymbol{\mu}^u = \boldsymbol{\mu}^* + u(\boldsymbol{\mu}^t - \boldsymbol{\mu}^*), \forall u \in [0, 1]$, obviously $\boldsymbol{\mu}^u \in \Pi_{i=1}^M \mathbb{B}(\boldsymbol{\mu}_i^*, \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|) \subset \Pi_{i=1}^M \mathbb{B}(\boldsymbol{\mu}_i^*, a)$. By Taylor's theorem,

$$\begin{aligned} \|\mathbb{E}(w_1(X; \boldsymbol{\mu}_1^t) - w_1(X; \boldsymbol{\mu}_1^*))(X - \boldsymbol{\mu}_1^t)\| &= \left\| \mathbb{E}\left[\int_{u=0}^1 \nabla_u w_1(X; \boldsymbol{\mu}^u) du (X - \boldsymbol{\mu}_1^t)\right] \right\| \\ \leq U_1 \|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\|_2 + \sum_{i \neq 1} U_i \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2 &\leq \max_{i \in [M]}\{U_i\} \sum_i \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2 \end{aligned} \tag{6}$$

where

$$\begin{aligned} U_1 &= \sup_{u \in [0,1]} \|\mathbb{E} w_1(X; \boldsymbol{\mu}^u)(1 - w_1(X; \boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\ U_i &= \sup_{u \in [0,1]} \|\mathbb{E} w_1(X; \boldsymbol{\mu}^u) w_i(X; \boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_2^u)^T\|_{op} \end{aligned}$$

Bounding them with careful analysis on Gaussian distribution yields the result. The technical details are deferred to Appendix B. □

5 Sample-based Convergence

In this section we present the proof sketch for sample-based convergence of gradient EM. The main ingredient in proof of Theorem 2 is the result of the following theorem, which develops an uniform upper bound for the differences between sample-based gradient and population gradient on each cluster center. Iteratively applying this bound, we can bound the error in $\boldsymbol{\mu}$ for the sample updates.

Theorem 5 (Uniform bound for sample-based gradient EM)**.** *Denote* $\mathbb{A}$ *as the contraction region* $\Pi_{i=1}^M \mathbb{B}(\boldsymbol{\mu}_i^*, a)$. *Under the condition of Theorem 1, with probability at least* $1 - \exp(-cd \log n)$,

$$\sup_{\boldsymbol{\mu} \in \mathbb{A}} \left\| G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu}) \right\| < \epsilon^{unif}(n); \qquad \forall i \in [M]$$

where $\epsilon^{unif}(n) = \tilde{O}(\max\{n^{-1/2} M^3 (1 + R_{\max})^3 \sqrt{d} \max\{1, \log(\kappa)\}, (1 + R_{\max}) d/\sqrt{n}\})$.

When the function class has bounded differences (changing one data point changes the function by a bounded amount almost surely), as in the case in many risk minimization problems in supervised learning, the Rademacher complexity can be used to achieve concentration. Now recall the function class we define in Eq. (3) is not bounded, hence the classical result does not apply. The key step missing here is the bounded difference requirement of McDiarmid's inequality. However, in our case we have functions which only have (almost) bounded differences with high probability, albeit not almost surely. So we use an extension of McDiarmid's inequality by [3]. For convenience, we restate a weaker version of the theorem using our notation below.

Theorem 6 ([3])**.** *Consider independent random variable $X = (X_1, \cdots, X_n)$ in the product probability space $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$, where $\mathcal{X}_i$ is the probability space for X_i. Also consider a function $g : \mathcal{X} \to \mathbb{R}$. If there exists a subset $\mathcal{Y} \subset \mathcal{X}$, and a scalar $c > 0$, such that*

$$|g(x) - g(y)| \leq L, \forall x, y \in \mathcal{Y}, x_j = y_j, \forall j \neq i.$$

Denote $p = 1 - P(X \in \mathcal{Y})$,then $P(g(X) - \mathbb{E}[g(X)|X \in \mathcal{Y}] \geq \epsilon) \leq p + \exp\left(-\frac{2(\epsilon - npL)_+^2}{nL^2}\right)$.

It is worth pointing out that in Theorem 6, the concentration is shown in reference to the conditional expectation $\mathbb{E}[g(X)|X \in \mathcal{Y}]$ when the data points are in the bounded difference set. So to fully achieve the type of bound given by McDiarmid's inequality, we need to further bound the difference of the conditional expectation and the full expectation. Combining the two parts we will be able to show that, the empirical difference is upper bounded using the Rademacher complexity.

Lemma 2. *Let u be a unit vector. $X_i, i = 1, \cdots, n$ are i.i.d. samples from GMM$(\pi, \boldsymbol{\mu}^*, I_d)$. Define*

$$g(X) = \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n w_1(X_i; \boldsymbol{\mu}) \langle X_i - \boldsymbol{\mu}_1, u \rangle - \mathbb{E} w_1(X; \boldsymbol{\mu}) \langle X - \boldsymbol{\mu}_1, u \rangle.$$

Then with probability $1 - \exp(-cd \log n)$ for some constant $C > 0$, $g(X) = \tilde{O}(\max\{R_n(\mathcal{F}_1^u), (1 + R_{\max}) d / \sqrt{n}\})$.

Now we derive the Rademacher complexity under the given function class. Note that when the function class is a contraction, or Lipschitz with respect to another function (usually of a simpler form), one can use the Ledoux-Talagrand contraction lemma [10] to reduce the Rademacher complexity of the original function class to the Rademacher complexity of the simpler function class. This is essential in getting the Rademacher complexities for complicated function classes. As we mention in Section 3, our function class in Eq. (3) is unfortunately not Lipschitz due to the fact that it involves all cluster centers even for the gradient on one cluster. We get around this problem by introducing a vector valued function, and show that the functions in Eq. (3) are Lipschitz in terms of the vector-valued function. In other words, the absolute difference in the function when the parameter changes is upper bounded by the norm of the vector difference of the vector-valued function. Then we use the recent result from [12], which proves a vector contraction theorem. For convenience, we restate it below in accordance of our notation.

Lemma 3 (Theorem 3 [12])**.** *Let X be nontrivial, symmetric and subgaussian. Then there exists a constant $C < \infty$, depending only on the distribution of X, such that for any countable set $\mathcal{S}$ and function $h_i : \mathcal{S} \to \mathbb{R}$, $f_i : \mathcal{S} \to \mathbb{R}^k$, $i \in [n]$ satisfying $\forall s, s' \in \mathcal{S}, |h_i(s) - h_i(s')| \leq L\|f(s) - f(s')\|$. If ϵ_{ik} is an independent doubly indexed Rademacher sequence, we have,*

$$\mathbb{E} \sup_{s \in \mathcal{S}} \sum_i \epsilon_i h_i(s) \leq \mathbb{E}\sqrt{2} L \sup_{s \in \mathcal{S}} \sum_{i,k} \epsilon_{ik} f_i(s)_k,$$

where $f_i(s)_k$ is the k-th component of $f_i(s)$.

Remark 3. *Note that in contrast to Lemma 3, we have a $\mathcal{S}$ as a subset of a separable Banach Space. We can easily modify the above theorem to introduce the following lemma that suits our setting using standard tools from measure theory.*

This equips us to prove Proposition 1. We show the proof sketch below and the full proof is to be found in Appendix C.

Proof of Proposition 1. For any unit vector u, the Rademacher complexity of $\mathcal{F}_1^u$ is

$$\begin{aligned} R_n(\mathcal{F}_1^u) =& \mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \epsilon_i w_1(X_i;\boldsymbol{\mu})\langle X_i - \boldsymbol{\mu}_1, u\rangle \\ \leq& \underbrace{\mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \epsilon_i w_1(X_i;\boldsymbol{\mu})\langle X_i, u\rangle}_{(D)} + \underbrace{\mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \epsilon_i w_1(X_i;\boldsymbol{\mu})\langle \boldsymbol{\mu}_1, u\rangle}_{(E)} \end{aligned} \tag{7}$$

We bound the two terms separately. Define $\eta_j(\boldsymbol{\mu}) : \mathbb{R}^{Md} \to \mathbb{R}^M$ to be a vector valued function with the k-th coordinate

$$[\eta_j(\boldsymbol{\mu})]_k = \frac{\|\boldsymbol{\mu}_1\|^2}{2} - \frac{\|\boldsymbol{\mu}_k\|^2}{2} + \langle X_j, \boldsymbol{\mu}_k - \boldsymbol{\mu}_1\rangle + \log\left(\frac{\pi_k}{\pi_1}\right)$$

It can be shown that $$|w_1(X_j;\boldsymbol{\mu}) - w_1(X_j;\boldsymbol{\mu}')| \leq \frac{\sqrt{M}}{4}\left\|\eta_j(\boldsymbol{\mu}) - \eta_j(\boldsymbol{\mu}')\right\| \tag{8}$$

Now let $\psi_1(X_j;\boldsymbol{\mu}) = w_1(X_j;\boldsymbol{\mu})\langle X_j, u\rangle$. With Lipschitz property (27) and Lemma 3, we have

$$\mathbb{E}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{j=1}^n \epsilon_j w_i(X_j;\boldsymbol{\mu})\langle X_j, u\rangle\right] \leq \mathbb{E}\left[\frac{\sqrt{2}\sqrt{M}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^n\sum_{k=1}^M \epsilon_{jk}[\eta_j(\boldsymbol{\mu})]_k\right]$$

The right hand side can be bounded with tools regarding independent sum of sub-gaussian random variables. Similar techniques apply to (E) term. Adding things up we get the final bound. □

Combining the pieces we can now prove Theorem 5.

Proof sketch of Theorem 5. Denote $Z_i = \sup_{\boldsymbol{\mu}\in\mathbb{A}}\left\|G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu})\right\|$. Consider a $1/2-$covering of the unit sphere $\mathcal{S}^{d-1}$ $\{u^{(1)},\cdots,u^{(K)}\}$, then $Z \leq 2\max_{j=1,\cdots,K} Z_{u^{(j)}}$.
By Lemma 2, we have with probability at least $1-\exp(-cd\log n)$, $Z_{u_j} = \tilde{O}(\max\{R_n(\mathcal{F}_1^u), (1+R_{\max})d/\sqrt{n}\})$. Plugging in the Rademacher complexity from Proposition 1, and applying union bound, we have

$$Z_1 \leq 2\max_j Z_{u_j} \leq \tilde{O}(\max\{n^{-1/2}M^3(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}, (1+R_{\max})d/\sqrt{n}\})$$

with probability at least $1-\exp(2d - cd\log n) = 1-\exp(-c'd\log n)$. □

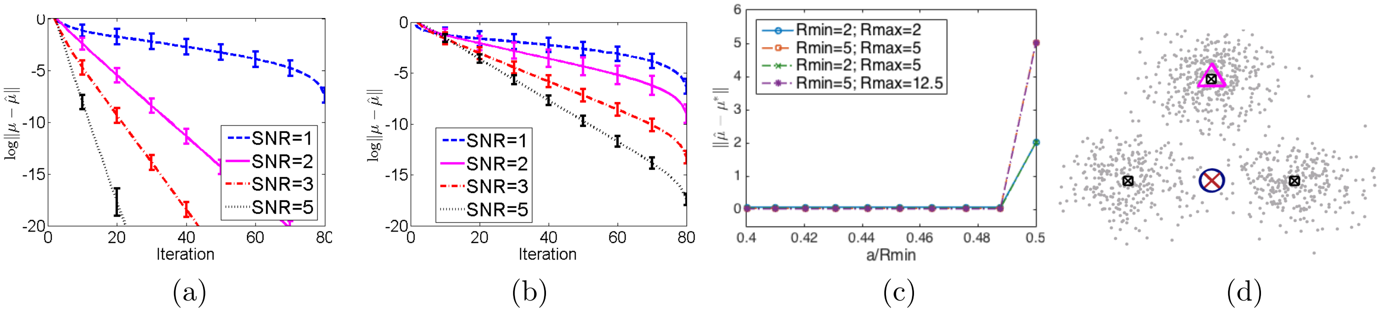
Remark 4. *It is worth pointing out that, the first part of the bound comes from the Rademacher complexity, which is optimal in terms of the order of n and d. However the extra factor of $\sqrt{d}$ and $\log(n)$ comes from the altered McDiarmids' inequality, tightening which is part of ongoing work.*

6 Experiments

In this section we collect some numerical results. In all experiments we set the covariance matrix for each mixture component as identity matrix I_d and define signal-to-noise ratio (SNR) as $R_{\min}$.

Convergence Rate We first evaluate the convergence rate and compare with those given in Theorem 3 and Theorem 4. For this set of experiments, we use a mixture of 3 Gaussians in 2 dimensions. In both experiments $R_{\max}/R_{\min} = 1.5$. In different settings of $\boldsymbol{\pi}$, we apply gradient EM with varying SNR from 1 to 5. For each choice of SNR, we perform 10 independent trials with $N = 12,000$ data points. The average of $\log \left\|\boldsymbol{\mu}^t - \hat{\boldsymbol{\mu}}\right\|$ and the standard deviation are plotted versus iterations. In Figure 1 (a) and (b) we plot balanced $\boldsymbol{\pi}$ ($\kappa = 1$) and unbalanced $\boldsymbol{\pi}$ ($\kappa > 1$) respectively.

All settings indicate the linear convergence rate as shown in Theorem 3. As SNR grows, the parameter γ in GS condition decreases and thus yields faster convergence rate. Comparing left two panels in Figure 1, increasing imbalance of cluster weights κ slows down the local convergence rate as shown in Theorem 3.



(a) (b) (c) (d)

Figure 1: (a, b): The influence of SNR on optimization error in different settings. The figures represent the influence of SNR when the GMM have different cluster centers and weights. (a) $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. (b) $\boldsymbol{\pi} = (0.6, 0.3, 0.1)$. (c) plots statistical error with different initializations arbitrarily close to the boundary of the contraction region. (d) shows the suboptimal stationary point when two centers are initialized from the midpoint of the respective true cluster centers.

Contraction Region To show the tightness of the contraction region, we generate a mixture with $M = 3, d = 2$, and initialize the clusters as follows. We use $\boldsymbol{\mu}_2^0 = \frac{\boldsymbol{\mu}_2^* + \boldsymbol{\mu}_3^*}{2} - \epsilon$, $\boldsymbol{\mu}_3^0 = \frac{\boldsymbol{\mu}_2^* + \boldsymbol{\mu}_3^*}{2} + \epsilon$, for shrinking ϵ, i.e. increasing $a/R_{\min}$ and plot the error on the Y axis. Figure 1-(c) shows that gradient EM converges when initialized arbitrarily close to the boundary, thus confirming our near optimal contraction region. Figure 1-(d) shows that when $\epsilon = 0$, i.e. $a = \frac{R_{\min}}{2}$, gradient EM will be trapped at a sub-optimal stationary point.

7 Concluding Remarks

In this paper, we obtain local convergence rates and a near optimal contraction radius for population and sample-based gradient EM for GMM with general cluster number and weights. For our sample based analysis, we face new challenges because of structural differences from the two component equal weight setting, which are alleviated via the use of non-traditional tools like a vector valued contraction argument and martingale concentrations bounds where bounded differences hold only with high probability.

A Accompanying Lemmas

In this subsection, we collect some lemmas on Gaussian distribution and basic properties of Gaussian mixture model. Most of them can be derived with fundamental analysis techniques. The following lemma from [18] bounds the covering number of a unit sphere.

Lemma 4 (Lemma 5.2 [18])**.** *Let $\mathcal{S}^{n-1}$ be the unit Euclidean sphere equipped with Euclidean metric. Denote $\mathcal{N}(\mathcal{S}^{n-1}, \epsilon)$ as the covering number with ϵ-net, then*

$$\mathcal{N}(\mathcal{S}^{n-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^n$$

Specifically, when $\epsilon = 1/2$, we have

$$\mathcal{N}(\mathcal{S}^{n-1}, \frac{1}{2}) \leq \exp(2n)$$

The following lemma is useful while carrying out spherical coordinate transformation.

Lemma 5. *(1) The volume for a d-dimensional r-ball is $\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d$;*

(2) $\int_0^\pi \sin^k(x)dx = \frac{\sqrt{\pi}\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2}+1)}$, and

$$\int_{\theta_{d-1}=0}^{2\pi} \int_{\theta_{d-2}=0}^{\pi} \cdots \int_{\theta_1=0}^{\pi} \sin^{d-2}(\theta_1) \cdots \sin(\theta_{d-2}) d\theta_1 \cdots d\theta_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

(3) If $X \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$, then

$$\mathbb{E}_X \|X - \boldsymbol{\mu}\|^p = 2^{\frac{p}{2}} \frac{\Gamma\left(\frac{p+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \sigma^p$$

Proof. (1, 2) can be proven by elementary integration. Now we prove (3). By spherical coordinate transformation,

$$\mathbb{E}_X \|X - \boldsymbol{\mu}\|^p = (2\pi\sigma^2)^{-\frac{d}{2}} \int_{u=0}^{\infty} u^{p+d-1} e^{-\frac{u^2}{2\sigma^2}} du \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} = 2^{\frac{p}{2}} \frac{\Gamma\left(\frac{p+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \sigma^p$$

□

Lemma 6 (Gamma tail bound [2])**.** *If $X \sim Gamma(v, c)$, then $P(X > \sqrt{2vt} + ct) \leq e^{-t}$. Or equivalently,*

$$P(X > t) \leq \exp\left(-\frac{v}{c^2}\left(1 + \frac{ct}{v} - \sqrt{1 + \frac{2ct}{v}}\right)\right)$$

In particular, if $\frac{ct}{v} \geq 4$,

$$P(X > t) \leq \exp\left(-\frac{v}{c^2}\sqrt{\frac{ct}{v}}\right) = \exp\left(-\sqrt{\frac{vt}{c^3}}\right)$$

Lemma 7. *For $\forall d > 0$, if $r \geq 2\sqrt{d+1}$, then*

$$\int_r^\infty u^d e^{-\frac{u^2}{2}} du \leq 2^{\frac{d-1}{2}} \Gamma\left(\frac{d+1}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d+1}\right)$$

For $p \in \{0, 1, 2\}$, when $r \geq 2\sqrt{d+p}$,

$$\int_r^\infty (u+x)^p u^{d-1} e^{-\frac{u^2}{2}} du \leq 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) (x+d)^p \exp\left(-\frac{r}{2}\sqrt{d}\right)$$

Proof. By changing of variables $v = \frac{u^2}{2}$ and integration by parts, we have

$$\begin{aligned}\int_r^\infty u^d e^{-\frac{u^2}{2}} du =& 2^{\frac{d-1}{2}} \int_{\frac{r^2}{2}}^\infty v^{\frac{d-1}{2}} e^{-v} dv \\ =& 2^{\frac{d-1}{2}} \Gamma\left(\frac{d+1}{2}\right) P(V > \frac{r^2}{2})\end{aligned}$$

where $V \sim \text{Gamma}(\frac{d+1}{2}, 1)$. By Lemma 6, if $r^2 \geq 4(1+d)$,

$$P\left(V > \frac{r^2}{2}\right) \leq \exp\left(-\frac{r}{2}\sqrt{d+1}\right)$$

Hence we have the first inequality. For the second, when $p = 0$, it follows directly from first part. When $p = 1$,

$$\begin{aligned}\int_r^\infty (u+x)^p u^{d-1} e^{-\frac{u^2}{2}} du =& \int_r^\infty u^d e^{-\frac{u^2}{2}} du + x \int_r^\infty u^{d-1} e^{-\frac{u^2}{2}} du \\ \leq& 2^{\frac{d-1}{2}} \Gamma\left(\frac{d+1}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d+1}\right) + x 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right) \\ \leq& 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) (x+d) \exp\left(-\frac{r}{2}\sqrt{d}\right)\end{aligned}$$

where we use $\Gamma\left(\frac{d+1}{2}\right) < \Gamma\left(\frac{d}{2}+1\right) = \frac{d}{2}\Gamma\left(\frac{d}{2}\right)$, and $\exp\left(-\frac{r}{2}\sqrt{d+1}\right) < \exp\left(-\frac{r}{2}\sqrt{d}\right)$ in the last step. When $p = 2$,

$$\begin{aligned}&\int_r^\infty (u+x)^2 u^{d-1} e^{-\frac{u^2}{2}} du = \int_r^\infty u^{d+1} e^{-\frac{u^2}{2}} du + 2x \int_r^\infty u^d e^{-\frac{u^2}{2}} du \\ &+ x^2 \int_r^\infty u^{d-1} e^{-\frac{u^2}{2}} du \\ &\leq 2^{\frac{d}{2}} \Gamma\left(\frac{d}{2}+1\right) \exp\left(-\frac{r}{2}\sqrt{d+2}\right) + 2x \cdot 2^{\frac{d-1}{2}} \Gamma\left(\frac{d+1}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d+1}\right) + x^2 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right) \\ &\leq (d + \sqrt{2}dx + x^2) 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right) \\ &\leq (x+d)^2 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right)\end{aligned}$$

□

Using Lemma 7, we can get an easy to use tail bound for Euclidean norm of a Gaussian vector.

Lemma 8. *If $X \sim \mathcal{N}(0, I_d)$, for $r \geq 2\sqrt{d}$, we have*

$$P(\|X\| \geq r) \leq \exp(-\frac{r\sqrt{d}}{2})$$

Proof. By spherical coordinate transformation,

$$\begin{aligned}P(\|X\| \geq r) =& \int (2\pi)^{-d/2} \exp(-\|x\|^2/2) dx \\ =& (2\pi)^{-d/2} \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_r^\infty r^{d-1} e^{-r^2/2} dr \\ \leq& \exp\left(-\frac{r}{2}\sqrt{d}\right)\end{aligned}$$

□

Lemma 9. *If $X \sim GMM(\pi, \boldsymbol{\mu}^*, \sigma^2 I_d)$, then X is a sub-gaussian random vector with sub-gaussian norm $\sigma + \sum_{i=1}^M \pi_i \|\boldsymbol{\mu}_i^*\|$.*

Proof. For any unit vector u, consider the random variable $X_u = \langle X, u \rangle$. By the definition in [18], it suffices to show that X_u has a sub-gaussian norm upper bounded by $\sigma + \sum_{i=1}^M \pi_i \|\boldsymbol{\mu}_i^*\|$.

$$\|X_u\|_{\phi_2} = \sup_{p \geq 1} (\mathbb{E}|X_u|^p)^{1/p}$$

For any $p \geq 1$, let Z be the latent variable in the mixture model, we have

$$\begin{aligned}
p^{-1/2} \left(\mathbb{E}|X_u|^p\right)^{1/p} =& p^{-1/2} \left(\sum_{i=1}^M \mathbb{E}[|X_u|^p | Z = i] \cdot P(Z = i) \right)^{1/p} \\
\leq& p^{-1/2} \sum_{i=1}^M \pi_i \left(\mathbb{E}[|X_u|^p | Z = i]\right)^{1/p} \\
\overset{(i)}{\leq}& p^{-1/2} \sum_{i=1}^M \pi_i \left(\mathbb{E}[|X_u - \boldsymbol{\mu}_i^*|^p | Z = i]^{1/p} + \|\boldsymbol{\mu}_i^*\| \right) \\
\leq& p^{-1/2} \left(\sum_{i=1}^M \pi_i p^{1/2} \sigma + \|\boldsymbol{\mu}_i^*\| \right) \leq \sigma + \sum_{i=1}^M \pi_i \|\boldsymbol{\mu}_i^*\|
\end{aligned}$$

where (i) follows from Minkovski's inequality. □

The following lemma characterize the relation between $\|\boldsymbol{\mu}_{\max}^*\|$ and $R_{\max}$.

Lemma 10. *If $X \sim GMM(\pi, \boldsymbol{\mu}^*, \sigma^2 I_d)$ with $\mathbb{E}X = 0$, let $\|\boldsymbol{\mu}_{\max}^*\| = \max_i \|\boldsymbol{\mu}_i^*\|$, then*

$$\|\boldsymbol{\mu}_{\max}^*\| \leq R_{\max} \leq 2\|\boldsymbol{\mu}_{\max}^*\|$$

Proof. We first prove $\|\boldsymbol{\mu}_{\max}^*\| \leq R_{\max}$ by contradiction. Assume $\|\boldsymbol{\mu}_{\max}^*\| > R_{\max}$, by definition of $R_{\max}$, all the cluster centers lies in the ball $\mathbb{B}(\|\boldsymbol{\mu}_{\max}^*\|, R_{\max})$, but the origin is outside of the ball, which contradicts the fact that $\mathbb{E}X = \sum_i \pi_i \boldsymbol{\mu}_i^* = 0$.
The second inequality follows from triangle inequality, assume $R_{\max}$ is achieved at R_{ij}, then

$$R_{\max} \leq \|\boldsymbol{\mu}_i^*\| + \|\boldsymbol{\mu}_j^*\| \leq 2\|\boldsymbol{\mu}_{\max}^*\|.$$

□

Lemma 11. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz if and only if there exist a constant L such that the restriction of f on a certain coordinate is Lipschitz with constant L.*

Proof. We prove $n = 2$ and general n can be proved in the same way.

$$\begin{aligned}
|f(x_1, x_2) - f(y_1, y_2)| \leq& |f(x_1, x_2) - f(y_1, x_2)| + |f(y_1, x_2) - f(y_1, y_2)| \\
\leq& L|x_1 - x_2| + L|y_1 - y_2| \\
\leq& \sqrt{2} L \|x - y\|
\end{aligned}$$

□

B Proofs in Section 4

Proof of Lemma 1. By (2), $\nabla_{\boldsymbol{\mu}_i} q(\boldsymbol{\mu}) = \mathbb{E}_X w_i(X; \boldsymbol{\mu}^*)(X - \boldsymbol{\mu}_i)$. Without loss of generality, we only show the claim for $i = 1$. That is equivalent of saying, if $X \sim \text{GMM}(\pi, \boldsymbol{\mu}^*)$, we have $\mathbb{E}[w_1(X; \boldsymbol{\mu}^*)(X - \boldsymbol{\mu}_1^*)] = 0$. Denote $\mathcal{N}(\boldsymbol{\mu}_i^*, \Sigma)$ as $\mathcal{N}_i$ and its distribution as $\phi_i(X)$. Decompose the left hand side with respect to the mixture components, we have

$$\begin{aligned}\mathbb{E}[w_1(X)X] =& \sum_i \pi_i \mathbb{E}_{X\sim\mathcal{N}_i}[w_1(X)X] \\ =& \sum_i \pi_i \int \phi_i(X) \frac{\pi_1\phi_1(X)}{\sum_k \pi_k \phi_k(X)} X dx \\ =& \pi_1 \mathbb{E}_{X\sim\mathcal{N}_1} X = \pi_1 \boldsymbol{\mu}_1^*\end{aligned}$$

Similarly $\mathbb{E}[w_1(X)] = \pi_1$. Hence $\nabla_{\boldsymbol{\mu}_1} q(\boldsymbol{\mu}) = \mathbb{E}_X w_1(X; \boldsymbol{\mu}^*)(X - \boldsymbol{\mu}_1) = \pi_1(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_1)$.
This completes the proof. □

Proof of Theorem 3. Define By Lemma 1, the GS condition is equivalent to

$$\left\|\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu})\right\| \le \gamma \|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$$

By triangle inequality,

$$\begin{aligned}\left\|\boldsymbol{\mu}_1^{t+1} - \boldsymbol{\mu}_1^*\right\| =& \left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^* + s\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t)\right\| \\ \le& \left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^* + s\nabla q(\boldsymbol{\mu})\right\| + s\left\|\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu})\right\| \\ \le& \frac{\pi_{\max} - \pi_{\min}}{\pi_{\max} + \pi_{\min}} \left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\right\| + \frac{2}{\pi_{\max} + \pi_{\min}} \gamma \left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\right\| \\ \le& \frac{\pi_{\max} - \pi_{\min} + 2\gamma}{\pi_{\max} + \pi_{\min}} \left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\right\|\end{aligned}$$

To see why the last inequality hold, notice that $q(\boldsymbol{\mu})$ has largest eigenvalue $-\pi_{\min}$ and smallest eigenvalue $-\pi_{\max}$. Apply the classical result for gradient descent, with step size $s = \frac{2}{\pi_{\max}+\pi_{\min}}$ guarantees

$$\left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^* + s\nabla q(\boldsymbol{\mu})\right\| \le \frac{\pi_{\max} - \pi_{\min}}{\pi_{\max} + \pi_{\min}} \left\|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\right\|$$

□

B.1 Proofs of Theorem 4

We start with two lemmas.

Lemma 12. *For $X \sim GMM(\pi, \boldsymbol{\mu}^*, I_d)$, if $R_{\min} = \tilde{\Omega}(\sqrt{d})$, and $\boldsymbol{\mu}_i \in \mathbb{B}(\boldsymbol{\mu}_i^*, a), \forall i \in [M]$ where*

$$a \le \frac{R_{\min}}{2} - \sqrt{d}\max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3}).$$

Then for $p = 0, 1, 2$ and $\forall i \in [M]$, we have

$$\mathbb{E}_X w_i(X; \boldsymbol{\mu})(1 - w_i(X; \boldsymbol{\mu}))\|X - \boldsymbol{\mu}_i\|^p \le 2M\left(\frac{3}{2}R_{\max} + d\right)^p \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{d}/8\right)$$

Using the same techniques, for the cross terms, we have the following lemma.

Lemma 13. *Assume $X \sim GMM(\pi, \boldsymbol{\mu}^*, I_d)$, and $\boldsymbol{\mu}_i \in \mathbb{B}(\boldsymbol{\mu}_i^*, a), \forall i \in [M]$. Under the same conditions as in Lemma 12, we have for $\forall i \neq j \in [M]$,*

$$\mathbb{E}_X[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_i\| \cdot \|X-\boldsymbol{\mu}_j\|] \leq (1+2\kappa)\left(\frac{3}{2}R_{\max}+d\right)^2 \exp\left(-\left(\frac{R_{\min}}{2}-a\right)^2\sqrt{d}/8\right)$$

Proof of Lemma 12. Without loss of generality, we prove the claim for $i = 1$. Recall the definition of $w_i(X;\boldsymbol{\mu})$ from Equation 1. For $p \in \{0, 1, 2\}$,

$$\begin{aligned}
&\mathbb{E}_X w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
=&\sum_{i\in[M]} \pi_i \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)} w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
\leq&\pi_1 \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*)} w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p + \sum_{i\neq 1}\pi_i \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)} w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p
\end{aligned} \tag{9}$$

First let us look at the first term. Define event $\mathcal{E}_r^{(1)} = \{X : X \sim \mathcal{N}(\boldsymbol{\mu}_1^*); \|X-\boldsymbol{\mu}_1^*\| \leq r\}$ for some $r > 0$. We will see later that we need $r < \frac{R_{\min}}{2} - a$. Then for $X \in \mathcal{E}_r^{(1)}$ using triangle inequality, we have

$$\|X-\boldsymbol{\mu}_i\| \begin{cases} \leq \|X-\boldsymbol{\mu}_i^*\| + \|\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_i\| \leq r+a & i=1 \\ \geq \|\boldsymbol{\mu}_i-\boldsymbol{\mu}_1^*\| - \|X-\boldsymbol{\mu}_1^*\| \geq \|\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_1^*\| - \|\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_i\| - r \geq R_{\min}-r-a & i\neq 1 \end{cases} \tag{10}$$

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*)} w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
=&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p|\mathcal{E}_r^{(1)}]P(\mathcal{E}_r^{(1)}) \\
&+\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p|\mathcal{E}_r^{(1)c}]P(\mathcal{E}_r^{(1)c})
\end{aligned}$$

In view of the fact that $w_1(X;\boldsymbol{\mu})$ is monotonically decreasing w.r.t. $\|X-\boldsymbol{\mu}_i\|$ and increasing w.r.t. $\|X-\boldsymbol{\mu}_1\|$, we have

$$\begin{aligned}
1-w_1(X;\boldsymbol{\mu}) \leq& \frac{(1-\pi_1)\exp\left(-\frac{(R_{\min}-r-a)^2}{2}\right)}{\pi_1\exp\left(-\frac{(r+a)^2}{2}\right) + (1-\pi_1)\exp\left(-\frac{(R_{\min}-r-a)^2}{2}\right)} \\
\leq& \frac{1-\pi_1}{\pi_1}\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)
\end{aligned}$$

Also notice that $w_1(X;\boldsymbol{\mu}) \leq 1$, we have

$$\begin{aligned}
&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p|\mathcal{E}_r^{(1)}]P(\mathcal{E}_r^{(1)}) \\
\leq&\frac{1-\pi_1}{\pi_1}\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)(r+a)^p
\end{aligned}$$

For $\mathcal{E}_r^{(1)c}$, note $w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu})) \leq \frac{1}{4}$, we have for $p = 1$,

$$\begin{aligned}
&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\||\mathcal{E}_r^{(1)c}]P(\mathcal{E}_r^{(1)c}) \\
\leq&\frac{1}{4}\int_{u=r}^{\infty}(u+a)(2\pi)^{-\frac{d}{2}}\exp\left(-\frac{u^2}{2}\right)\cdot\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}u^{d-1}du \\
\leq&\frac{1}{4}(2\pi)^{-\frac{d}{2}}\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}\int_{u=r}^{\infty}(u+a)\exp\left(-\frac{u^2}{2}\right)u^{d-1}du \\
\overset{(i)}{\leq}&\frac{a+d}{4}\exp\left(-\frac{r}{2}\sqrt{d}\right)
\end{aligned}$$

The inequality (i) follows from Lemma 7 when $r > 2\sqrt{d+1}$. Similarly, for $p = 2$,

$$\begin{aligned}&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^2|\mathcal{E}_r^{(1)c}]P(\mathcal{E}_r^{(1)c})\\ \leq&\frac{2^{-\frac{d}{2}-1}}{\Gamma\left(\frac{d}{2}\right)}\int_r^\infty (u+a)^2u^{d-1}e^{-\frac{u^2}{2}}du \overset{(ii)}{\leq} \frac{(a+d)^2}{4}\exp\left(-\frac{r}{2}\sqrt{d}\right)\end{aligned}$$

The inequality (ii) follows from Lemma 7 when $r > 2\sqrt{d+1}$ and $p = 2$. Therefore for the first mixture we have,

$$\begin{aligned}&\pi_1\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*)}w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p\\ \leq&(1-\pi_1)(r+a)^p\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)+\pi_1\frac{(a+d)^p}{4}\exp\left(-\frac{r}{2}\sqrt{d}\right)\end{aligned} \tag{11}$$

Next we bound $\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)}w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p$ for $i \neq 1$. For some $0 < r < \frac{R}{2} - a$, we have

$$\begin{aligned}&\pi_i\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)}w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p\\ =&\int_X \frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\cdot\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j\pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX\\ =&\underbrace{\int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,r)}\frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\cdot\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j\pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX}_{I_1^{(p)}}+\underbrace{\int_{X\notin\mathbb{B}(\boldsymbol{\mu}_i^*,r)}\frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\cdot\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j\pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX}_{I_2^{(p)}}\end{aligned} \tag{12}$$

When $\|X-\boldsymbol{\mu}_i^*\| \leq r$, since by assumption $\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*\| \leq a$,

$$\begin{aligned}\frac{\phi(X;\boldsymbol{\mu}_i^*)}{\phi(X;\boldsymbol{\mu}_i)}=&\exp\left(\frac{\|X-\boldsymbol{\mu}_i\|^2}{2}-\frac{\|X-\boldsymbol{\mu}_i^*\|^2}{2}\right)\\ =&\exp\left(\left(X-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^*}{2}\right)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*)\right)\end{aligned} \tag{13}$$

Since by Cauchy-Schwarz we have $|(X-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^*}{2})^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*)| = |(X-\boldsymbol{\mu}_i^*+\frac{\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_i}{2})^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*)| \leq (r+a/2)a$, we have:

$$\exp\left(-(r+\frac{a}{2})a\right) \leq \frac{\phi(X;\boldsymbol{\mu}_i^*)}{\phi(X;\boldsymbol{\mu}_i)} \leq \exp\left((r+\frac{a}{2})a\right) \tag{14}$$

For such X, $\phi(X;\boldsymbol{\mu}_1) \le (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{(R_{\min}-r-a)^2}{2}\right)$, and we have

$$\begin{aligned}
I_1^{(p)} =& \int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,r)} \frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j \pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX \\
\le& \int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,r)} \frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\pi_i\phi(X;\boldsymbol{\mu}_i)\exp\left((r+\frac{a}{2})a\right)}{\sum_j \pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX \\
\le& \pi_1 \exp\left((r+\frac{a}{2})a\right)\int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,r)} \phi(X;\boldsymbol{\mu}_1)\|X-\boldsymbol{\mu}_1\|^p dX \\
\le& \pi_1(2\pi)^{-d/2}\exp\left((r+\frac{a}{2})a\right)(R_{\max}+a+r)^p \exp\left(-\frac{(R_{\min}-r-a)^2}{2}\right)\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}r^d \\
\le& \frac{\pi_1 2^{-d/2}}{\Gamma(\frac{d}{2}+1)}\exp\left((r+\frac{a}{2})a - \frac{(R_{\min}-r-a)^2}{2}\right)(R_{\max}+a+r)^p r^d \\
\le& \pi_1 2^{1-d}\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-r/R_{\min})^2\right)\right)(R_{\max}+a+r)^p r^d
\end{aligned}$$

The last inequality follows from the fact that $\Gamma\left(\frac{d}{2}+1\right) \ge ([\frac{d}{2}])! \ge 2^{\frac{d}{2}-1}$. On the other hand, for I_2, since $w_1(X;\boldsymbol{\mu}) \le 1$, taking spherical coordinate transformation we have,

$$\begin{aligned}
I_2^{(p)} \le& \int_{\|X-\boldsymbol{\mu}_i^*\|\ge r} \pi_i\phi(X;\boldsymbol{\mu}_i^*)\|X-\boldsymbol{\mu}_1\|^p dX \\
\le& \pi_i \int_{\|X-\boldsymbol{\mu}_i^*\|\ge r} (2\pi)^{-d/2}\exp(-\frac{\|X-\boldsymbol{\mu}_i^*\|^2}{2})\|X-\boldsymbol{\mu}_1\|^p dX \\
\le& \frac{\pi_i 2^{1-d/2}}{\Gamma(\frac{d}{2})}\int_{u=r}^{\infty} u^{d-1}\exp\left(-\frac{u^2}{2}\right)(u+R_{\max}+a)^p du
\end{aligned}$$

Apply Lemma 7, when $r \ge 2\sqrt{d+2}$, for $p \in \{0,1,2\}$

$$I_2^{(p)} \le \pi_i\left(R_{\max}+a+d\right)^p \exp\left(-\frac{r}{2}\sqrt{d}\right) \tag{15}$$

Summing up I_1 and I_2, for any $0 < r < R_{\min}/2$, from (12) we get:

$$\begin{aligned}
&\pi_i \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)} w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p \\
\le& \pi_1 2^{1-d}\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-r/R_{\min})^2\right)\right)(R_{\max}+a+r)^p r^d + \pi_i\left(R_{\max}+a+d\right)^p\exp\left(-\frac{r}{2}\sqrt{d}\right)
\end{aligned} \tag{16}$$

Now plugging Eq. (11) and Eq. (16) into Eq. (9) gives,

$$
\begin{aligned}
&\mathbb{E}_X w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
\leq&(1-\pi_1)(r+a)^p \exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)+\pi_1\frac{(a+d)^p}{4}\exp\left(-\frac{r}{2}\sqrt{d}\right) \\
&+\pi_1(M-1)2^{1-d}\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-r/R_{\min})^2\right)\right)(R_{\max}+a+r)^p r^d \\
&+(1-\pi_1)\left(R_{\max}+a+d\right)^p\exp\left(-\frac{r}{2}\sqrt{d}\right) \\
\leq&\underbrace{(1-\pi_1)(r+a)^p\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)}_{(A)}+\underbrace{(R_{\max}+a+d)^p\exp\left(-\frac{r}{2}\sqrt{d}\right)}_{(B)} \\
&+\underbrace{2\pi_1(M-1)\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-r/R_{\min})^2\right)+d\log(r/2)\right)(R_{\max}+a+r)^p}_{(C)}
\end{aligned}
$$

Note that in order to have a negative term inside exponential of (A), we require $r+a<\frac{R_{\min}}{2}$. In order to ensure the same for (C), we need:

$$a<\frac{R_{\min}}{2}\left(1-\frac{r}{R_{\min}}\right)^2 \tag{17}$$

If $r^2\geq 2d\log(r/2)$, then we have:

$$
\begin{aligned}
&\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-r/R_{\min})^2\right)+d\log(r/2)\right)\leq\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-r/R_{\min})^2\right)+r^2/2\right) \\
&\leq\exp\left(R_{\min}a-\left(\frac{R_{\min}^2}{2}-rR_{\min}+\frac{r^2}{2}\right)+\frac{r^2}{2}\right) \\
&=\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)
\end{aligned}
$$

Therefore, $(A)+(C)\leq(1-\pi_1+2\pi_1(M-1))(R_{\max}+a+r)^p\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)$
Finally, if $r\leq R_{\min}\frac{R_{\min}/2-a}{R_{\min}+\sqrt{d}/2}$, we have:

$$\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2r-2a)\right)\leq\exp(-\frac{r}{2}\sqrt{d})$$

Hence,

$$
\begin{aligned}
(A)+(B)+(C)\leq&(2-\pi_1+2\pi_1(M-1))\left(\frac{3}{2}R_{\max}+d\right)^p\exp\left(-\frac{r}{2}\sqrt{d}\right) \\
\leq&2M\left(\frac{3}{2}R_{\max}+d\right)^p\exp\left(-\frac{r}{2}\sqrt{d}\right)
\end{aligned}
$$

Set

$$r=\frac{R_{\min}/2-a}{4},\quad a\leq\frac{R_{\min}}{2} \tag{18}$$

then Eq (17) and $a+r\leq\frac{R_{\min}}{2}$ are automatically satisfied. When $R_{\min}\geq\frac{\sqrt{d}}{6}$, we have $r\leq R_{\min}\frac{R_{\min}/2-a}{R_{\min}+\sqrt{d}/2}$. Finally in order to meet the constraints

$$r\geq 2\sqrt{d+2}\Leftarrow r\geq 3\sqrt{d} \tag{19}$$

$$r^2\geq 2d\log r/2 \tag{20}$$

we need

$$\begin{aligned} \frac{R_{\min}/2 - a}{4} &\geq \max(\sqrt{2d[\log(R_{\min}/4)]_+}, 2\sqrt{3}\sqrt{d}) \\ a &\leq \frac{R_{\min}}{2} - \sqrt{d}\max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3}) \end{aligned}$$

The right hand side of last inequality is non-negative when $R_{\min} = \tilde{\Omega}(\sqrt{d})$. Under these conditions, with Eq. (18) plugged in, we have

$$\mathbb{E}_X w_1(X;\boldsymbol{\mu})(1 - w_1(X;\boldsymbol{\mu}))\|X - \boldsymbol{\mu}_1\|^p \leq 2M\left(\frac{3}{2}R_{\max} + d\right)^p \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{d}/8\right)$$

□

Proof of Lemma 13. For any $r \leq \frac{R_{\min}}{2} - a$, define $\mathcal{E}_0 = \{X : \exists i, \text{ such that } Z_X = i, \|X - \boldsymbol{\mu}_i^*\| > r\}$ and $\mathcal{E}_k = \{X : Z_X = k, \|X - \boldsymbol{\mu}_k^*\| \leq r\}$.

$$\begin{aligned} &\mathbb{E}_X\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\| \cdot \|X - \boldsymbol{\mu}_j\|\right] \\ \leq &\underbrace{\mathbb{E}_X\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\|\|X - \boldsymbol{\mu}_j\||\mathcal{E}_0\right]P(\mathcal{E}_0)}_{I_0} \\ &+ \sum_{k\in[M]} \underbrace{\pi_k \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_k^*)}\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\|\|X - \boldsymbol{\mu}_j\|\|\|X - \boldsymbol{\mu}_k\| \leq r\right]}_{I_k} \end{aligned}$$

First we look at I_0, this again can be decomposed as the sum over mixtures. Similarly as in Eq. (15), we have

$$I_0 \leq (R_{\max} + a + d)^2 \exp\left(-\frac{r}{2}\sqrt{d}\right)$$

For I_k, by Eq. (14),

$$\begin{aligned} I_k = &\int_X \frac{\pi_i\phi(X;\boldsymbol{\mu}_i)\pi_j\phi(X;\boldsymbol{\mu}_j)\pi_k\phi(X;\boldsymbol{\mu}_k^*)}{(\sum_\ell \pi_\ell\phi(X;\boldsymbol{\mu}_t))^2}\|X - \boldsymbol{\mu}_i\| \cdot \|X - \boldsymbol{\mu}_j\|dX \\ \leq &\int_X \frac{\pi_i\phi(X;\boldsymbol{\mu}_i)\pi_j\phi(X;\boldsymbol{\mu}_j)\pi_k\phi(X;\boldsymbol{\mu}_k)\exp((r + a/2)a)}{(\sum_\ell \pi_\ell\phi(X;\boldsymbol{\mu}_\ell))^2}\|X - \boldsymbol{\mu}_i\| \cdot \|X - \boldsymbol{\mu}_j\|dX \\ \leq &\kappa\pi_k 2\pi^{-\frac{d}{2}}\exp(-\frac{R_{(\min} - r - a)^2}{2})\exp((r + a/2)a)(R_{\max} + r + a)^2\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2} + 1\right)}r^d \\ \leq &\pi_k\kappa 2^{-d/2}\frac{1}{\Gamma\left(\frac{d}{2} + 1\right)}r^d\exp\left((r + a/2)a - \frac{(R_{\min} - r - a)^2}{2}\right)(R_{\max} + r + a)^2 \\ \leq &2\pi_k\kappa\exp\left(R_{\min}\left(a - \frac{R_{\min}}{2}\left(1 - \frac{r}{R_{\min}}\right)^2\right) + d\log(r/2)\right)(R_{\max} + r + a)^2 \end{aligned} \tag{21}$$

Adding up I_k's and I_0, we have

$$\begin{aligned} &\mathbb{E}_X\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\|\|X - \boldsymbol{\mu}_j\|\right] \\ \leq &(R_{\max} + a + d)^2 \exp\left(-\frac{r}{2}\sqrt{d}\right) \\ &\quad + 2\kappa\exp\left(R_{\min}\left(a - \frac{R_{\min}}{2}\left(1 - \frac{r}{R_{\min}}\right)^2\right) + d\log(r/2)\right)(R_{\max} + r + a)^2 \end{aligned}$$

Take $r = \frac{1}{4}\left(\frac{R_{\min}}{2} - a\right)$, we have $R_{\min}\left(a - \frac{R_{\min}}{2}\left(1 - \frac{r}{R_{\min}}\right)^2\right) + d\log(r/2) \leq -\frac{r}{2}\sqrt{d}$. Therefore,

$$
\begin{aligned}
&\mathbb{E}_X[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\| \cdot \|X - \boldsymbol{\mu}_j\|] \\
\leq &(1+2\kappa)\left(\frac{3}{2}R_{\max} + d\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{d}/8\right)
\end{aligned}
$$

□

Proof of Theorem 4. Consider the difference of the gradient corresponding to $\boldsymbol{\mu}_i$, without loss of generality, assume $i = 1$.

$$
\nabla_{\boldsymbol{\mu}_1} Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu}^t) = \mathbb{E}(w_1(X;\boldsymbol{\mu}^t) - w_1(X;\boldsymbol{\mu}^*))(X - \boldsymbol{\mu}_1^t) \tag{22}
$$

For any given X, consider the function $\boldsymbol{\mu} \to w_1(X;\boldsymbol{\mu})$, we have

$$
\nabla_{\boldsymbol{\mu}} w_1(X;\boldsymbol{\mu}) = \begin{pmatrix} w_1(X;\boldsymbol{\mu})(1 - w_1(X;\boldsymbol{\mu}))(X - \boldsymbol{\mu}_1)^T \\ -w_1(X;\boldsymbol{\mu})w_2(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_2)^T \\ \vdots \\ -w_1(X;\boldsymbol{\mu})w_M(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_M)^T \end{pmatrix} \tag{23}
$$

Let $\boldsymbol{\mu}^u = \boldsymbol{\mu}^* + u(\boldsymbol{\mu}^t - \boldsymbol{\mu}^*), \forall u \in [0,1]$, obviously $\boldsymbol{\mu}^u \in \otimes_{i=1}^M \mathbb{B}(\boldsymbol{\mu}_i^*, \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|) \subset \otimes_{i=1}^M \mathbb{B}(\boldsymbol{\mu}_i^*, a)$. By Taylor's theorem,

$$
\begin{aligned}
\|\mathbb{E}(w_1(X;\boldsymbol{\mu}_1^t) - w_1(X;\boldsymbol{\mu}_1^*))(X - \boldsymbol{\mu}_1^t)\| &= \left\|\mathbb{E}\left[\int_{u=0}^1 \nabla_u w_1(X;\boldsymbol{\mu}^u)du(X - \boldsymbol{\mu}_1^t)\right]\right\| \\
= \Bigg\| \int_{u=0}^1 &\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^u)^T(\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*)(X - \boldsymbol{\mu}_1^t)du \\
&- \sum_{i\neq 1}\int_{u=0}^1 \mathbb{E}w_1(X;\boldsymbol{\mu}^u)w_i(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_2^u)^T(\boldsymbol{\mu}_2^t - \boldsymbol{\mu}_2^*)(X - \boldsymbol{\mu}_1^t)du\Bigg\| \\
\leq U_1\|\boldsymbol{\mu}_1^t &- \boldsymbol{\mu}_1^*\|_2 + \sum_{i\neq 1} U_i \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2
\end{aligned} \tag{24}
$$

where

$$
\begin{aligned}
U_1 &= \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\
U_i &= \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)w_i(X;\boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_2^u)^T\|_{op}
\end{aligned}
$$

For U_1 by triangle inequality we have,

$$
\begin{aligned}
U_1 \leq &\sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^u)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\
&+ \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(\boldsymbol{\mu}_1^u - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\
\leq &\sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^u)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\
&+ a \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^u)\|
\end{aligned} \tag{25}
$$

We now develop an uniform bound for the operator norm. For any $u \in [0,1]$, there exists a rotation matrix O, such that all $R\boldsymbol{\mu}_i^u, i \in [M]$ have non-zero entries in the leading $\min\{d, M\}$ coordinates, and zeros for the remaining $[d-M]_+$ coordinates. Denote $\tilde{X} := OX$, then $\tilde{X}|Z = i \sim \mathcal{N}(O\boldsymbol{\mu}_i^*, I_d)$. Let

$$O\boldsymbol{\mu}_i^u = [\tilde{\boldsymbol{\mu}}_i^u, 0_{[d-M]_+}] \text{ and } O\boldsymbol{\mu}_i^* = [v_i^{\min\{d,M\}}, v_i^{[d-M]_+}], \quad \tilde{\boldsymbol{\mu}}_i^u \in \mathbb{R}^{\min\{d,M\}}$$

For ease of notation, we assume $d \geq M$ for now, the other case can be derived without much modification. We can rewrite

$$(X - \boldsymbol{\mu}_1^u)(X - \boldsymbol{\mu}_1^u)^T = O^T \begin{bmatrix} (\tilde{X}^M - \tilde{\boldsymbol{\mu}_1^u})(\tilde{X}^M - \tilde{\boldsymbol{\mu}_1^u})^T & (\tilde{X}^M - \tilde{\boldsymbol{\mu}_1^u})(\tilde{X}^{d-M})^T \\ (\tilde{X}^{d-M})(\tilde{X}^M - \tilde{\boldsymbol{\mu}_1^u})^T & (\tilde{X}^{d-M})(\tilde{X}^{d-M})^T \end{bmatrix} O$$

Note by the rotation, $w_i(X; \boldsymbol{\mu})$ only depend on the first M coordinates. And by isotropicity, $\tilde{X}^M$ and $\tilde{X}^{d-M}$ are independent. By $\mathbb{E}\tilde{X}^{d-M} = 0$ (since we assume that the centroid of the means is at zero, and a rotation does not change that) and $\mathbb{E}\tilde{X}^{d-M}(\tilde{X}^{d-M})^T = I_{d-M} + \sum_i \pi_i (v_i^{d-M})(v_i^{d-M})^T$, we have,

$$\begin{aligned}
&\|\mathbb{E}w_1(X; \boldsymbol{\mu}^u)(1 - w_1(X; \boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^u)(X - \boldsymbol{\mu}_1^u)^T\|_{op} = \left\| \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \right\|_{op} \\
&\leq \max\{\|D_1\|_{op}, \|D_2\|_{op}\}
\end{aligned}$$

D_1 and D_2 are defined below. Applying Lemma 12 with dimension $\min\{d, M\}$, when $R_{\min} = \Omega(\sqrt{\min\{d, M\}})$,

$$\begin{aligned}
\|D_1\|_{op} =& \|\mathbb{E}w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u)(1 - w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u))(\tilde{X}^{\min\{d,M\}} - \tilde{\boldsymbol{\mu}_1^u})(\tilde{X}^{\min\{d,M\}} - \tilde{\boldsymbol{\mu}_1^u})^T\|_{op} \\
\leq& 2M \left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)
\end{aligned}$$

For D_2, by independence and Lemma 12, when $R_{\min} = \Omega(\sqrt{\min\{d, M\}})$,

$$\begin{aligned}
\|D_2\|_{op} =& \left\|\mathbb{E}w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u)(1 - w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u)) \left(I_{[d-M]_+} + \sum_i \pi_i (v_i^{[d-M]_+})(v_i^{[d-M]_+})^T\right)\right\|_{op} \\
=& \left\| \left(\mathbb{E}_{\tilde{X}_{\min\{d,M\}}} w_1(\tilde{X}_{\min\{d,M\}}; \tilde{\boldsymbol{\mu}}^u)(1 - w_1(\tilde{X}_{\min\{d,M\}}; \tilde{\boldsymbol{\mu}}^u))\right) \right. \\
& \left. \cdot \mathbb{E}_{X_{[d-M]_+}} \left(I_{[d-M]_+} + \sum_i \pi_i (v_i^{[d-M]_+})(v_i^{[d-M]_+})^T\right) \right\|_{op} \\
\leq& (R_{\max}^2 + 1) 2M \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)
\end{aligned}$$

Combining the two and plugging in Eq. (25),

$$\begin{aligned}
U_1 \leq& 2M \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right) \cdot \\
& \left(\max\left\{\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)^2, (R_{\max}^2 + 1)\right\} + a\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)\right) \\
\leq& 2M \left(2R_{\max} + \min\{d, M\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)
\end{aligned}$$

The max will always be achieved at the first term as $\min\{d, M\} \geq 1$. Similarly, with the same rotation, for $U_i, i \neq 1$,

$$U_i \leq \sup_u \|\mathbb{E}w_1(X; \boldsymbol{\mu}^u)w_i(X; \boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_1^u)(X - \boldsymbol{\mu}_i^u)^T\|_{op} + a\|\mathbb{E}w_1(X; \boldsymbol{\mu}^u)w_i(X; \boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_i^u)\|$$

By Lemma 13, when $R_{\min} = \Omega(\sqrt{\min\{d, M\}})$, we have

$$\begin{aligned}
U_i \leq & \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right) \cdot \\
& \left(\max\left\{(1 + 2\kappa)\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)^2, 2M(R_{\max}^2 + 1)\right\} + 2Ma\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)\right) \\
\leq & \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right) \\
& \cdot \left(\max\{(1 + 2\kappa), 2M\}\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right) + 2Ma\right) \\
\leq & \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)^2 \cdot \max\{3M, M + 2\kappa + 1\} \\
\leq & M(2\kappa + 4)\left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right)
\end{aligned}$$

The second inequality is because $R_{\max}^2 + 1 \leq \left(\frac{3}{2}R_{\max} + \min\{d, M\}\right)^2$ and the third inequality is because $2a \leq \frac{3}{2}R_{\max} + \min\{d, M\}$. Taking back to Eq. (24), and summing over $i \in [M]$, we have

$$\begin{aligned}
& \|\nabla_{\boldsymbol{\mu}_i} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla_{\boldsymbol{\mu}_i} q(\boldsymbol{\mu})\| \\
\leq & M(2\kappa + 4)\left(2R_{\max} + \min\{d, M\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\min\{d, M\}}/8\right) \sum_{i=1}^M \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|
\end{aligned}$$

This completes the proof. □

B.2 Proof of Theorem 1

Proof of Theorem 1. By Theorem 4 and Theorem 3, it suffices to check $\gamma \leq \pi_{\min}$. Solving the inequality we have

$$a \leq \frac{R_{\min}}{2} - \frac{2\sqrt{2}}{\sqrt[4]{\min\{d, M\}}}\sqrt{\log\left(\frac{M^2(2\kappa + 4)(2R_{\max} + \min\{d, M\})^2}{\pi_{\min}}\right)}$$

Combined with the condition in Theorem 4, we have

$$\begin{aligned}
a \leq & \frac{R_{\min}}{2} - \max\left\{\frac{2\sqrt{2}}{\sqrt[4]{\min\{d, M\}}}\sqrt{\log\left(\frac{M^2(2\kappa + 4)(2R_{\max} + \min\{d, M\})^2}{\pi_{\min}}\right)},\right. \\
& \left.\sqrt{\min\{d, M\}}\max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3})\right\} \\
= & \frac{R_{\min}}{2} - \sqrt{\min\{d, M\}}o(R_{\min})
\end{aligned}$$

because

$$\begin{aligned}
&\max\{c\sqrt{\log(c_1\frac{M^2\kappa}{\pi_{\min}} + 2\log\left(2R_{\max} + \min\{d, M\}\right)},\\
&\qquad\sqrt{\min\{d, M\}}\max\{c_2\sqrt{\log(R_{\min}/4)_+}, 8\sqrt{3}\}\}\\
&\leq \max\{c\sqrt{\log(c_1\frac{M^2\kappa}{\pi_{\min}} + c_2 R_{\max} + c_3\min\{d, M\})},\\
&\qquad c'\sqrt{\min\{d, M\}}\sqrt{\log(R_{\max} + e)}\}\\
&\leq \sqrt{\min\{d, M\}}O\left(\sqrt{\log\left(\max\left\{\frac{M^2\kappa}{\pi_{\min}}, R_{\max}, \min\{d, M\}\right\}\right)}\right)
\end{aligned}$$

The condition in Theorem 4 can be rewritten as

$$a \leq \frac{R_{\min}}{2} - \sqrt{\min\{d, M\}}O\left(\sqrt{\log\left(\max\left\{\frac{M^2\kappa}{\pi_{\min}}, R_{\max}, \min\{d, M\}\right\}\right)}\right)$$

□

C Proofs for sample-based gradient EM

In this section we develop the error bound for sample-based gradient EM. Our proof is based on the Rademacher complexity theory and some new tools for contraction result.

Proof of Proposition 1. For any unit vector u, the Rademacher complexity of $\mathcal{F}$ is

$$\begin{aligned}
R_n(\mathcal{F}) =& \mathbb{E}_X\mathbb{E}_\epsilon \sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \epsilon_i w_1(X_i;\boldsymbol{\mu})\langle X_i - \boldsymbol{\mu}_1, u\rangle\\
\leq& \underbrace{\mathbb{E}_X\mathbb{E}_\epsilon \sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \epsilon_i w_1(X_i;\boldsymbol{\mu})\langle X_i, u\rangle}_{(D)} + \underbrace{\mathbb{E}_X\mathbb{E}_\epsilon \sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \epsilon_i w_1(X_i;\boldsymbol{\mu})\langle \boldsymbol{\mu}_1, u\rangle}_{(E)}
\end{aligned} \tag{26}$$

We bound the two terms separately. Define $\eta_j(\boldsymbol{\mu}) : \mathbb{R}^{Md} \to \mathbb{R}^M$ to be a vector valued function with the k-th coordinate

$$[\eta_j(\boldsymbol{\mu})]_k = \frac{\|\boldsymbol{\mu}_1\|^2}{2} - \frac{\|\boldsymbol{\mu}_k\|^2}{2} + \langle X_j, \boldsymbol{\mu}_k - \boldsymbol{\mu}_1\rangle + \log\left(\frac{\pi_k}{\pi_1}\right)$$

We claim

$$|w_1(X_j;\boldsymbol{\mu}) - w_1(X_j;\boldsymbol{\mu}')| \leq \frac{\sqrt{M}}{4}\left\|\eta_j(\boldsymbol{\mu}) - \eta_j(\boldsymbol{\mu}')\right\| \tag{27}$$

This vectorized Lipschitz condition simply follows from the fact that

$$\begin{aligned}
w_1(X_j, \boldsymbol{\mu}) =& \frac{1}{1 + \sum_{k=2}^M \exp([\eta_j(\boldsymbol{\mu})]_k)}\\
\frac{\partial w_1(X_j, \boldsymbol{\mu})}{\partial[\eta_j(\boldsymbol{\mu})]_k} =& \frac{\exp([\eta_j(\boldsymbol{\mu})]_k)}{(1 + \sum_{k=2}^M \exp([\eta_j(\boldsymbol{\mu})]_k))^2} \leq \frac{1}{4}
\end{aligned}$$

so $w_1(X_j, \boldsymbol{\mu})$ is $\frac{1}{4}$-Lipschitz continuous w.r.t. $[\eta_j(\boldsymbol{\mu})]_k$. By Lemma 11, $w_1(X_j, \boldsymbol{\mu})$ is $\frac{\sqrt{M}}{4}$ Lipschitz w.r.t $\eta_j(\boldsymbol{\mu})$ Now let $\psi_j(\boldsymbol{\mu}) = w_1(X_j;\boldsymbol{\mu})\langle X_j, u\rangle$.

With Lipschitz property (27) and by Lemma 3, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j w_1(X_j;\boldsymbol{\mu})\langle X_j,u\rangle\right] \leq \mathbb{E}\left[\frac{1}{n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}[\eta_j(\boldsymbol{\mu})]_k\frac{\sqrt{2M}}{4}\langle X_j,u\rangle\right]\\
=&\mathbb{E}\left[\frac{\sqrt{2}M^{\frac{1}{2}}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=2}^{M}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle+\log(\frac{\pi_k}{\pi_1})\right)\langle X_j,u\rangle\right]\\
\leq&\underbrace{\mathbb{E}\left[\frac{\sqrt{2M}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\log(\frac{\pi_k}{\pi_1})\right)\langle X_j,u\rangle\right]}_{(D.1)}\\
&+\underbrace{\mathbb{E}\left[\frac{\sqrt{2M}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle\langle X_j,u\rangle\right]}_{(D.2)}
\end{aligned}
\tag{28}
$$

To bound (D.1), note that the sum over $k = 1, \cdots, M$ can be considered as an inner product of two vectors in $\mathbb{R}^M$. The supremum of $\|\boldsymbol{\mu}\|$ can be bounded as $\max_{\boldsymbol{\mu}\in\mathbb{A}}\|\boldsymbol{\mu}_i\| \leq \|\boldsymbol{\mu}^*_{\max}\| + a \leq \frac{3}{2}R_{\max}$.

$$
\begin{aligned}
(D.1)=&\mathbb{E}\left[\frac{\sqrt{2M}}{4}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\begin{pmatrix}\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_1\|^2}{2}+\log(\frac{\pi_1}{\pi_1})\\ \vdots\\ \frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_M\|^2}{2}+\log(\frac{\pi_M}{\pi_1})\end{pmatrix}^T\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\langle X_j,u\rangle\\ \vdots\\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jM}\langle X_j,u\rangle\end{pmatrix}\right]\\
\leq& cM(9R_{\max}^2/4+\log(\kappa))\mathbb{E}\left\|\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\langle X_j,u\rangle\\ \vdots\\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jM}\langle X_j,u\rangle\end{pmatrix}\right\|
\end{aligned}
\tag{29}
$$

By Lemma 9, and $\|u\| = 1$, we know $\langle X_j, u\rangle$ is sub-Gaussian with parameter upper bounded by $1 + R_{\max}$. So each element of the vector in Equation 29 is the average of n independent mean 0 sub-Gaussian random variables with sub-gaussian norm upper bounded by $1 + R_{\max}$ (since w.l.o.g we have assumed that $\sigma = 1$ and $\max_i \|\mu\| \leq R_{\max}$, by Lemma 10). Consequently, $\forall k \in [M]$, $\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}\langle X_j, u_1\rangle\right| \leq c(1+R_{\max})/\sqrt{n}$ for some global constant c [18], and

$$
(D.1) \leq cM^{3/2}(9R_{\max}^2/4+\log(\kappa))(1+R_{\max})\frac{1}{\sqrt{n}} \leq cM^{3/2}(1+R_{\max})^3\max\{1,\log(\kappa)\}\frac{1}{\sqrt{n}}
$$

On the other hand, for $(D.2)$, we have

$$
\begin{aligned}
(D.2)=&\mathbb{E}\left[\frac{\sqrt{2M}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle\langle X_j,u\rangle\right]\\
=&\mathbb{E}\left[\frac{\sqrt{2M}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{k=1}^{M}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^T\left(\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right)u\right]\\
\leq&\sum_{k=1}^{M}\mathbb{E}\left[\frac{\sqrt{2M}}{4}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\|\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right\|_{op}\right]\\
\leq&\sum_{k=1}^{M}\frac{\sqrt{2M}}{2}\|\boldsymbol{\mu}_{\max}\|\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right\|_{op}\right]
\end{aligned}
\tag{30}
$$

For each $k \in [M]$, the operator norm $\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\|_{op}$ can be bounded by the same discretization technique with the 1/2-covering of the unit sphere. To be specific, since for any matrix A, $\|A\|_{op} = \sup_{u \in \mathcal{S}^{d-1}} \|Au\|$,

$$\forall u, \exists u_j \ s.t. \ \|Au\| \leq \|Au_j\| + \|A\|_{op}\|u - u_j\| \leq \max_j \|Au_j\| + \frac{1}{2}\|A\|_{op}$$

Taking $\sup_{u \in \mathcal{S}^{d-1}}$ on the left side, we get $\|A\|_{op} \leq 2\max_j \|Au_j\|$. Therefore $\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\|_{op} \leq 2\max_\ell \frac{1}{n}\sum_{j=1}^n \epsilon_{jk}\langle X_j, u_\ell\rangle^2$. The square of sub-gaussian random variable $\langle X_j, u_\ell\rangle$ is sub-exponential, from Lemma 5.14 in [18] we know

$$\mathbb{E}\left[\exp\left(\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}\langle X_j, u\rangle^2 t\right)\right] \leq \exp\left(\frac{c_4 t^2 (1 + R_{\max})^4}{n}\right)$$

With the 1/2-covering number of $\mathcal{S}^{d-1}$ bounded by $\exp(2d)$, we have

$$\mathbb{E}\left[\exp\left(t \cdot \|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\|_{op}\right)\right] \leq \exp\left(2d + \frac{c_5 t^2 (1 + R_{\max})^4}{n}\right)$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\right\|_{op}\right] &= \frac{1}{t}\log\left(\exp\left(t\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\right\|_{op}\right]\right)\right), \quad \forall t > 0 \\
&\leq \frac{1}{t}\log\left(\mathbb{E}\left[\exp\left(t\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\right\|_{op}\right)\right]\right) \\
&\leq \frac{2d}{t} + \frac{ct(1 + R_{\max})^4}{n}
\end{aligned}$$

Taking $t = c\frac{\sqrt{nd}}{(1+R_{\max})^2}$,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk} X_j X_j^T\right\|_{op}\right] \leq c\sqrt{\frac{d}{n}}(1 + R_{\max})^2$$

Plugging back to Eq. (30), and use $\sup_{\boldsymbol{\mu} \in \mathbb{A}} \|\boldsymbol{\mu}\| \leq \sup_k \|\boldsymbol{\mu}_k^*\| + a \leq \frac{3}{2}R_{\max}$, we have

$$(D.2) \leq \frac{cM(1 + R_{\max})^3\sqrt{d}}{\sqrt{n}}$$

Plugging the bound back to Eq. (28), we have

$$(D) \leq \frac{cM^{3/2}(1 + R_{\max})^3\sqrt{d}\max\{1, \log(\kappa)\}}{\sqrt{n}}$$

Apply Lemma 3 on the (E) term in Eq. (26), we have

$$\begin{aligned}(E) =&\mathbb{E}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j w_i(X_j;\boldsymbol{\mu})\langle\boldsymbol{\mu}_i,u\rangle\right]\\ \leq&\mathbb{E}\left[\frac{\sqrt{2M}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle+\log(\frac{\pi_k}{\pi_1})\right)\langle\boldsymbol{\mu}_i,u\rangle\right]\\ \leq&\underbrace{\frac{\sqrt{2M}}{4}\mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\log\frac{\pi_k}{\pi_1}\right)\langle\boldsymbol{\mu}_i,u\rangle\right]}_{E.1}\\ &\qquad+\underbrace{\frac{\sqrt{2M}}{4}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle\langle\boldsymbol{\mu}_i,u\rangle\right]}_{E.2}\end{aligned}$$

We will now bound $(E.1)$ and $(E.2)$.

$$\begin{aligned}(E.1)\leq&\frac{\sqrt{2M}}{4}\mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\log\frac{\pi_k}{\pi_1}\right)\sup_{\boldsymbol{\mu}\in\mathbb{A}}\langle\boldsymbol{\mu}_i,u\rangle\right]\\ \leq&\frac{\sqrt{2M}}{4}R_{\max}\mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\begin{pmatrix}\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_1\|^2}{2}+\log(\frac{\pi_1}{\pi_1})\\ \vdots\\ \frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_M\|^2}{2}+\log(\frac{\pi_M}{\pi_1})\end{pmatrix}^T\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\\ \vdots\\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jM}\end{pmatrix}\right]\\ \leq&cMR_{\max}(9R_{\max}^2/4+\log\kappa)E_\epsilon\left\|\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\\ \vdots\\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jM}\end{pmatrix}\right\|\end{aligned}\tag{31}$$

Note that each element of the vector in Equation 31 is the average of n i.i.d mean 0 Radamacher random variables, which are essentially sub-gaussian radnom variables with subgaussian norm upper bounded by 1. Consequently, $\forall k\in[M]$, $\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}\right|\leq c'/\sqrt{n}$ for some global constant c [18], and

$$(E.1)\leq c'M^{3/2}R_{\max}(9R_{\max}^2/4+\log\kappa)/\sqrt{n}$$

As for (E.2), we have

$$\begin{aligned}(E.2)\leq&\frac{\sqrt{2M}}{4}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{M}\epsilon_{jk}\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle\sup_{\boldsymbol{\mu}\in\mathbb{A}}\langle\boldsymbol{\mu}_i,u\rangle\right]\\ \leq&\frac{3\sqrt{2M}}{8}R_{\max}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{k=1}^{M}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^T\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right)\right]\\ \leq&\frac{3\sqrt{2M}}{8}R_{\max}\sum_{k=1}^{M}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^T\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right)\right]\\ \leq&\frac{9\sqrt{2M}}{8}R_{\max}^2\sum_{k=1}^{M}\mathbb{E}_{X,\epsilon}\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right\|\end{aligned}\tag{32}$$

In Eq (32), the vector $\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_j$ is the average of n independent mean zero isotropic subgaussian random vectors. Another using of the discretizing technique along with the moment generating function with $t \geq 0$ gives:

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_j\right\| &\leq 2\max_{\ell}\langle\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_j, u_\ell\rangle \\
E\left[\exp t\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_j\right\|\right] &\leq \sum_{\ell} E\left[\exp\left(2\frac{t}{n}\sum_{j=1}^n \epsilon_{jk}\langle X_j, u_\ell\rangle\right)\right] \leq \exp\left(2d + \frac{c'(1+R_{\max})^2t^2}{n}\right) \\
E\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_j\right\| &\leq \frac{c'' + 2d + \frac{c'(1+R_{\max})^2t^2}{n}}{t} \qquad \text{Using Jensen's inequality}
\end{aligned}$$

Taking $t = \Theta\sqrt{nd}/(1+R_{\max})$,

$$(E.2) \leq cM^{3/2}R_{\max}^2(1+R_{\max})\sqrt{d}/\sqrt{n}$$

Thus, combing (E.1) and (E.2) we get:

$$(E) \leq \frac{cM^{3/2}(1+R_{\max})^3\max\{1,\log(\kappa)\}\sqrt{d}}{\sqrt{n}}$$

The final bound follows by combining (D) and (E):

$$R_n(\mathcal{F}) \leq \frac{cM^{3/2}(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}}{\sqrt{n}}$$

□

Proof of Lemma 2. Consider for some $r > 0$, the set in which X lies in the r-ball of its corresponding center. If Z_i denotes the hidden cluster assignment of X_i, we denote $\mathcal{Y}_r^i = \{X_1, \cdots, X_n : \|X_i - \boldsymbol{\mu}_{Z_i}^*\| \leq r\}$.

$$\mathcal{Y}_r := \{X_1, \ldots X_n : \|X_i - \boldsymbol{\mu}_{Z_i}^*\| \leq r, \ , \forall i \in [n]\} = \cap_i \mathcal{Y}_r^i$$

By Lemma 8 and union bound, for $r = \Omega(\sqrt{d})$,

$$p := P(\mathbf{X} \notin \mathcal{Y}_r) \leq \sum_{i=1}^n P(X \in (\mathcal{Y}_r^i)^c) \leq cn\exp\left(-\frac{r\sqrt{d}}{2}\right). \tag{33}$$

Let $m_r := \mathbb{E}[g(X)|X \in \mathcal{Y}_r]$, we want to show m_r is close to $\mathbb{E}[g(\mathbf{X})]$ and is close to $g(\mathbf{X})$ with high probability.
Let $\mathbf{X}$ and $\mathbf{X'}$ be two samples which only differ on one data-point, then

$$\begin{aligned}
g(\mathbf{X}) - g(\mathbf{X}') = &\sup_{\boldsymbol{\mu}\in\mathbb{A}}\left(\frac{1}{n}\sum_{i=1}^n w_1(X_i;\boldsymbol{\mu})\langle X_i - \boldsymbol{\mu}_1, u\rangle - \mathbb{E}_X w_1(X;\boldsymbol{\mu})\langle X - \boldsymbol{\mu}_1, u\rangle\right) \\
&- \sup_{\boldsymbol{\mu}\in\mathbb{A}}\left(\frac{1}{n}\sum_{i=1}^n w_1(X_i';\boldsymbol{\mu})\langle X_i' - \boldsymbol{\mu}_1, u\rangle - \mathbb{E}_X w_1(X';\boldsymbol{\mu})\langle X' - \boldsymbol{\mu}_1, u\rangle\right)
\end{aligned}$$

Assume $\tilde{\boldsymbol{\mu}}$ be the maximizer for the supremum of X, then

$$\begin{aligned}
g(\mathbf{X}) - g(\mathbf{X}') \overset{(i)}{\leq} & \frac{1}{n}(\sum_{i=1}^{n} w_1(X_i; \tilde{\boldsymbol{\mu}})\langle X_i - \tilde{\boldsymbol{\mu}}_1, u\rangle - \mathbb{E}w_1(X; \tilde{\boldsymbol{\mu}})\langle X - \tilde{\boldsymbol{\mu}}_1, u\rangle) \\
& - \frac{1}{n}\sum_{i=1}^{n}(w_1(X_i'; \tilde{\boldsymbol{\mu}})\langle X_i' - \tilde{\boldsymbol{\mu}}_1, u\rangle - \mathbb{E}w_1(X'; \tilde{\boldsymbol{\mu}})\langle X' - \tilde{\boldsymbol{\mu}}_1, u\rangle) \\
= & \frac{1}{n} w_1(X_i; \tilde{\boldsymbol{\mu}})\langle X_i - \tilde{\boldsymbol{\mu}}_1, u\rangle - w_1(X_i'; \tilde{\boldsymbol{\mu}})\langle X_i' - \tilde{\boldsymbol{\mu}}_1, u\rangle
\end{aligned}$$

where (i) is by definition of supremum. The inequality holds when we change the order of X and X', hence for $X, X' \in \mathcal{Y}_r$,

$$\begin{aligned}
|g(\mathbf{X}) - g(\mathbf{X}')| \leq & \frac{1}{n}|w_1(X_i; \tilde{\boldsymbol{\mu}})\langle X_i - \tilde{\boldsymbol{\mu}}_1, u\rangle - w_1(X_i'; \tilde{\boldsymbol{\mu}})\langle X_i' - \tilde{\boldsymbol{\mu}}_1, u\rangle| \\
\leq & \frac{2}{n} \sup_{\boldsymbol{\mu} \in \mathbb{A}, X \in \mathcal{Y}_r} |w_1(X_i; \boldsymbol{\mu})\langle X_i - \boldsymbol{\mu}_1, u\rangle| \\
\leq & \frac{2}{n} \sup_{X \in \mathcal{Y}_r} (\|X - \boldsymbol{\mu}_{Z_X}^*\| + R_{\max}) \\
\leq & \frac{2(r + R_{\max})}{n} := L
\end{aligned}$$

By Theorem 6, we have

$$\begin{aligned}
P(g(\mathbf{X}) - m_r \geq \epsilon) \leq & p + \exp\left(-2\frac{(\epsilon - nLp)_+^2}{nL^2}\right) \\
\leq & c_1 n \exp(-c\sqrt{d}r) + \exp\left(-2\frac{(\epsilon - nL \cdot c_1 n \exp(-c\sqrt{d}r))_+^2}{nL^2}\right) \\
= & \underbrace{cn \exp(-c\sqrt{d}r)}_{P_1} + \underbrace{\exp\left(-\frac{c_1 n(\epsilon - c_2 n(r + R_{\max})\exp(-c\sqrt{d}r))_+^2}{(r + R_{\max})^2}\right)}_{P_2} \qquad (34)
\end{aligned}$$

Let $r = \Theta((1 + R_{\max})\log^2(n)\sqrt{d})$ and $\epsilon = c_0(1 + R_{\max})d\log^{5/2}(n)/\sqrt{n}$. Since, for large n,

$$n(r + R_{\max})\exp(-cr\sqrt{d}) \leq c_2(1 + R_{\max})\exp(\log n + \log\log n - c\log^2 n) = o((1 + R_{\max})/\sqrt{n})$$

for some constant c_0, which yields for large n, $(\epsilon - c_2 n(r + R_{\max})\exp(-cr\sqrt{d}))_+ \geq \epsilon/2$. Finally, for large n, we can have the following bounds on P_1 and P_2.

$$\begin{aligned}
P_1 &= O\left(\exp(\log n - c(1 + R_{\max})^2(\log n)^2)\right) = O\left(\exp(-c'(1 + R_{\max})^2 d\log n)\right) \\
P_2 &\leq \exp\left(-\frac{cn\epsilon^2}{d(\log^2 n(1 + R_{\max}))^2}\right) = O\left(\exp(-c'' d\log n)\right) \qquad (35)
\end{aligned}$$

where c, c', c''' are some global constants. The last line uses the fact $r + R_{\max} = O(\sqrt{d}(1 + R_{\max})\log^2 n)$. Now we bound the difference between $\mathbb{E}g(X)$ and the conditional expectation m_r. By the total expectation theorem,

$$\begin{aligned}
& \mathbb{E}g(\mathbf{X}) = m_r P(\mathbf{X} \in \mathcal{Y}_r) + \mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)] \\
& \mathbb{E}[g(\mathbf{X})](P(\mathbf{X} \in \mathcal{Y}_r) + P(\mathbf{X} \notin \mathcal{Y}_r)) = m_r P(\mathbf{X} \in \mathcal{Y}_r) + \mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)] \\
& \mathbb{E}g(\mathbf{X}) - m_r = \frac{\mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)] - \mathbb{E}[g(\mathbf{X})]P(X \notin \mathcal{Y}_r)}{P(\mathbf{X} \in \mathcal{Y}_r)} \qquad (36) \\
\Rightarrow & |m_r - \mathbb{E}g(\mathbf{X})| \leq \frac{p|\mathbb{E}g(\mathbf{X})| + |\mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)]|}{1 - p}
\end{aligned}$$

p is defined in Eq (33). Note that by Proposition 1, and the symmetrization result, $\mathbb{E}g(X) \leq 2R_n(\mathcal{F}) \leq cn^{-1/2}M^3\sqrt{d}(1+R_{\max})^3 \max\{1, \log(\kappa)\}$. On the other hand, as $g(\mathbf{X})$ is the sup over a class of quantity, which is centered at zero. So $g(X) \geq 0$. We also have $1(\mathbf{X} \in \cup_i (\mathcal{Y}_r^i)^c) \leq \sum_{i=1}^n 1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)$. Hence,

$$\mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)] = \mathbb{E}[g(\mathbf{X})\sum_{i=1}^n 1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)] \leq \sum_{i=1}^n \mathbb{E}[g(\mathbf{X})1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)]$$

Note for each sample X_i and $\boldsymbol{\mu}$, $\left|\sup_{\boldsymbol{\mu}\in\mathbb{A}} w_1(X_i;\boldsymbol{\mu})\langle X_i - \boldsymbol{\mu}_1, u\rangle\right| \leq \sup_{\boldsymbol{\mu}\in\mathbb{A}} w_1(X_i;\boldsymbol{\mu})\|X_i - \boldsymbol{\mu}^*_{Z_i}\| + \|\boldsymbol{\mu}^*_{Z_i} - \boldsymbol{\mu}_1\| \leq \|X_i - \boldsymbol{\mu}^*_{Z_i}\| + 2R_{\max}$. Thus,

$$\begin{aligned}
|g(\mathbf{X})| =& |\sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{j=1}^n w_1(X_j;\boldsymbol{\mu})\langle X_j - \boldsymbol{\mu}_1, u\rangle - \mathbb{E}_X w_1(X;\boldsymbol{\mu})\langle X - \boldsymbol{\mu}_1, u\rangle| \\
\leq& \frac{1}{n}\sum_{j=1}^n (\|X_j - \boldsymbol{\mu}^*_{Z_j}\| + 2R_{\max}) + \mathbb{E}_X\|X - \boldsymbol{\mu}^*_{Z_X}\| + 2R_{\max} \\
\leq& \frac{1}{n}\sum_{j=1}^n \|X_j - \boldsymbol{\mu}^*_{Z_j}\| + \mathbb{E}_X\|X - \boldsymbol{\mu}^*_{Z_X}\| + 4R_{\max}
\end{aligned}$$

Therefore we have,

$$\begin{aligned}
&\mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)] \leq \sum_{i=1}^n \mathbb{E}_{\mathbf{X}}[(\frac{1}{n}\sum_{j=1}^n \|X_j - \boldsymbol{\mu}^*_{Z_j}\| + \mathbb{E}_X\|X - \boldsymbol{\mu}^*_{Z_X}\| + 4R_{\max})1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)] \\
&\leq \sum_{i=1}^n \mathbb{E}_{\mathbf{X}}[(\frac{1}{n}\sum_{j=1}^n \|X_j - \boldsymbol{\mu}^*_{Z_j}\| 1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)]] + (\mathbb{E}_X\|X - \boldsymbol{\mu}^*_{Z_X}\| + 4R_{\max})P(\mathbf{X} \in (\mathcal{Y}_r^i)^c)] \qquad (37) \\
&\leq \sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n \mathbb{E}_{\mathbf{X}}[\|X_j - \boldsymbol{\mu}^*_{Z_j}\| 1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)] + c'(R_{\max} + d)p
\end{aligned}$$

where the last inequality follows from Lemma 5. Note that when $j \neq i$, the expectation factors due to independence of the sample points and by Lemma 8,

$$\mathbb{E}_{\mathbf{X}}[\|X_j - \boldsymbol{\mu}^*_{Z_j}\| 1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)] = \mathbb{E}_{X_j}\|X_j - \boldsymbol{\mu}^*_{Z_j}\| \cdot P(\|X_i - \boldsymbol{\mu}^*_{Z_i}\| \geq r) \leq cde^{-\frac{r\sqrt{d}}{2}}$$

When $j = i$, from Lemma 7,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\|X_j - \boldsymbol{\mu}^*_{Z_j}\| 1(\mathbf{X} \in (\mathcal{Y}_r^i)^c)] \leq& cn\int_{v=r}^{\infty} v \cdot v^{d-1}(v + R_{\max} + a)\exp(-v^2/2)dv \cdot \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \\
\leq& c_1 d \exp\left(-\frac{r\sqrt{d}}{2}\right)
\end{aligned}$$

Putting back to Eq. (37), we have

$$\begin{aligned}
\mathbb{E}[g(\mathbf{X})1(\mathbf{X} \notin \mathcal{Y}_r)] \leq& c_1 nd \exp\left(-\frac{r\sqrt{d}}{2}\right) + c_2 d \exp\left(-\frac{r\sqrt{d}}{2}\right) + c_3 n(R_{\max} + d)\exp\left(-\frac{r\sqrt{d}}{2}\right) \\
\leq& cn(R_{\max} + d)\exp\left(-\frac{r\sqrt{d}}{2}\right)
\end{aligned}$$

Following from Eq. (36), we have

$$|m_r - \mathbb{E}g(X)| \leq \frac{c_1 n \exp(-\frac{r}{2}\sqrt{d}) R_n(\mathcal{F}) + c_2 n(R_{\max} + d)\exp(-\frac{r}{2}\sqrt{d})}{1 - c_3 n \exp(-c_4 r\sqrt{d})} \tag{38}$$

Recall that we take $r = \Theta(\sqrt{d}(1 + R_{\max})\log^2 n)$, for large enough n, we have $1 - c_3 n \exp(-cr\sqrt{d}) \geq 1/2$, and $ne^{-cr\sqrt{d}} \leq C/n$. Finally for the second part of the numerator in Eq. (38) we have:

$$\begin{aligned} n(R_{\max} + d)\exp(-\sqrt{d}r/2) &\leq (R_{\max} + 1)\exp(\log n + \log d - \Theta(d(1 + R_{\max})\log^2 n)) \\ &\leq C'(R_{\max} + 1)/\sqrt{n}. \end{aligned}$$

Eq (38) becomes,

$$m_r \leq 2R_n(\mathcal{F})(1 + O(1/n)) + O((R_{\max} + 1)/\sqrt{n}) \tag{39}$$

Thus using Eqs (34), (35) and (39) the final bound becomes:

$$\begin{aligned} &P(g(X) \leq 2R_n(\mathcal{F})(1 + O(1/n)) + O((R_{\max} + 1)/\sqrt{n}) + (1 + R_{\max})d\sqrt{\log^5 n/n}) \\ \geq& 1 - P_1 - P_2 \\ \geq& 1 - c\exp\left(-c' \min((1 + R_{\max})^2 d\log n, d\log n)\right) \geq 1 - \exp\left(-cd\log n\right) \end{aligned}$$

Finally we have,

$$P(g(\mathbf{X}) = \tilde{O}(\max\{R_n(\mathcal{F}), (1 + R_{\max})dlog^{5/2}(n)/\sqrt{n}\})) \geq 1 - \exp\left(-cd\log n\right)$$

□

In learning theory, we have the following symmetrization lemma.

Lemma 14 (See e.g. [15])**.** *Let $\mathcal{F}$ be a function class with domain X. Let $\{X_1, X_2, \cdots, X_n\}$ be a set of sample generated by a distribution $\mathbb{P}$ on X. Assume σ_i are i.i.d. Rademacher variables, then*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}(\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i))\right) \leq 2R_n(\mathcal{F})$$

Here $R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\right]$ is the Rademacher complexity.

Proof of Theorem 5. Denote $Z_i = \sup_{\boldsymbol{\mu}\in\mathbb{A}} \left\|G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu})\right\|$ where

$$G(\boldsymbol{\mu}) = \begin{pmatrix} \mathbb{E}w_1(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_1) \\ \mathbb{E}w_2(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_2) \\ \vdots \\ \mathbb{E}w_M(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_3) \end{pmatrix}.$$

Assume $\mathcal{S}^{n-1}$ is n dimensional unit sphere, then $Z_i = \sup_{u\in\mathcal{S}^{d-1}} \sup_{\boldsymbol{\mu}\in\mathbb{A}} \langle G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu}), u\rangle$. Without loss of generality, we assume $i = 1$, the proof for other clusters follows simlilarly. Denote $Z_u = \sup_{\boldsymbol{\mu}\in A} \langle G^{(1)}(\boldsymbol{\mu}) - G_n^{(1)}(\boldsymbol{\mu}), u\rangle$. Let $\{u^{(1)}, u^{(2)}, \cdots, u^{(K)}\}$ be a 1/2-covering of the unit sphere $\mathcal{S}^{d-1}$, then $\forall v \in \mathcal{S}^{d-1}, \exists j \in [K]$, s.t. $\left\|v - u^{(j)}\right\| \leq \frac{1}{2}$. Hence we have

$$Z_v \leq Z_{u^{(j)}} + |Z_v - Z_{u^{(j)}}| \leq \max_j Z_{u^j} + Z\left\|v - u^{(j)}\right\|$$

As a result, $Z \leq 2\max_{j=1,\cdots,K} Z_{u^{(j)}}$. Therefore it is sufficient to bound Z_u for a fixed $u^{(j)} \in \mathcal{S}^{d-1}$. By Lemma 4, covering number $K \leq \exp(2d)$.
By Lemma 2, we have with probability at least $1-\exp(-cd\log n)$, $Z_{u_j} = \tilde{O}(\max\{R_n(\mathcal{F}_1^u), (1+R_{\max})d/\sqrt{n}\})$. Plugging in the Rademacher complexity from Proposition 1, and applying union bound, we have

$$Z_1 \leq 2\max_j Z_{u_j} \leq \tilde{O}(\max\{n^{-1/2}M^3(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}, (1+R_{\max})d/\sqrt{n}\})$$

with probability at least $1-\exp(2d-cd\log n) = 1-\exp(-c'd\log n)$. □

Proof of Theorem 2. We show the result by induction. When $t=1$,

$$\begin{aligned}\left\|\boldsymbol{\mu}^1 - \boldsymbol{\mu}^*\right\|_2 = \left\|G_n(\boldsymbol{\mu}^0) - \boldsymbol{\mu}^*\right\| &\leq \left\|G(\boldsymbol{\mu}^0) - \boldsymbol{\mu}^*\right\| + \left\|G_n(\boldsymbol{\mu}^0) - G(\boldsymbol{\mu}^0)\right\| \\ &\leq \zeta\left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\| + \epsilon^{\text{unif}}(n,\delta)\end{aligned}$$

If $\left\|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\right\| < a$ and $\epsilon^{\text{unif}}(n,\delta) \leq (1-\zeta)a$, we have $\left\|\boldsymbol{\mu}_i^{t+1} - \boldsymbol{\mu}_i^*\right\| \leq a$. So $\boldsymbol{\mu}^t$ lies in the contraction region for $\forall t \geq 0$.
Then iteratively we get

$$\begin{aligned}\left\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\right\| &\leq \zeta\left\|\boldsymbol{\mu}^{t-1} - \boldsymbol{\mu}^*\right\| + \epsilon^{\text{unif}}(n,\delta) \\ &\leq \zeta^t\left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\| + \sum_{i=0}^{t-1}\zeta^i\epsilon^{\text{unif}}(n,\delta) \\ &\leq \zeta^t\left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\| + \frac{1}{1-\zeta}\epsilon^{\text{unif}}(n,\delta)\end{aligned}$$

with probability at least $1-\delta$. □

References

[1] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 02 2017.

[2] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[3] Richard Combes. An extension of mcdiarmid's inequality. *arXiv preprint arXiv:1511.05240*, 2015.

[4] Denis Conniffe. Expected maximum log likelihood estimation. *The Statistician*, pages 317–329, 1987.

[5] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[6] Sanjoy Dasgupta and Leonard J Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 152–159. Morgan Kaufmann Publishers Inc., 2000.

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

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

[9] Kenneth Lange. A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 425–437, 1995.

[10] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes.* Springer Science & Business Media, 2013.

[11] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[12] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[13] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[14] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*, 2016.

[15] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning.* MIT press, 2012.

[16] Iftekhar Naim and Daniel Gildea. Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. *arXiv preprint arXiv:1206.6427*, 2012.

[17] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

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

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

[20] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2676–2684. Curran Associates, Inc., 2016.

[21] Lei Xu and Michael I Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

[22] Bowei Yan and Purnamrita Sarkar. On robustness of kernel clustering. In *Advances in Neural Information Processing Systems*, pages 3090–3098, 2016.