Ten Steps of EM Suffice for Mixtures of Two Gaussians

Constantinos Daskalakis
EECS and CSAIL, MIT
costis@mit.edu

Christos Tzamos
EECS and CSAIL, MIT
tzamos@mit.edu

Manolis Zampetakis
EECS and CSAIL, MIT
mzampet@mit.edu

Abstract

The Expectation-Maximization (EM) algorithm is a widely used method for maximum likelihood estimation in models with latent variables. For estimating mixtures of Gaussians, its iteration can be viewed as a soft version of the k-means clustering algorithm. Despite its wide use and applications, there are essentially no known convergence guarantees for this method. We provide global convergence guarantees for mixtures of two Gaussians with known covariance matrices. We show that the population version of EM, where the algorithm is given access to infinitely many samples from the mixture, converges geometrically to the correct mean vectors, and provide simple, closed-form expressions for the convergence rate. As a simple illustration, we show that, in one dimension, ten steps of the EM algorithm initialized at infinity result in less than 1% error estimation of the means. In the finite sample regime, we show that, under a random initialization, $\tilde{O}(d/\epsilon^2)$ samples suffice to compute the unknown vectors to within ϵ in Mahalanobis distance, where d is the dimension. In particular, the error rate of the EM based estimator is $\tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ where n is the number of samples, which is optimal up to logarithmic factors.

1 Introduction

The *Expectation-Maximization (EM) algorithm* [DLR77, Wu83, RW84] is one of the most widely used heuristics for maximizing likelihood in statistical models with latent variables. Consider a probability distribution $p_{\boldsymbol{\lambda}}$ sampling $(\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{X}$ is a vector of observable random variables, $\boldsymbol{Z}$ a vector of non-observable random variables and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ a vector of parameters. Given independent samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the observed random variables, the goal of maximum likelihood estimation is to select $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ maximizing the log-likelihood of the samples, namely $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Unfortunately, computing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$ involves summing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ over all possible values of $\boldsymbol{z}_i$, which commonly results in a log-likelihood function that is non-convex with respect to $\boldsymbol{\lambda}$ and therefore hard to optimize. In this context, the EM algorithm proposes the following heuristic:

- Start with an initial guess $\boldsymbol{\lambda}^{(0)}$ of the parameters.
- For all $t \geq 0$, until convergence:
  - (E-Step) For each sample i, compute the posterior $Q_i^{(t)}(\boldsymbol{z}) := p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{Z} = \boldsymbol{z} | \boldsymbol{X} = \boldsymbol{x}_i)$.
  - (M-Step) Set $\boldsymbol{\lambda}^{(t+1)} := \arg\max_{\boldsymbol{\lambda}} \sum_i \sum_{\boldsymbol{z}} Q_i^{(t)}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z})}{Q_i^{(t)}(\boldsymbol{z})}$.

Intuitively, the E-step of the algorithm uses the current guess of the parameters, $\boldsymbol{\lambda}^{(t)}$, to form beliefs, $Q_i^{(t)}$, about the state of the (non-observable) $\boldsymbol{Z}$ variables for each sample i. Then the M-step uses the new beliefs about the state of $\boldsymbol{Z}$ for each sample to maximize with respect to $\boldsymbol{\lambda}$ a lower bound on $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Indeed, by the concavity of the log function, the objective function used in the M-step of the algorithm is a lower bound on the true log-likelihood for all values of $\boldsymbol{\lambda}$, and it equals the true log-likelihood for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. From these observations, it follows that the above alternating procedure improves the true log-likelihood until convergence.

Despite its wide use and practical significance, little is known about whether and under what conditions EM converges to the true maximum likelihood estimator. A few works establish local convergence of the algorithm to stationary points of the log-likelihood function [Wu83, Tse04, CH08], and even fewer local convergence to the MLE [RW84, BWY17]. Besides local convergence, it is also known that badly initialized EM may settle far from the MLE both in parameter and in likelihood distance [Wu83]. The lack of theoretical understanding of the convergence properties of EM is intimately related to the non-convex nature of the optimization it performs.

Our paper aims to illuminate why EM works well in practice and develop techniques for understanding its behavior. We do so by analyzing one of the most basic and natural, yet still challenging, statistical models EM may be applied to, namely balanced mixtures of two multi-dimensional Gaussians with equal and known covariance matrices. In particular, we study the convergence of EM when applied to the following family of parametrized density functions:

$$p_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma),$$

where Σ is a known covariance matrix, $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ are unknown (vector) parameters, and $\mathcal{N}(\boldsymbol{\mu}, \Sigma; \boldsymbol{x})$ represents the Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ, i.e.

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp\left(-0.5(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right).$$

Our main contribution is to provide global convergence guarantees for EM applied to the above family of distributions. We establish our result for both the "population version" of the algorithm, and the finite-sample version, as described below.

Analysis of Population EM for Mixtures of Two Gaussians. To elucidate the optimization features of the algorithm and avoid analytical distractions arising due to sampling error, it has been standard practice in the literature of theoretical analyses of EM to consider the "population version" of the algorithm, where the EM iterations are performed assuming access to infinitely many samples from a distribution $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}$ as above. With infinitely many samples, we can identify the mean, $\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, of $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}$, and re-parametrize the density around the mean as follows:

$$p_{\boldsymbol{\mu}}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \Sigma). \tag{1.1}$$

We first study the convergence of EM when we perform iterations with respect to the parameter $\boldsymbol{\mu}$ of $p_{\boldsymbol{\mu}}(\boldsymbol{x})$ in (1.1). Starting with an initial guess $\boldsymbol{\lambda}^{(0)}$ for the unknown mean vector $\boldsymbol{\mu}$, the t-th iteration of EM amounts to the following update:

$$\boldsymbol{\lambda}^{(t+1)} = M(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}) \triangleq \frac{\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}\boldsymbol{x}\right]}{\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}\right]}, \tag{1.2}$$

where we have compacted both the E- and M-step of EM into one update.

The intuition behind the EM update formula is as follows. First, we take expectations with respect to $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ because we are studying the population version of EM, hence we assume access to infinitely many samples from $p_{\boldsymbol{\mu}}$. For each sample $\boldsymbol{x}$, the ratio $\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}$ is our belief, at step t, that $\boldsymbol{x}$ was sampled from the first Gaussian component of $p_{\boldsymbol{\mu}}$, namely the one for which our current estimate of its mean vector is $\boldsymbol{\lambda}^{(t)}$. (The complementary probability is our present belief that $\boldsymbol{x}$ was sampled from the other Gaussian component.) Given these beliefs for all vectors $\boldsymbol{x}$, the update (1.2) is the result of the M-step of EM. Intuitively, our next guess $\boldsymbol{\lambda}^{(t+1)}$ for the mean vector of the first Gaussian component is a weighted combination over all samples $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ where the weight of every $\boldsymbol{x}$ is our belief that it came from the first Gaussian component.

Our main result for population-EM is the following:

Informal Theorem (Population EM Analysis)**.** *Whenever the initial guess $\boldsymbol{\lambda}^{(0)}$ is not equidistant to $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, EM converges geometrically to either $\boldsymbol{\mu}$ or $-\boldsymbol{\mu}$, with convergence rate that improves as $t \to \infty$. We provide a simple, closed form expression of the convergence rate as a function of $\boldsymbol{\lambda}^{(t)}$ and $\boldsymbol{\mu}$. If the initial guess $\boldsymbol{\lambda}^{(0)}$ is equidistant to $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, EM converges to the unstable fixed point $\mathbf{0}$.*

A formal statement is provided as Theorem 2 in Section 4. We start with the proof of the single-dimensional version, presented as Theorem 1 in Section 3. As a simple illustration of our result, we show in Section 5 that, in one dimension, when our original guess $\lambda^{(0)} = +\infty$ and the signal-to-noise ratio $\mu/\sigma = 1$, 10 steps of the EM algorithm result in 1% error.

Despite the simplicity of the case we consider, no global convergence results were known prior to our work, even for the population EM. [BWY17] studied the same setting proving only local convergence, i.e. convergence only when the initial guess is close to the true parameters. They argue that the population EM update is contracting close to the true parameters. Unfortunately, the EM update is non-contracting outside a small neighborhood of the true parameters so this argument cannot be used for a global convergence guarantee.

In this work, we study the problem under arbitrary starting points and completely characterize the fixed points of EM. We show that other than a measure-zero subset of the space (namely points that are equidistant from the centers of the two Gaussians), any initialization of the EM algorithm converges to the true centers of the Gaussians, providing explicit bounds for the convergence rate.

To achieve this, we follow an orthogonal approach to [BWY17]: Instead of trying to directly compute the number of steps required to reach convergence *for a specific instance of the problem*, we study *the sensitivity of the EM iteration as the instance varies*. The intuition is that if the EM update is sensitive to updating the instance, then changing the instance should also attract the update towards the cha that keeping the instance fixed, o articular, we gain a handle on the ed by Eq. (3.2).

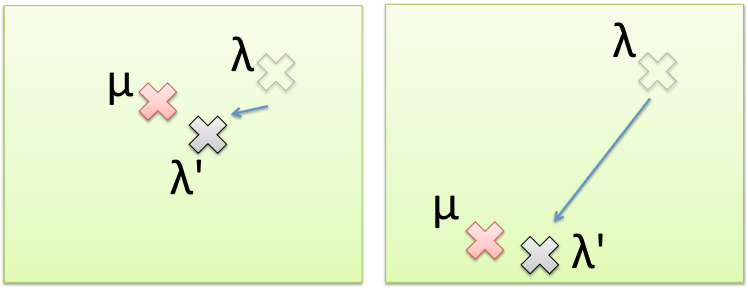

Figure 1: Sensitivity of the EM update when changing the true parameters. Large sensitivity implies large progress towards the true parameters. See Eq. (3.2) for a quantification.

Analysis of Finite-Sample EM for Mixtures of Two Gaussians. The finite sample analysis proceeds in three steps. First, in the finite sample regime we do not know the average of the two mean vectors, $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, exactly. We show that, with $\tilde{O}(d/\epsilon^2)$ samples, we can approximate the average to within Mahalanobis distance ϵ. We then chain two coupling arguments. The first compares the progress towards the true mean made by the correctly centered population EM update to that of the incorrectly centered population EM update. The second compares the progress towards the true mean made by the incorrectly centered population EM update with the progress made by the incorrectly centered finite sample EM update. See Figure 2 and Theorem 3. Given the error incurred in the approximation of the center $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, we propose to stabilize the sample-based EM iteration by including in the sample for each sampled point $\boldsymbol{x}_i$ its symmetric point $-\boldsymbol{x}_i$. This is the sample based version that we analyze, although our analysis goes through without this stabilization. Our result is the following, formally given as Theorem 3 in Section 6.

Informal Theorem (Finite Sample EM Analysis)**.** *Whenever $\epsilon < SNR$, $\tilde{O}(d/\epsilon^2 \cdot \mathrm{poly}(1/SNR))$ samples suffice to approximate $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ to within Mahalanobis distance ϵ using the EM algorithm. In particular, the error rate of the EM based estimator is $\tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ where n is the number of samples, which is optimal up to logarithmic factors.*[1]

Bootstrapping EM for Faster Convergence. We note that, in multiple dimensions, care must be taken in initializing the EM algorithm, even in the infinite sample regime, as the convergence guarantee depends on the angle between the current iterate and the true mean vector. While a randomly chosen unit vector will have projection of $\Theta(1/\sqrt{d})$ in the direction of $\boldsymbol{\mu}$, we argue that we can boostrap EM to turn this projection larger than a constant. This allows us to work with similar convergence rates as in the single-dimensional case, namely only SNR (and not dimension) dependent. Our initialization procedure is described in Section 6.3.

[1] Note that even if SNR is arbitrarily large (so that the two Gaussian components are "perfectly separated") the problem degenerates to finding the mean of one Gaussian whose optimal rate is $\Omega\left(\sqrt{\frac{d}{n}}\right)$.

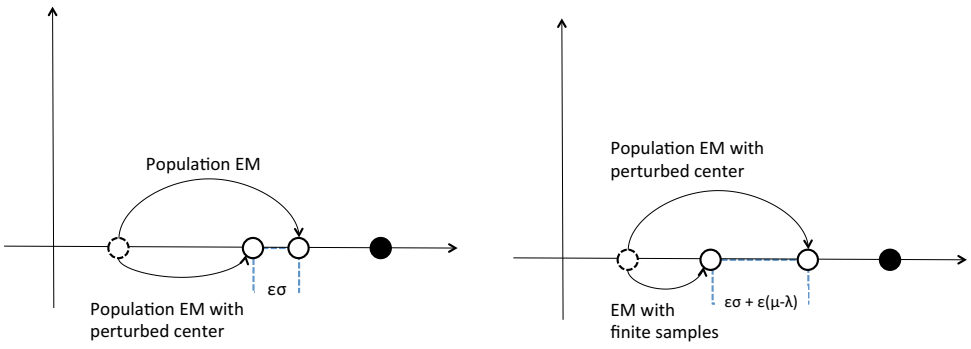

Figure 2: (a) Coupling correctly and incorrectly centered population EM updates. We show that, starting from the same iterate, the correctly and incorrectly centered population EM updates will land to close-by points. This quantified by Eq. (6.13). (b) Coupling incorrectly centered population EM and finite sample EM updates. We show that, starting from the same iterate, the incorrectly centered population EM update and the finite sample update land to close-by points. This quantified by Lemma 10.

Informal Theorem (EM Initialization)**.** *EM can be boostrapped so that the number of iterations required to approximate* $\boldsymbol{\mu}_1$ *and* $\boldsymbol{\mu}_2$ *to within Mahalanobis distance* ϵ *depends logarithmically in the dimension.*

Related Work on Learning Mixtures of Gaussians. We have already outlined the literature on the Expectation-Maximization algorithm. Several results study its local convergence properties and there are known cases where badly initialized EM fails to converge. See above.

There is also a large body of literature on learning mixtures of Gaussians. A long line of work initiated by Dasgupta [Das99, AK01, VW04, AM05, KSV05, DS07, CR08, BV08, CDV09] provides rigorous guarantees on recovering the parameters of Gaussians in a mixture under separability assumptions, while later work [KMV10, MV10, BS10] has established guarantees under minimal information theoretic assumptions. More recent work [HP15] provides tight bounds on the number of samples necessary to recover the parameters of the Gaussians as well as improved algorithms, while another strand of the literature studies proper learning with improved running times and sample sizes [SOAJ14, DK14]. Finally, there has been work on methods exploiting general position assumptions or performing smoothed analysis [HK13, GHK15].

In practice, the most common algorithm for learning mixtures of Gaussians is the Expectation-Maximization algorithm, with the practical experience that it performs well in a broad range of scenarios despite the lack of theoretical guarantees. Recently, Balakrishnan, Wainwright and Yu [BWY17] studied the convergence of EM in the case of an equal-weight mixture of two Gaussians with the same and known covariance matrix, showing local convergence guarantees. In particular, they show that when EM is initialized close to the actual parameters, then it converges. In this work, we revisit the same setting considered by [BWY17] but establish *global convergence guarantees*. We show that, for any initialization of the parameters, the EM algorithm converges geometrically to the true parameters. We also provide a simple and explicit formula for the rate of convergence.

Concurrent and independent work by Xu, Hsu and Maleki [XHM16] has also provided global and geometric convergence guarantees for the same setting, as well as a slightly more general setting where the mean of the mixture is unknown, but they do not provide explicit convergence rates. They also do not provide an analysis of the finite-sample regime.

2 Preliminary Observations

In this section we illustrate some simple properties of the EM update (1.2) and simplify the formula. First, it is easy to see that plugging in the values $\boldsymbol{\lambda} \in \{-\boldsymbol{\mu}, \mathbf{0}, \boldsymbol{\mu}\}$ into $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ results into

$$M(-\boldsymbol{\mu}, \boldsymbol{\mu}) = -\boldsymbol{\mu} \quad ; \quad M(\mathbf{0}, \boldsymbol{\mu}) = \mathbf{0} \quad ; \quad M(\boldsymbol{\mu}, \boldsymbol{\mu}) = \boldsymbol{\mu}. \tag{2.1}$$

In particular, for all $\boldsymbol{\mu}$, these values are certainly fixed points of the EM iteration. Next, we rewrite $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ as follows:

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\boldsymbol{x}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(-\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\boldsymbol{x}\right]}{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(-\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\right]}.$$

It is easy to observe that by symmetry this simplifies to

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)-\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\boldsymbol{x}\right]}{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\right]} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)-\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\boldsymbol{x}\right].$$

Simplifying common terms in the density functions $\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)$, we get that

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\exp\left(\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right)-\exp\left(-\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right)}{\exp\left(\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right)+\exp\left(-\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right)}\boldsymbol{x}\right].$$

We thus get the following expression for the EM iteration

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\tanh(\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x})\boldsymbol{x}\right]. \tag{2.2}$$

3 Single-dimensional Convergence

In the single dimensional case the EM algorithm takes the following form according to (2.2).

$$\lambda^{(t+1)} = M(\lambda^{(t)}, \mu) = \mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[\tanh\left(\frac{\lambda^{(t)}x}{\sigma^2}\right)x\right] \tag{3.1}$$

Observe that the function $M(\lambda, \mu)$ is increasing with respect to λ. Indeed the partial derivative of M with respect to λ is

$$\frac{\partial M(\lambda,\mu)}{\partial\lambda} = \mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[\tanh'\left(\frac{\lambda^{(t)}x}{\sigma^2}\right)\frac{x^2}{\sigma^2}\right]$$

which is strictly greater than zero since the $\tanh'$ function is strictly positive.

We will show next that the fixed points we identified at (2.1) are the only fixed points of $M(\cdot, \mu)$. When initialized with $\lambda^{(0)} > 0$ (resp. $\lambda^{(0)} < 0$), the EM algorithm converges to $\mu > 0$ (resp. to $-\mu < 0$). The point $\lambda = 0$ is an unstable fixed point.

Theorem 1. *In the single dimensional case, when $\lambda^{(0)}, \mu > 0$, the parameters $\lambda^{(t)}$ satisfy*

$$\left|\lambda^{(t+1)} - \mu\right| \le \kappa^{(t)}\left|\lambda^{(t)} - \mu\right| \qquad \text{where } \kappa^{(t)} = \exp\left(-\frac{\min(\lambda^{(t)}, \mu)^2}{2\sigma^2}\right)$$

Moreover $\kappa^{(t)}$ is a decreasing function of t.

Proof. For simplicity we will use λ for $\lambda^{(t)}$, λ' for $\lambda^{(t+1)}$ and we will assume that $X \sim \mathcal{N}(0, \sigma^2)$.

By a simple change of variables we can see that

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

The main idea is to use the Mean Value Theorem with respect to the second coordinate of the function M on the interval $[\lambda, \mu]$.

$$\frac{M(\lambda,\mu) - M(\lambda,\lambda)}{\mu - \lambda} = \left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi} \text{ with } \xi \in (\lambda, \mu)$$

But we know that $M(\lambda, \lambda) = \lambda$ and $M(\lambda, \mu) = \lambda'$ and therefore we get

$$\lambda' - \lambda \geq \left(\min_{\xi \in [\lambda,\mu]} \left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi}\right)(\mu - \lambda)$$

which is equivalent to

$$\left|\lambda' - \mu\right| \leq \left(1 - \min_{\xi \in [\lambda,\mu]} \left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi}\right)|\lambda - \mu| \tag{3.2}$$

where we have used the fact that $\lambda' < \mu$ which is comes from the fact that $M(\lambda, \mu)$ is increasing with respect to λ and that $M(\mu, \mu) = \mu$.

The only thing that remains to complete our proof is to prove a lower bound of the partial derivative of M with respect to μ.

$$\left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi} = \mathbb{E}\left[\frac{\lambda}{\sigma^2}\tanh'\left(\frac{\lambda(X+\xi)}{\sigma^2}\right)(X+\xi) + \tanh\left(\frac{\lambda(X+\xi)}{\sigma^2}\right)\right]$$

The first term is non-negative, Lemma 1. The second term is at least $1 - \exp\left[-\frac{\min(\xi,\lambda)\cdot\xi}{2\sigma^2}\right]$, Lemma 2 and the theorem follows. □

Lemma 1. *Let $\alpha, \beta > 0$ and $X \sim \mathcal{N}(\alpha, \sigma^2)$ then $\mathbb{E}\left[\tanh'\left(\beta X/\sigma^2\right) X\right] \geq 0$.*

Proof.

$$\mathbb{E}\left[\tanh'\left(\frac{\beta X}{\sigma^2}\right)X\right] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\tanh'\left(\frac{\beta y}{\sigma^2}\right) y \exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right)dy$$

But now we can see that since $\tanh'$ is an even function and since for any $y > 0$ we have $\exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right) \geq \exp\left(-\frac{(-y-\alpha)^2}{2\sigma^2}\right)$ then

$$-\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{0}\tanh'\left(\frac{\beta y}{\sigma^2}\right) y \exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right)dy \leq \frac{1}{\sqrt{2\pi}\sigma}\int_{0}^{\infty}\tanh'\left(\frac{\beta y}{\sigma^2}\right) y \exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right)dy$$

which means that $\mathbb{E}\left[\tanh'\left(\beta X/\sigma^2\right) X\right] \geq 0$. □

Lemma 2. *Let $\alpha, \beta > 0$ and $X \sim \mathcal{N}(\alpha, \sigma^2)$ then $\mathbb{E}\left[\tanh\left(\beta X/\sigma^2\right)\right] \geq 1 - \exp\left[-\frac{\min(\alpha,\beta)\cdot\alpha}{2\sigma^2}\right]$.*

Proof. Note that $\mathbb{E}\left[\tanh\left(\beta X/\sigma^2\right)\right]$ is increasing as a function of β as its derivative with respect to β is positive by Lemma 1. It thus suffices to show that $\mathbb{E}\left[\tanh\left(\beta X/\sigma^2\right)\right] \geq 1 - \exp\left[-\frac{\alpha\beta}{2\sigma^2}\right]$ when $\beta \leq \alpha$. We have that

$$\begin{aligned}
\mathbb{E}\left[1-\tanh\left(\beta X/\sigma^2\right)\right] &= \mathbb{E}\left[\frac{2}{\exp(2\beta X/\sigma^2)+1}\right] \leq \mathbb{E}\left[\frac{1}{\exp(\beta X/\sigma^2)}\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\frac{\exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right)}{\exp(\beta x/\sigma^2)}dx = \frac{\exp\left(\frac{(\alpha-\beta)^2-\alpha^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\alpha+\beta)^2}{2\sigma^2}\right)dx \\
&= \exp\left(\frac{(\alpha-\beta)^2-\alpha^2}{2\sigma^2}\right) \leq \exp\left(-\frac{\alpha\beta}{2\sigma^2}\right)
\end{aligned}$$

which completes the proof. □

4 Multi-dimensional Convergence

In the multidimensional case, the EM algorithm takes the form of (2.2). In this case, we will quantify our approximation guarantees using the *Mahalanobis distance* $\|\cdot\|_\Sigma$ between vectors with respect to matrix Σ, defined as follows:

$$\|\boldsymbol{x}-\boldsymbol{y}\|_\Sigma = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{y})}.$$

We will show that the fixed points identified in (2.1) are the only fixed points of $M(\cdot,\boldsymbol{\mu})$. When initialized with $\boldsymbol{\lambda}^{(0)}$ such that $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma < \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$ (resp. $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma > \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$), the EM algorithm converges to $\boldsymbol{\mu}$ (resp. to $-\boldsymbol{\mu}$). The algorithm converges to $\boldsymbol{\lambda}=\mathbf{0}$ when initialized with $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma = \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$. In particular,

Theorem 2. *Whenever $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma < \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$, i.e. the initial guess is closer to $\boldsymbol{\mu}$ than $-\boldsymbol{\mu}$, the estimates $\boldsymbol{\lambda}^{(t)}$ of the EM algorithm satisfy*

$$\left\|\boldsymbol{\lambda}^{(t+1)}-\boldsymbol{\mu}\right\|_\Sigma \leq \kappa^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}\right\|_\Sigma, \quad \text{where } \kappa^{(t)} = \exp\left(-\frac{\min\left(\boldsymbol{\lambda}^{(t),T}\Sigma^{-1}\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\lambda}^{(t)}\right)^2}{2\boldsymbol{\lambda}^{(t),T}\Sigma^{-1}\boldsymbol{\lambda}^{(t)}}\right).$$

Moreover, $\kappa^{(t)}$ is a decreasing function of t. The symmetric things hold when $\left\|\boldsymbol{\lambda}^{(0)}-\boldsymbol{\mu}\right\|_\Sigma > \left\|\boldsymbol{\lambda}^{(0)}+\boldsymbol{\mu}\right\|_\Sigma$. When the initial guess is equidistant to $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, then $\boldsymbol{\lambda}^{(t)}=\mathbf{0}$ for all $t>0$.

Proof. For simplicity we will use $\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}^{(t)}$, $\boldsymbol{\lambda}'$ for $\boldsymbol{\lambda}^{(t+1)}$.

By applying the following change of variables $\boldsymbol{\lambda} \leftarrow \Sigma^{-1/2}\boldsymbol{\lambda}$ and $\boldsymbol{\mu} \leftarrow \Sigma^{-1/2}\boldsymbol{\mu}$ we may assume that $\Sigma = I$ where I is the identity matrix. Therefore the iteration of EM becomes

$$M(\boldsymbol{\lambda},\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},I)}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle)\boldsymbol{x}\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(\boldsymbol{x}+\boldsymbol{\mu})\right]$$

Let $\hat{\boldsymbol{\lambda}}$ be the unit vector in the direction of $\boldsymbol{\lambda}$, $\hat{\boldsymbol{\lambda}}^\perp$ be the unit vector that belongs to the plane of $\boldsymbol{\mu},\boldsymbol{\lambda}$ and is perpendicular to $\boldsymbol{\lambda}$, and let $\{\boldsymbol{v}_1=\hat{\boldsymbol{\lambda}}, \boldsymbol{v}_2=\hat{\boldsymbol{\lambda}}^\perp, \boldsymbol{v}_3, ..., \boldsymbol{v}_d\}$ be a basis of $\mathbb{R}^d$. We have:

$$\langle\boldsymbol{v_i},\boldsymbol{\lambda}'\rangle = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I)}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(\langle\boldsymbol{v_i},\boldsymbol{x}\rangle+\langle\boldsymbol{v_i},\boldsymbol{\mu}\rangle)\right] \tag{4.1}$$

Since the Normal distribution is rotation invariant we can equivalently write:

$$\langle\boldsymbol{v_i},\boldsymbol{\lambda}'\rangle = \mathbb{E}_{\alpha_1,...,\alpha_d\sim\mathcal{N}(0,1)}\left[\tanh(\langle\boldsymbol{\lambda},\sum_j\alpha_j\boldsymbol{v}_j\rangle+\langle\boldsymbol{\lambda},\boldsymbol{\mu}\rangle)(\langle\boldsymbol{v_i},\sum_j\alpha_j\boldsymbol{v}_j\rangle+\langle\boldsymbol{v_i},\boldsymbol{\mu}\rangle)\right]$$

which simplifies to

$$\langle \boldsymbol{v_i}, \boldsymbol{\lambda}' \rangle = \mathbb{E}_{\alpha_1,\ldots,\alpha_d \sim \mathcal{N}(0,1)} \left[\tanh(\alpha_1 \left\| \boldsymbol{\lambda} \right\| + \langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle)(a_i + \langle \boldsymbol{v_i}, \boldsymbol{\mu} \rangle) \right] =$$

$$\mathbb{E}_{\alpha_1 \sim \mathcal{N}(0,1)} \left[\tanh(\alpha_1 \left\| \boldsymbol{\lambda} \right\| + \langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle) \cdot (\mathbb{E}_{\alpha_2,\ldots,\alpha_d \sim \mathcal{N}(0,1)} \left[a_i \right] + \langle \boldsymbol{v_i}, \boldsymbol{\mu} \rangle) \right] \tag{4.2}$$

We now consider different cases for i to further simplify Equation (4.2).

– When $i = 1$, we have that $\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}' \rangle = \mathbb{E}_{y \sim \mathcal{N}(0,1)} \left[\tanh(\left\| \boldsymbol{\lambda} \right\| (y + \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle)) \left(y + \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \right) \right]$. This is equivalent with an iteration of EM in one dimension and thus from Theorem 1 we get that

$$|\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle - \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}' \rangle| \leq \kappa |\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle - \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \rangle| \tag{4.3}$$

where

$$\kappa = \exp\left(-\frac{\min(\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \rangle, \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle)^2}{2} \right) = \exp\left(-\frac{\min(\langle \boldsymbol{\lambda}, \boldsymbol{\lambda} \rangle, \langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle)^2}{2 \langle \boldsymbol{\lambda}, \boldsymbol{\lambda} \rangle} \right)$$

– When $i = 2$, $\mathbb{E}_{\alpha_2,\ldots,\alpha_d \sim \mathcal{N}(0,1)} \left[a_i \right] + \langle \boldsymbol{v_i}, \boldsymbol{\mu} \rangle = \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle$ and thus

$$\langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda}' \rangle = \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle \mathbb{E}_{y \sim \mathcal{N}(0,1)} \left[\tanh(\left\| \boldsymbol{\lambda} \right\| (y + \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle)) \right]$$

Let κ as defined before and using Lemma 2 we get that

$$\langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle \geq \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda}' \rangle \geq (1 - \kappa) \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle \tag{4.4}$$

– When $i \geq 3$, $\mathbb{E}_{\alpha_2,\ldots,\alpha_d \sim \mathcal{N}(0,1)} \left[a_i \right] + \langle \boldsymbol{v_i}, \boldsymbol{\mu} \rangle = 0$ and thus $\langle \boldsymbol{v_i}, \boldsymbol{\lambda}' \rangle = 0$.

We can now bound the distance of $\boldsymbol{\lambda}'$ from $\boldsymbol{\mu}$:

$$\left\| \boldsymbol{\lambda}' - \boldsymbol{\mu} \right\| = \sqrt{\sum_i \langle \boldsymbol{v}_i, \boldsymbol{\lambda}' - \boldsymbol{\mu} \rangle^2} = \sqrt{\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}' - \boldsymbol{\mu} \rangle^2 + \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda}' - \boldsymbol{\mu} \rangle^2}$$

$$\overset{(4.3),\ (4.4)}{\leq} \sqrt{\kappa^2 \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda} - \boldsymbol{\mu} \rangle^2 + \kappa^2 \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\lambda} - \boldsymbol{\mu} \rangle^2} \leq \kappa \left\| \boldsymbol{\lambda} - \boldsymbol{\mu} \right\|$$

We now have to prove that this convergence rate κ decreases as the iterations increase. This is implied by the following lemmas which show that $\min(\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \rangle, \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle) \leq \min(\langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\lambda}' \rangle, \langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle)$

Lemma 3. *If* $\left\| \boldsymbol{\lambda} \right\| \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$ *then* $\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \leq \left\| \boldsymbol{\lambda}' \right\|$ *and* $\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \leq \langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle$.

Proof. The analysis above implies that $\boldsymbol{\lambda}'$ can be written in the form $\boldsymbol{\lambda}' = \alpha \cdot \hat{\boldsymbol{\lambda}} + \beta \cdot \hat{\boldsymbol{\lambda}}^{\perp}$, where $\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \leq \alpha \leq \left\| \boldsymbol{\lambda} \right\|$ and $0 \leq \beta \leq \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle$. It is easy to see that the first inequality holds since $\left\| \boldsymbol{\lambda}' \right\| \geq \alpha \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$. For the second, we write $\langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle$ as:

$$\langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle = \frac{\langle \boldsymbol{\lambda}', \boldsymbol{\mu} \rangle}{\left\| \boldsymbol{\lambda}' \right\|} = \frac{\alpha \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle + \beta \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle}{\sqrt{\alpha^2 + \beta^2}} = \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \frac{1 + \frac{\langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle}{\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle} \frac{\beta}{\alpha}}{\sqrt{1 + \left(\frac{\beta}{\alpha} \right)^2}} \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle \frac{1 + \left(\frac{\beta}{\alpha} \right)^2}{\sqrt{1 + \left(\frac{\beta}{\alpha} \right)^2}} \geq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$$

where we used the fact that $\frac{\langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle}{\langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle} \geq \frac{\beta}{\alpha}$ which follows by the bounds on α and β. □

Lemma 4. *If* $\left\| \boldsymbol{\lambda} \right\| \leq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$ *then* $\left\| \boldsymbol{\lambda} \right\| \leq \left\| \boldsymbol{\lambda}' \right\| \leq \langle \hat{\boldsymbol{\lambda}}', \boldsymbol{\mu} \rangle$.

Proof. We have that $\boldsymbol{\lambda}' = \alpha \cdot \hat{\boldsymbol{\lambda}} + \beta \cdot \hat{\boldsymbol{\lambda}}^{\perp}$, where $\left\| \boldsymbol{\lambda} \right\| \leq \alpha \leq \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle$ and $0 \leq \beta \leq \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle$. We also have $\langle \boldsymbol{\lambda}', \boldsymbol{\mu} \rangle = \alpha \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu} \rangle + \beta \langle \hat{\boldsymbol{\lambda}}^{\perp}, \boldsymbol{\mu} \rangle \geq \alpha^2 + \beta^2 = \left\| \boldsymbol{\lambda}' \right\|^2 \geq \alpha^2 \geq \left\| \boldsymbol{\lambda} \right\|^2$ so the lemma follows. □

Finally substituting back in the basis that we started before changing coordinates to make the covariance matrix identity we get the result as stated at the theorem. □

5 An Illustration of the Speed of Convergence

Using our results in the previous sections we can calculate explicit speeds of convergence of EM to its fixed points. In this section, we present some results with this flavor. For simplicity, we start with single dimensional case, and discuss the multi-dimensional case in the end of this section.

Let us consider a mixture of two single-dimensional Gaussians whose signal-to-noise ratio $\eta = \mu/\sigma$ is equal to 1. There is nothing special about the value of 1, except that it is a difficult case to consider since the Gaussian components are not separated, as shown in Figure 3. When the SNR is

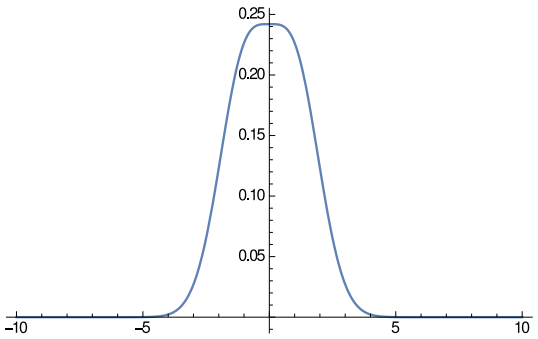

Figure 3: The density of $\frac{1}{2}\mathcal{N}(x;1,1) + \frac{1}{2}\mathcal{N}(x;-1,1)$.

larger, the numbers presented below still hold and in reality the convergence is even faster. When the SNR is even smaller than one, the numbers change, but gracefully, and they can be calculated in a similar fashion.

We will also assume a completely agnostic initialization of EM, setting $\lambda^{(0)} \to +\infty$.[2] To analyze the speed of convergence of EM to its fixed point μ, we first make the observation that in one step we already get to $\lambda^{(1)} \leq \mu + \sigma\sqrt{\frac{2}{\pi}}$. To see this we can plug $\lambda^{(0)} \to \infty$ into equation (3.1) to get:

$$\lambda^{(1)} = \mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[\text{sign}(x)x\right] = \mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[|x|\right],$$

which equals the mean of the Folded Normal Distribution. A well-known bound for this mean is $\mu + \sqrt{\frac{2}{\pi}}\sigma$. Therefore the distance from the true mean after one step is $\left|\lambda^{(1)} - \mu\right| \leq \sqrt{\frac{2}{\pi}}\sigma$.

Now, using Theorem 1, we conclude that in all subsequent steps the distance to μ shrinks by a factor of at least $e^{+1/2}$. This means that, if we want to estimate μ to within additive error $1\%\sigma$, then we need to run EM for at most $\lceil 2 \cdot \ln 100 + \ln\frac{2}{\pi}\rceil = 9$ additional steps. Accounting for the first step, 10 iterations of the EM algorithm in total suffice to get to within error 1%, even when our initial guess of the mean is infinitely away from the true value!

In Figure 4 we illustrate the speed of convergence of EM as implied by Theorem 2 in multiple dimensions. The plot was generated for a Gaussian mixture with $\boldsymbol{\mu} = (2\ 2)$ and $\Sigma = I$, but the behavior illustrated in this figure is generic (up to a transformation of the space by $\Sigma^{-\frac{1}{2}}$). As implied by Theorem 2, the rate of convergence depends on the distance of $\boldsymbol{\lambda}^{(t)}$ from the origin $\mathbf{0}$ and the angle $\langle\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}\rangle$. The figure shows the directions of the EM updates for every point, and the factor by which the distance to the fixed point decays, with deeper colors corresponding to faster decays. There are three fixed points. Any point that is equidistant from $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ is updated to $\mathbf{0}$ in one step and stays there thereafter. Points that are closer to $\boldsymbol{\mu}$ are pushed towards $\boldsymbol{\mu}$, while points that are closer to $-\boldsymbol{\mu}$ are pushed towards $-\boldsymbol{\mu}$.

[2] In the multi-dimensional setting, this would corrrespond to a very large magnitude $\boldsymbol{\lambda}^{(0)}$ chosen in a random direction.

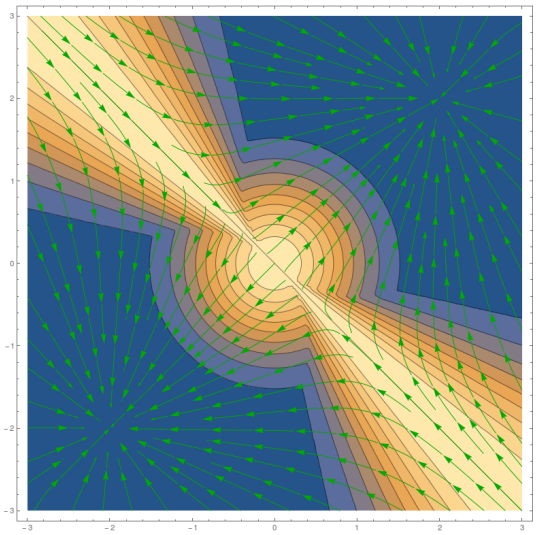

Figure 4: Illustration of the Speed of Convergence of EM in Multiple Dimensions as Implied by Theorem 2.

Remark 1 (General Speed of Convergence)**.** *The analysis given above for $SNR = 1$, generalizes to arbitrary SNRs, at the cost of a factor of $O(1/SNR^2)$ in the number of iterations. It also generalizes to obtain an arbitrary approximation ϵ, at a cost of a factor of $O(\log 1/\epsilon)$. In multiple dimensions, we could run EM from a random initialization. The number of iterations for an approximation of ϵ in Mahalanobis distance would depend on the angle of the initial iterate with $\boldsymbol{\mu}$. Under a random initialization, the cosine of that angle is expected to be $\Theta(1/\sqrt{d})$, resulting in a total $O(d/SNR^2 \cdot \log(1/\epsilon))$ number of iterations. We show that we can boostrap EM to obtain a better initialization, starting from a random one, improving the angle to $\Omega(1)$, after $O(\log d/SNR^2)$ iterations. With a constant angle, EM takes $O(1/SNR^2 \cdot \log(1/\epsilon))$ iterations to give ϵ error, as in the single dimension, overall improving exponentially the dependence on d. We describe our bootstrapping operation in the context of our analysis of the finite sample EM in Section 6.3.*

6 Sample Based Model

The main goal of this section is to prove convergence guarantees for the EM algorithm, when we have a finite sample. Similarly to the Section 4 we willquantify our approximation guarantees using the Mahalanobis distance $\|\cdot\|_\Sigma$ between two vectors with respect to matrix Σ, which we remind is defined as follows:

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\Sigma = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{y})}.$$

Also for the simplicity of the notation, it is useful to define the *Mahalanobis inner product* between two vectors with respect to matric Σ as follows:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_\Sigma = \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{y}.$$

Towards our goal, we encounter two challenges.

The first is that we cannot assume that we exactly know the mean of the mixture distribution $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Our only access to this mean is via samples. We therefore use samples to estimate it. Then, we translate the origin to our estimate, and write the EM iteration for finding the mean of one of the two mixture components with respect to this origin. Given the error incurred in the approximation of $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, we propose to stabilize the sample-based EM iteration by including

in the sample for each sampled point $\boldsymbol{x}_i$ its symmetric point $-\boldsymbol{x}_i$. This is the sample based version that we analyze, although our analysis goes through without this stabilization.

The other challenge has to do with the speed of convergence, as discussed in Remark 1. Recall that the convergence of EM is a function of the angle $\langle \hat{\boldsymbol{\lambda}}^{(0)}, \hat{\boldsymbol{\mu}} \rangle_\Sigma$, where $\hat{\boldsymbol{\lambda}}^{(0)}$ is the unit vector in the direction of $\boldsymbol{\lambda}^{(0)}$. In high dimensions choosing a random starting point $\boldsymbol{\lambda}^{(0)}$ leads to an inner product that has value approximately $1/\sqrt{d}$. Thus the first step is to find a starting point $\boldsymbol{\lambda}^{(0)}$ such that the mean vector $\boldsymbol{\mu}$ (after the translation by the estimation of $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$) has enough projection in the direction of $\boldsymbol{\lambda}^{(0)}$. How can we do this? We actually bootstrap the EM algorithm to get such a good initialization. We show that, if we run EM starting from a random vector with small ℓ_2 norm, then the EM algorithm will output a vector $\boldsymbol{\lambda}$ with $\langle \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}} \rangle_\Sigma \geq 1/2$ after $O(\log d/SNR^2)$ iterations, where SNR is defined as $SNR = \|\boldsymbol{\mu}\|_\Sigma$.

At this point we multiply $\boldsymbol{\lambda}$ with a large positive constant M and we continue the EM iteration with $\boldsymbol{\lambda}^{(0)} = M\boldsymbol{\lambda}$ using fresh samples for any iteration from now on. This stage needs only logarithmic number of steps with respect to $1/\varepsilon$, d and polynomial in $1/SNR$. At each of these steps we prove that the sample based iteration is very well concentrated around its expectation and thus after few steps and $\tilde{O}((d/\varepsilon^2)\mathrm{poly}(1/SNR)\log 1/\eta)$ samples we will find an estimation $\boldsymbol{\lambda}$ such that

$$\mathbb{P}\left(\|\boldsymbol{\lambda} - \boldsymbol{\mu}\|_\Sigma \geq \varepsilon\right) \leq \eta.$$

Our goal is to prove that $\|\boldsymbol{\lambda} - \boldsymbol{\mu}\|_\Sigma \leq \varepsilon$ holds with *high probability*, which means $\eta = \mathrm{poly}\left(\frac{\varepsilon^2}{d}\right)$. Also for any vector $\boldsymbol{v} \in \mathbb{R}^d$ we use $\hat{\boldsymbol{v}}$ to refer to the unit vector in the direction of $\boldsymbol{v}$. We present our results in the following order

1. **Centering:** Find an estimation $\boldsymbol{c}$ of $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ using $\tilde{O}((d/\varepsilon^2)) \cdot \log 1/\eta$ samples such that

$$\mathbb{P}\left(\left\|\boldsymbol{c} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right\|_\Sigma \geq \varepsilon\right) \leq \eta/3.$$

   We use $\boldsymbol{\delta}$ to refer to the error in our estimation $\boldsymbol{c} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}$.

2. **Initialization:** We bootstrap EM to find a good initialization. In particular, starting from a randomly chosen vector, we run EM for $O(\log d/SNR^2)$ iterations using $\tilde{O}((d/\varepsilon^2)) \cdot \log 1/\eta$ samples to get a unit vector $\hat{\boldsymbol{\lambda}}$ such that

$$\mathbb{P}\left(\langle \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}} \rangle_\Sigma \geq 1/2\right) \leq \eta/3.$$

3. **Main Execution:** Setting $\boldsymbol{\lambda}^{(0)} = M\hat{\boldsymbol{\lambda}}$ for some large constant M and running EM for $t = O((1/SNR^2)\log(1/\varepsilon))$ iterations, using $\tilde{O}((d/\varepsilon^2 SNR^4)) \cdot \log 1/\eta$ fresh samples at each iteration, we get a vector $\boldsymbol{\lambda}^{(t)}$ such that

$$\mathbb{P}\left(\left\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\mu}_1\right\|_\Sigma \geq \varepsilon\right) \leq \eta/3.$$

Combining the above steps all together we get our main theorem for this section.

Theorem 3. *If $\varepsilon \leq SNR$ and using $\tilde{O}(d/\varepsilon^2) \cdot \mathrm{poly}(1/SNR) \cdot \log 1/\eta$ samples we get an estimation $\boldsymbol{\lambda}^{(t)}$ such that there is a constant η that satisfies*

$$\mathbb{P}\left[\left\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\mu}_1\right\|_\Sigma \geq \varepsilon\right] \leq \eta.$$

We start now proving the lemmas for each of the steps described above.

6.1 Centering

Lemma 5. *For any $\eta < 1$, using $\tilde{O}(d/\varepsilon^2 \log(1/\eta))$ samples there exists an estimator $\boldsymbol{c}$ of the mean $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ such that if $\boldsymbol{\delta} = \boldsymbol{c} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ then*

$$\mathbb{P}\left[\|\boldsymbol{\delta}\|_\Sigma \geq \varepsilon\right] \leq \eta$$

Proof. We start by making the transformation $x \mapsto \Sigma^{-1/2}x$ to the space, so that the covariance matrix is the identity I. Finally we will take back this transformation and the Euclidean norm becomes the corresponding Mahalanobis.

We will get an estimate $\boldsymbol{c}$ of the mean $(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2$ by drawing $O(d/\varepsilon^2)$ samples from the mixture and working in each axis direction separately. We will compute the estimate $\hat{c}_i$ in axis direction $\boldsymbol{e}_i$ as the average of the first and third quartile of the empirical distribution given by the samples. It suffices to show that with high probability every quartile is at most $\frac{\varepsilon}{\sqrt{d}}$ away from the true quartile of the distribution $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2} \cdot \boldsymbol{e}_i$.

Let $p_{-\mu,\mu}$ be the mixture of Gaussian distributions obtained by centering $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2} \cdot \boldsymbol{e}_i$ around 0.

The cumulative distribution F of $p_{-\mu,\mu}$ is given by $F(x) = \frac{1}{2}\Phi\left(x+\mu\right) + \frac{1}{2}\Phi\left(x-\mu\right)$. By the DKW inequality [DL12] with n samples we have that the empirical distribution F_n satisfies:

$$\mathbb{P}\Big(\sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right| > \varepsilon/\sqrt{d}\Big) \leq 2e^{-2n\varepsilon^2/d}$$

In particular, for x such that $F_n(x) = 1/4$, with high probability $F(x) = 1/4 \pm \varepsilon/\sqrt{d}$. Moreover, for x^* such that $F(x^*) = 1/4$, we have that $|x - x^*| \leq \frac{\varepsilon/\sqrt{d}}{\min_{\xi\in[x,x^*]} F'(\xi)}$ by the mean value theorem.

Since for $\xi \in [x, x^*]$, $F(\xi) \in [1/4 - \varepsilon/\sqrt{d}, 1/4 + \varepsilon/\sqrt{d}]$, this implies that $\Phi\left(\xi+\mu\right) \in [1/4 - \varepsilon/\sqrt{d}, 1/2 + 2\varepsilon/\sqrt{d}]$ and thus $F'(\xi) \geq \frac{1}{2}\Phi'\left(\xi+\mu\right)$. But $\Phi'(z) \geq \frac{1}{5}$ when $\Phi(z) \in [1/5, 4/5]$ which shows that $|x - x^*| = O(\varepsilon/\sqrt{d})$. Similarly, this holds for the 3rd quartile as well and thus the same bound holds for the mean as well of the two quartiles with probability $1 - 2e^{-2n\varepsilon^2/d}$. Setting $n = O(d \log d/\varepsilon^2)$, we get that the bound is violated with probability $O(1/d)$. Taking a union bound for all d axis directions, we get that for all i, $|\boldsymbol{\delta}_i| = O(\varepsilon/\sqrt{d})$ with constant probability. This implies that $\|\boldsymbol{\delta}\| \leq \varepsilon$ with constant probability. By using a factor of $\log 1/\eta$ more samples it is easy to see that the error probability reduces to η. □

6.2 Sample Based EM Iteration

Using the estimation of the center $\boldsymbol{c}$ that we found in the previous section, we translate all our data and parameters so that $\boldsymbol{c} \mapsto \boldsymbol{0}$. After this centering, the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ become $\boldsymbol{\mu} + \boldsymbol{\delta}$, $-\boldsymbol{\mu} + \boldsymbol{\delta}$ and the covariance matrix remains the same.

We will use $\tilde{\boldsymbol{\lambda}}^{(t)}$ to refer to the estimation of $\boldsymbol{\mu}$ after t steps of finite sample stabilized EM Iteration. By stabilized we mean that for any sample $\boldsymbol{x}_i$ that we get, we include in our data set the vector $-\boldsymbol{x}_i$ too. This way one EM iteration is simpler and easier to analyze but the results hold even without this stabilization. Following exactly the same steps as in Section 2 we can see that

$$\tilde{\boldsymbol{\lambda}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tanh\left(\langle \tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_\Sigma\right)\boldsymbol{x}_i. \tag{6.1}$$

6.3 Initialization of EM

We rewrite the sample based EM iteration is the following form

$$\tilde{\boldsymbol{\lambda}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tanh\left(\langle \tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_\Sigma\right)\boldsymbol{x}_i = \sum_{i=1}^{n} \tanh\left(\tilde{\lambda}^{(t)}\langle \hat{\tilde{\boldsymbol{\lambda}}}^{(t)}, \boldsymbol{x}_i\rangle_\Sigma\right)\boldsymbol{x}_i$$

where $\tilde{\lambda}^{(t)} = \left\|\tilde{\boldsymbol{\lambda}}^{(t)}\right\|_{\Sigma}$ and $\hat{\tilde{\boldsymbol{\lambda}}}^{(t)} = \frac{\tilde{\boldsymbol{\lambda}}^{(t)}}{\|\tilde{\boldsymbol{\lambda}}^{(t)}\|_{\Sigma}}$ is the unit vector in the direction of $\tilde{\boldsymbol{\lambda}}^{(t)}$.

The basic idea of bootstraping that we describe in this section, is that if $\tilde{\lambda}^{(t)}$ is very small then the tanh function is very close to be linear. This linear approximation of tanh gives the following approximate form of the EM update

$$\tilde{\boldsymbol{\lambda}}^{(t+1)} \approx \frac{1}{n}\sum_{i=1}^{n}\langle\tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_{\Sigma}\boldsymbol{x}_i = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\Sigma^{-1}\tilde{\boldsymbol{\lambda}}^{(t)} = \hat{\Sigma}_p\Sigma^{-1}\tilde{\boldsymbol{\lambda}}^{(t)} \tag{6.2}$$

where $\hat{\Sigma}_p$ is the empirical covariance matrix of the mixture $p_{\boldsymbol{\mu}+\boldsymbol{\delta},-\boldsymbol{\mu}+\boldsymbol{\delta}}$. Now the intuition suggests that the direction of the maximum eigenvector of the matrix $\hat{\Sigma}_p\Sigma^{-1}$ is a direction that is spanned by $\boldsymbol{\mu}+\boldsymbol{\delta}$, $-\boldsymbol{\mu}+\boldsymbol{\delta}$ and hence a direction with small angle to the direction of $\boldsymbol{\mu}$. Also observe that this approximate EM iteration is actually an iteration of the power method for the matrix $\hat{\Sigma}_p\Sigma^{-1}$! Because of the efficiency of the power method, we expect that after a few steps this iteration will find a direction $\hat{\boldsymbol{\lambda}}$ such that the inner product $\langle\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\lambda}}\rangle$ is large enough. This direction is a good initialization for the EM algorithm as we will see in the next section. We now formaly demonstrate the intuition we described for the approximate EM update.

We start by bounding the error we introduce by replacing the tanh function with its linear approximation. It is very easy to see that $\left|\tanh''(x)\right| \leq 1$. Now by Taylor expansion of tanh around 0 we get that

$$|\tanh(x) - x| \leq \max_{\xi}\left(\tanh''(\xi)\right)\frac{x^2}{2} \leq \frac{x^2}{2}.$$

We want $\tilde{\lambda}^{(t)}$ to be small enough such that the linear approximation of tanh is a good approximation. For this reason we pick for $\epsilon < 1$

$$\tilde{\lambda}^{(0)} = \sqrt{\frac{2}{\mathcal{S}}}\cdot\epsilon \qquad \text{where} \qquad \mathcal{S} = \sum_{i=1}^{n}\|\boldsymbol{x}_i\|_{\Sigma}^3\,. \tag{6.3}$$

Also we choose the direction $\hat{\tilde{\boldsymbol{\lambda}}}^{(0)}$ uniformly from the unit sphere. Also let's assume that after every step of EM we normalize the $\tilde{\boldsymbol{\lambda}}^{(t)}$ to ensure that $\tilde{\lambda}^{(t)}$ satisfies (6.3). This renormalization is non-necessary and it could be easily dropped by choosing $\tilde{\lambda}^{(0)}$ to be so small that after $O(\log d/SNR^2)$ steps (6.3) is still satisfied. Given (6.3) we have that

$$\begin{aligned}
\left\|\tilde{\boldsymbol{\lambda}}^{(t+1)} - \frac{1}{n}\sum_{i=1}^{n}\langle\tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_{\Sigma}\boldsymbol{x}_i\right\|_{\Sigma} &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\tanh\left(\tilde{\lambda}^{(t)}\langle\hat{\tilde{\boldsymbol{\lambda}}}^{(t)}, \boldsymbol{x}_i\rangle_{\Sigma}\right) - \langle\tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_{\Sigma}\right|\|\boldsymbol{x}_i\|_{\Sigma} \\
&\leq \frac{1}{2n}\sum_{i=1}^{n}\langle\tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_{\Sigma}^2\|\boldsymbol{x}_i\|_{\Sigma} \\
&\leq \frac{1}{2n}\left(\tilde{\lambda}^{(t)}\right)^2\sum_{i=1}^{n}\|\boldsymbol{x}_i\|_{\Sigma}^3 \implies
\end{aligned}$$

$$\left\|\tilde{\boldsymbol{\lambda}}^{(t+1)} - \frac{1}{n}\sum_{i=1}^{n}\langle\tilde{\boldsymbol{\lambda}}^{(t)}, \boldsymbol{x}_i\rangle_{\Sigma}\boldsymbol{x}_i\right\|_{\Sigma} \leq \frac{1}{n}\epsilon^2 \tag{6.4}$$

At this point the calculations become much easier if we assume that we have already done the mapping $x \mapsto \Sigma^{-1/2}x$ and when we are done we will take the inverse mapping and get the result in Mahalanobis distance. Equation (6.4) suggests that it suffices to analyze the convergence of the power method given by the following equation.

$$\tilde{\boldsymbol{\tau}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\langle \tilde{\boldsymbol{\tau}}^{(t)}, \boldsymbol{x}_i\rangle \boldsymbol{x}_i \overset{(6.2)}{=} \hat{\Sigma}_p \tilde{\boldsymbol{\tau}}^{(t)} \tag{6.5}$$

where $\hat{\Sigma}_p$ is the empirical covariance matrix of the mixture $p_{\boldsymbol{\mu}+\boldsymbol{\delta},-\boldsymbol{\mu}+\boldsymbol{\delta}}$. Before analyzing (6.5) lets see what happens if instead of the empirical covariance $\hat{\Sigma}_p$ we had the actual covariance Σ_p of the mixture distribution $p_{\boldsymbol{\mu}+\boldsymbol{\delta},-\boldsymbol{\mu}+\boldsymbol{\delta}}$. Then the iteration would be $\boldsymbol{\tau}^{(t+1)} = \Sigma_p \boldsymbol{\tau}^{(t)}$. The covariance matrix of the mixture is

$$\Sigma_p = I + \boldsymbol{\mu}\boldsymbol{\mu}^T. \tag{6.6}$$

Therefore the principal eigenvector of Σ_p is $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}$ with eigenvalue $1+\mu^2$, where $\mu = \|\boldsymbol{\mu}\|$. All the other eigenvectors have eigenvalue 1 and therefore the ratio of the largest to the lowest eigenvalue is $\rho = 1+\mu^2$.

To get the corresponding properties of $\hat{\Sigma}_p$ we observe that each $\boldsymbol{x}_i$ can be written as

$$\boldsymbol{x}_i = \boldsymbol{y}_i + z_i\boldsymbol{\delta} + \boldsymbol{\mu}$$

where $\boldsymbol{y}_i$ is distributed as $\mathcal{N}(0, I)$ and z_i is a Rademacher indicator variable that shows whether $\boldsymbol{x}_i$ is coming from the distribution $\mathcal{N}(\boldsymbol{\mu}+\boldsymbol{\delta}, I)$ or $\mathcal{N}(\boldsymbol{\mu}-\boldsymbol{\delta}, I)$. We notice that $\boldsymbol{y}_i$ and z_i are independent. We can now rewrite $\hat{\Sigma}_p$ in terms of $\boldsymbol{y}_i$ and z_i.

$$\hat{\Sigma}_p = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i\boldsymbol{y}_i^T + \frac{1}{n}\sum_{i=1}^{n} z_i\left(\boldsymbol{y}_i\boldsymbol{\delta}^T + \boldsymbol{\delta}\boldsymbol{y}_i^T\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{y}_i^T\right) + \frac{1}{n}\sum_{i=1}^{n}\left(z_i\boldsymbol{\delta}+\boldsymbol{\mu}\right)\left(z_i\boldsymbol{\delta}+\boldsymbol{\mu}\right)^T$$

For simplicity we define

$$\begin{aligned}
\hat{\Sigma}_1 &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i\boldsymbol{y}_i^T \\
\hat{\Sigma}_2 &= \frac{1}{n}\sum_{i=1}^{n} z_i\left(\boldsymbol{y}_i\boldsymbol{\delta}^T + \boldsymbol{\delta}\boldsymbol{y}_i^T\right) \\
\hat{\Sigma}_3 &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{y}_i^T\right) \\
\hat{\Sigma}_4 &= \frac{1}{n}\sum_{i=1}^{n}\left(z_i\boldsymbol{\delta}+\boldsymbol{\mu}\right)\left(z_i\boldsymbol{\delta}+\boldsymbol{\mu}\right)^T
\end{aligned}$$

It is easy to see that using $n = \tilde{O}(d/\varepsilon^2)$ samples, $\hat{\Sigma}_1$ satisfies the following lemma.

Lemma 6. *For $n = \tilde{O}(d/\varepsilon^2)$ let $\hat{\Sigma}_1 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i\boldsymbol{y}_i^T$, where $\boldsymbol{y}_i$ is drawn from $\mathcal{N}(0, I)$. For any direction $\hat{\boldsymbol{v}}$ in $\mathbb{R}^d$. Then*

$$\mathbb{P}\left(\left|\hat{\boldsymbol{v}}^T\hat{\Sigma}_1\hat{\boldsymbol{v}} - 1\right| > \varepsilon^2\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right).$$

We describe now a sketch of the proof of Lemma 6, using the following Lemma 25 from [DKT15].

Lemma 25 of Daskalakis et al. 2016. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ be two symmetric, positive semi-definite matrices, and let $(\lambda_1, \boldsymbol{v}_1), \ldots, (\lambda_d, \boldsymbol{v}_d)$ be the eigenvalue-eigenvector pairs of Σ. Suppose that*

- *For all $i \in \{1, \ldots d\}$, $\left|\left(\frac{\boldsymbol{v}_i}{\sqrt{\lambda_i}}\right)^T \left(\Sigma - \hat{\Sigma}\right) \left(\frac{\boldsymbol{v}_i}{\sqrt{\lambda_i}}\right)\right| \leq \epsilon$,*
- *For all $i, j \in \{1, \ldots d\}$, $\left|\left(\frac{\boldsymbol{v}_i}{\sqrt{\lambda_i}} + \frac{\boldsymbol{v}_j}{\sqrt{\lambda_j}}\right)^T \left(\Sigma - \hat{\Sigma}\right) \left(\frac{\boldsymbol{v}_i}{\sqrt{\lambda_i}} + \frac{\boldsymbol{v}_j}{\sqrt{\lambda_j}}\right)\right| \leq 4\epsilon$.*

Then for all $\boldsymbol{z} \in \mathbb{R}^d$, $\left|\boldsymbol{z}^T \left(\Sigma - \hat{\Sigma}\right) \boldsymbol{z}\right| \leq 3d\epsilon \boldsymbol{z}^T \Sigma \boldsymbol{z}$.

The projection of $\boldsymbol{y}_i$ in each combination of two eigenvectors of Σ is an one dimensional gaussian with variance 1. Therefore using $\tilde{O}(d/\varepsilon^2)$ samples we can estimate $\hat{\Sigma}_1$ in this direction with error at most $\varepsilon/\sqrt{d}$ with high probability. Therefore by a union bound on these fixed direction that depend only to Σ_1 we can satisfy the conditions of Lemma 25 of [DKT15]. Then from the implication of Lemma 25, Lemma 6 follows.

We continue with $\hat{\Sigma_3}$.

Lemma 7. *For $n = \tilde{O}(d/\varepsilon^2)$ let $\hat{\Sigma}_3 = \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{y}_i\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{y}_i^T\right)$, where $\boldsymbol{y}_i$ is drawn from $\mathcal{N}(0, I)$, z_i is a uniform Rademacher random variable and. For any direction $\hat{\boldsymbol{v}}$ in $\mathbb{R}^d$. Then*

$$\mathbb{P}\left(\left|\hat{\boldsymbol{v}}^T\hat{\Sigma}_3\hat{\boldsymbol{v}}\right| > 2\varepsilon\left|\hat{\boldsymbol{v}}^T\boldsymbol{\mu}\right|\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right)$$

Proof. Let $\hat{\boldsymbol{v}}$ an arbitrary direction in $\mathbb{R}^d$. Then we have that

$$\left|\hat{\boldsymbol{v}}^T\hat{\Sigma}_3\hat{\boldsymbol{v}}\right| = \frac{2}{n}\left|\sum_{i=1}^n (\hat{\boldsymbol{v}}^T\boldsymbol{y}_i)(\boldsymbol{\mu}^T\hat{\boldsymbol{v}})\right| \leq 2\left|\hat{\boldsymbol{v}}^T\boldsymbol{\mu}\right|\left|\hat{\boldsymbol{v}}^T\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\right)\right| \leq 2\left|\hat{\boldsymbol{v}}^T\boldsymbol{\mu}\right|\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\right\|$$

Now we consider the quantity $\boldsymbol{e}_j^T\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\right)$, where $\boldsymbol{e}_j$ is the unit jth vector. This is equivalent with having $\frac{1}{n}\sum_{i=1}^n y_i$ where y_i is drawn from $\mathcal{N}(0, 1)$. So we have that

$$\mathbb{P}\left[\left|\boldsymbol{e}_j^T\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\right)\right| \geq \frac{\varepsilon}{\sqrt{d}}\right] \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right)$$

Now by doing a union bound over all $\boldsymbol{e}_j$ we get that

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\right\| \geq \varepsilon\right] \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right)$$

Using these we can conclude that

$$\mathbb{P}\left(\left|\hat{\boldsymbol{v}}^T\hat{\Sigma}_3\hat{\boldsymbol{v}}\right| > 2\varepsilon\left|\hat{\boldsymbol{v}}^T\boldsymbol{\mu}\right|\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right)$$

□

For Σ_2, it is easy to observe that in $\hat{\Sigma}_2 = \frac{1}{n}\sum_{i=1}^n z_i\left(\boldsymbol{y}_i\boldsymbol{\delta}^T + \boldsymbol{\delta}\boldsymbol{y}_i^T\right)$, z_i are independent from $\boldsymbol{y}_i$ and so the product $z_i\boldsymbol{y}_i$ is a sample from standard multinormal distribution $\mathcal{N}(0, I)$. Therefore we can substitute $z_i\boldsymbol{y}_i$ with just $\boldsymbol{y}_i$. Now using exactly the same analysis as in Lemma 7 we can prove the following lemma.

Lemma 8. *For $n = \tilde{O}(d/\varepsilon^2)$ let $\hat{\Sigma}_2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{i=1}^n z_i \left(\boldsymbol{y}_i\boldsymbol{\delta}^T + \boldsymbol{\delta}\boldsymbol{y}_i^T\right)$, where $\boldsymbol{y}_i$ is drawn from $\mathcal{N}(0, I)$, z_i is a uniform Rademacher random variable and. For any direction $\hat{\boldsymbol{v}}$ in $\mathbb{R}^d$ and any $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\| \leq \varepsilon$. Then*

$$\mathbb{P}\left(\left|\hat{\boldsymbol{v}}^T\hat{\Sigma}_2\hat{\boldsymbol{v}}\right| > 2\varepsilon^2\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right)$$

For $\hat{\Sigma}_4$ we do straight forward calculations. For simplicity we use $\mu = \|\boldsymbol{\mu}\|_\Sigma$. Let $\hat{\boldsymbol{v}}$ be an arbitrary direction in $\mathbb{R}^d$.

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_4\hat{\boldsymbol{v}} = \frac{1}{n}\sum_{i=1}^n \left(z_i\boldsymbol{\delta}^T\hat{\boldsymbol{v}} + \boldsymbol{\mu}^T\hat{\boldsymbol{v}}\right)^2 \implies$$

$$\left(\left|\boldsymbol{\mu}^T\hat{\boldsymbol{v}}\right| - \varepsilon\right)^2 \leq \hat{\boldsymbol{v}}^T\hat{\Sigma}_4\hat{\boldsymbol{v}} \leq \left(\left|\boldsymbol{\mu}^T\hat{\boldsymbol{v}}\right| + \varepsilon\right)^2 \tag{6.7}$$

Now we calculate the variance of the samples in the direction of $\boldsymbol{\mu}$. We have

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}_p\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^T\hat{\Sigma}_1\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T\hat{\Sigma}_2\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T\hat{\Sigma}_3\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T\hat{\Sigma}_4\hat{\boldsymbol{\mu}}$$

Using Lemmas 6, 8, 7 and (6.7) we have that with probability at least $1 - \text{poly}(\varepsilon^2/d)$

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}_1\hat{\boldsymbol{\mu}} \geq 1 - \varepsilon^2$$

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}_2\hat{\boldsymbol{\mu}} \geq -2\varepsilon^2$$

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}_3\hat{\boldsymbol{\mu}} \geq -2\varepsilon\mu$$

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}_4\hat{\boldsymbol{\mu}} \geq (\mu - \varepsilon)^2$$

and so

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}_p\hat{\boldsymbol{\mu}} \geq 1 - 3\varepsilon^2 - 2\varepsilon\mu + (\mu - \varepsilon)^2.$$

Now let any other direction $\hat{\boldsymbol{v}}$ with $\hat{\boldsymbol{v}}^T\boldsymbol{\mu} = a\mu$, with $a \geq 0$. Again using Lemmas 6, 8, 7 and (6.7) we have that with probability at least $1 - \text{poly}(\varepsilon^2/d)$

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_1\hat{\boldsymbol{v}} \leq 1 + \varepsilon^2$$

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_2\hat{\boldsymbol{v}} \leq 2\varepsilon^2$$

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_3\hat{\boldsymbol{v}} \leq 2a\varepsilon\mu$$

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_4\hat{\boldsymbol{v}} \leq (a\mu + \varepsilon)^2$$

and so

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_p\hat{\boldsymbol{v}} \leq 1 + 3\varepsilon^2 + 2a\varepsilon\mu + (a\mu + \varepsilon)^2.$$

The principal eigenvector $\hat{\boldsymbol{v}}$ of $\hat{\Sigma}_p$ has to satisfy

$$\hat{\boldsymbol{v}}^T\hat{\Sigma}_p\hat{\boldsymbol{v}} \geq \hat{\boldsymbol{\mu}}^T\hat{\Sigma}_p\hat{\boldsymbol{\mu}} \implies$$

$$6\varepsilon^2 + 4a\varepsilon\mu + (a^2 - 1)\mu^2 \geq 0$$

Now using an ε such that $\varepsilon \leq \mu/10$ we have that the above implies $a \geq 3/4$. This proves the following Proposition that we use in the analysis of the initialization step.

Proposition 1. *Let $\hat{\Sigma}_p$ be the empirical covariance matrix computed from $n = \tilde{O}(d/\varepsilon^2)$ samples from the distribution $p_{\boldsymbol{\mu}+\boldsymbol{\delta},\boldsymbol{\mu}-\boldsymbol{\delta}}$. Then the principal eigenvector $\hat{\boldsymbol{v}}$ of $\hat{\Sigma}_p$ satisfies*

$$\mathbb{P}\left(\langle \hat{\boldsymbol{v}}, \hat{\boldsymbol{\mu}} \rangle < \frac{3}{4}\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right).$$

The last thing we need to prove to complete the analysis of the initialization is the gap between the first and the second eigenvalue of $\hat{\Sigma}_p$. We want to use this gap for the analysis of the convergence rate of the power method iteration that takes place in the first steps. Given Proposition refprop:principalVectorSp, we have that any direction $\hat{v}$ except from the principal one, has $\langle \hat{\boldsymbol{v}}, \hat{\boldsymbol{\mu}} \rangle \leq \frac{1}{4}$ with high probability. Let $\hat{\boldsymbol{v}}'$ be the eigenvector that corresponds to the second maximum eigenvalue of $\hat{\Sigma}_p$. Let $a = \hat{\boldsymbol{v}}'^T \boldsymbol{\mu}$, using Lemmas 6, 8, 7 and (6.7) we have that

$$\hat{\boldsymbol{v}}'^T \hat{\Sigma}_1 \hat{\boldsymbol{v}}' \leq 1 + \varepsilon^2$$

$$\hat{\boldsymbol{v}}'^T \hat{\Sigma}_2 \hat{\boldsymbol{v}}' \leq 2\varepsilon^2$$

$$\hat{\boldsymbol{v}}'^T \hat{\Sigma}_3 \hat{\boldsymbol{v}}' \leq 2a\varepsilon\mu \leq \frac{\varepsilon\mu}{2}$$

$$\hat{\boldsymbol{v}}'^T \hat{\Sigma}_4 \hat{\boldsymbol{v}}' \leq (a\mu + \varepsilon)^2 \leq \left(\frac{\mu}{4} + \varepsilon\right)^2 \implies$$

$$\hat{\boldsymbol{v}}'^T \hat{\Sigma}_p \hat{\boldsymbol{v}}' \leq 1 + 3\varepsilon^2 + \frac{1}{2}\varepsilon\mu + \left(\frac{1}{4}\mu + \varepsilon\right)^2.$$

On the other hand based on the fact that $\hat{\boldsymbol{v}}^T \hat{\Sigma}_p \hat{\boldsymbol{v}} \geq \hat{\boldsymbol{\mu}}^T \hat{\Sigma}_p \hat{\boldsymbol{\mu}}$ we conclude that

$$\hat{\boldsymbol{v}}^T \hat{\Sigma}_p \hat{\boldsymbol{v}} \geq 1 - 3\varepsilon^2 - 2\varepsilon\mu + (\mu - \varepsilon)^2.$$

Using also the hypothesis that $0 \leq \varepsilon \leq \mu/10$ we get that

$$\frac{\hat{\boldsymbol{v}}^T \hat{\Sigma}_p \hat{\boldsymbol{v}}}{\hat{\boldsymbol{v}}'^T \hat{\Sigma}_p \hat{\boldsymbol{v}}'} \geq \frac{1 - 3\varepsilon^2 - 2\varepsilon\mu + (\mu - \varepsilon)^2}{1 + 3\varepsilon^2 + \frac{1}{2}\varepsilon\mu + \left(\frac{1}{4}\mu + \varepsilon\right)^2} \geq \frac{1 + \frac{232}{400}\mu^2}{1 + \frac{101}{400}\mu^2} \geq \min\left\{1 + \frac{1}{4}\mu^2, 2\right\}.$$

Hence we get the following proposition.

Proposition 2. *Let $\hat{\Sigma}_p$ be the empirical covariance matrix computed from $n = \tilde{O}(d/\varepsilon^2)$ samples from the distribution $p_{\boldsymbol{\mu}+\boldsymbol{\delta},\boldsymbol{\mu}-\boldsymbol{\delta}}$. Let also $\hat{\rho}$ be the ratio of the magnitude of the first two eigenvalues of $\hat{\Sigma}_p$ then*

$$\mathbb{P}\left(\hat{\rho} < \min\left\{1 + \frac{1}{4}\mu^2, 2\right\}\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right).$$

It is well known and easy to prove that if we choose a random vector $\boldsymbol{\lambda}^{(0)}$ uniformly from the half unit sphere, defined by $\hat{\boldsymbol{\mu}}$, then we will have that $\langle \boldsymbol{\lambda}^{(0)}, \hat{\boldsymbol{\mu}} \rangle \geq 1/\sqrt{d}$ with high probability.

By standard analysis of the power method, Chapter 21.3 [SSBD14], we know that the number of iterations we need to get within a constant angle from the principal eigenvector, starting from angle $1/\sqrt{d}$ is $O(\log d/\log\hat{\rho})$ where $\hat{\rho}$ is, as we have said, the ratio of the first two eigenvalues of $\hat{\Sigma}_p$. Therefore after $O(\log d/\log\hat{\rho}) = O(\log d/\min(\mu^2, 1))$ steps the iteration of $\tilde{\boldsymbol{\tau}}$ will find a vector $\boldsymbol{\tau}$ such that $\langle \hat{\tau}, \hat{\boldsymbol{\mu}} \rangle$ is at least $2/3$.

Now we are ready to analyze the performance of $\tilde{\boldsymbol{\lambda}}^{(t)}$ as be described in the beginning of the section. Applying (6.4) repeatedly at every iteration we get that after $t = O(\log d/\log\hat{\rho})$ iterations it holds that

$$\left\| \frac{1}{n}\tilde{\boldsymbol{\lambda}}^{(t)} - \tilde{\tau}^{(t)} \right\| \leq \frac{1}{n}\epsilon^2 \left(1 + \hat{\rho} + \hat{\rho}^2 + \cdots + \hat{\rho}^k\right) = \frac{1}{n}\epsilon^2 \frac{\hat{\rho}^{k+1} - 1}{\hat{\rho} - 1}.$$

But $k+1$ is $O(\log d/\log\hat{\rho})$ and therefore

$$\left\|\frac{1}{n}\tilde{\boldsymbol{\lambda}}^{(t)}-\tilde{\tau}^{(t)}\right\|\leq\frac{1}{n}\epsilon^2\frac{\mathrm{poly}(d)-1}{\hat{\rho}-1}.$$

Finally since $\hat{\rho}$ is $\min\left(1+\frac{\mu^2}{4},2\right)$ and also $\varepsilon\leq\mu/10$ we only need to set ϵ polynomially with respect to ε^2/d and we will get that after the first $O(\log d/\min(\mu^2,1))$ iterations it holds that

$$\left\|\frac{1}{n}\tilde{\boldsymbol{\lambda}}^{(t)}-\tilde{\tau}^{(t)}\right\|\leq\eta$$

for some $\eta\leq\varepsilon/6$. Which implies that

$$\langle\hat{\tilde{\boldsymbol{\lambda}}},\hat{\boldsymbol{\mu}}\rangle\geq\frac{2}{3}-\frac{\eta}{\mu}\geq\frac{1}{2}.$$

This proves the following lemma and completes the proof of the initialization.

Lemma 9. *Starting for a guess $\lambda^{(0)}$ such that $\lambda^{(0)}\leq\sqrt{\frac{1}{18\mathcal{S}}}\mathrm{poly}\left(\frac{\varepsilon^2}{d}\right)$, with $\mathcal{S}=\sum_{i=1}^n\|\boldsymbol{x}_i\|_\Sigma^3$ and after $O(\frac{\log d}{\varepsilon^2})$ iterations of EM we get a vector $\boldsymbol{\lambda}$ such that*

$$\langle\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\mu}}\rangle_\Sigma\geq\frac{1}{2}.$$

The probability of failure is at most $\mathrm{poly}\left(\frac{\varepsilon^2}{d}\right)$.

6.4 Finite-Sample EM Analysis

We initialize EM at $\boldsymbol{\lambda}^{(0)}=M\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is the point from Lemma 9, for some large constant M.[3] With this initialization, we run EM for $t=O((1/\mu^2)\log(d/\varepsilon))$ steps using $O((d/\varepsilon^2\mu^4)\log 1/\delta)$ samples at each step, where for ease of notation we have set $\mu=SNR\equiv\|\boldsymbol{\mu}\|_\Sigma$.

To study our sample-based EM iteration (6.1) we will relate its progress to an appropriate population EM iteration. Note that this iteration differs from the population EM iteration that we discussed in Section 2 and analyzed in Section 4. The reason is that we have incurred an error $\boldsymbol{\delta}$ in the estimation of the mean of the distribution in Section 6.1. With respect to our estimated mean centering, the true means of the two Gaussian components are $\boldsymbol{\mu}+\boldsymbol{\delta}$, $-\boldsymbol{\mu}+\boldsymbol{\delta}$ rather than $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$. Another source of discrepancy comes from the fact that we included for each point $\boldsymbol{x}_i$ in our sample its symmetric point $-\boldsymbol{x}_i$. This implies that each $\boldsymbol{x}_i$ is coming with probability 1/2 from the mixture $p_{\boldsymbol{\mu}+\boldsymbol{\delta},-\boldsymbol{\mu}-\boldsymbol{\delta}}$ and with probability 1/2 from the mixture $p_{\boldsymbol{\mu}-\boldsymbol{\delta},-\boldsymbol{\mu}+\boldsymbol{\delta}}$. Given this, using again the same operations as in Section 2, we have that the corresponding population iteration, denoted by $\boldsymbol{\lambda}^{(t)}$, is

$$\boldsymbol{\lambda}^{(t+1)}=\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}+\boldsymbol{\delta},\Sigma)}\left[\tanh\left(\langle\boldsymbol{\lambda}^{(t)},\boldsymbol{x}\rangle_\Sigma\right)\boldsymbol{x}\right]+\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}-\boldsymbol{\delta},\Sigma)}\left[\tanh\left(\langle\boldsymbol{\lambda}^{(t)},\boldsymbol{x}\rangle_\Sigma\right)\boldsymbol{x}\right]. \tag{6.8}$$

Our proof follows two steps illustrated in Figures 2 and **??**:

- **Step 1:** First, we relate the population EM iteration defined by (6.8) to the vanilla population EM iteration defined by (2.2);
- **Step 2:** Then, we related the population EM iteration defined by (6.8) to the sample-based iteration.

[3] It is easy to find such constant by getting a small number of samples and keeping the one that has maximum magnitude.

Step 1: To analyze the convergence of (6.8), we use Theorem 2 for every component of the mixture. More precisely, let $\boldsymbol{\lambda}_1^{(t)}$ and $\boldsymbol{\lambda}_2^{(t)}$ be

$$\begin{aligned}\boldsymbol{\lambda}_1^{(t+1)} &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}+\boldsymbol{\delta},\Sigma)}\left[\tanh(\langle\boldsymbol{\lambda}^{(t)},\boldsymbol{x}\rangle_\Sigma)\boldsymbol{x}\right]\\ \boldsymbol{\lambda}_2^{(t+1)} &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}-\boldsymbol{\delta},\Sigma)}\left[\tanh(\langle\boldsymbol{\lambda}^{(t)},\boldsymbol{x}\rangle_\Sigma)\boldsymbol{x}\right].\end{aligned}$$

We know from Theorem 2 that

$$\begin{aligned}\left\|\boldsymbol{\lambda}_1^{(t+1)}-\boldsymbol{\mu}-\boldsymbol{\delta}\right\|_\Sigma &\le \kappa_1^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}-\boldsymbol{\delta}\right\|_\Sigma\\ \left\|\boldsymbol{\lambda}_2^{(t+1)}-\boldsymbol{\mu}+\boldsymbol{\delta}\right\|_\Sigma &\le \kappa_2^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}+\boldsymbol{\delta}\right\|_\Sigma\end{aligned}$$

where

$$\begin{aligned}\kappa_1^{(t)} &= \exp\left(-\frac{\min\left\{\left\|\boldsymbol{\lambda}^{(t)}\right\|_\Sigma^2,\langle\boldsymbol{\mu}+\boldsymbol{\delta},\boldsymbol{\lambda}^{(t)}\rangle_\Sigma\right\}^2}{2\left\|\boldsymbol{\lambda}^{(t)}\right\|_\Sigma^2}\right)\\ \kappa_2^{(t)} &= \exp\left(-\frac{\min\left\{\left\|\boldsymbol{\lambda}^{(t)}\right\|_\Sigma^2,\langle\boldsymbol{\mu}-\boldsymbol{\delta},\boldsymbol{\lambda}^{(t)}\rangle_\Sigma\right\}^2}{2\left\|\boldsymbol{\lambda}^{(t)}\right\|_\Sigma^2}\right).\end{aligned}$$

But we have that

$$\kappa_1^{(t)},\kappa_2^{(t)} \le \kappa'^{(t)} \triangleq \exp\left(-\frac{\min\left\{\left\|\boldsymbol{\lambda}^{(t)}\right\|_\Sigma^2,\langle\boldsymbol{\mu},\boldsymbol{\lambda}^{(t)}\rangle_\Sigma\right\}^2}{2\left\|\boldsymbol{\lambda}^{(t)}\right\|_\Sigma^2}+\frac{\varepsilon^2}{2}\right)$$

which implies that

$$\left\|\boldsymbol{\lambda}_1^{(t+1)}-\boldsymbol{\mu}-\boldsymbol{\delta}\right\|_\Sigma \le \kappa'^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}-\boldsymbol{\delta}\right\|_\Sigma \tag{6.9}$$

$$\left\|\boldsymbol{\lambda}_2^{(t+1)}-\boldsymbol{\mu}+\boldsymbol{\delta}\right\|_\Sigma \le \kappa'^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}+\boldsymbol{\delta}\right\|_\Sigma. \tag{6.10}$$

We are ready now to bound the convergence of $\boldsymbol{\lambda}^{(t)}=(\boldsymbol{\lambda}_1^{(t)}+\boldsymbol{\lambda}_2^{(t)})/2$

$$\begin{aligned}\left\|\boldsymbol{\lambda}^{(t+1)}-\boldsymbol{\mu}\right\|_\Sigma &= \left\|\frac{\boldsymbol{\lambda}_1^{(t+1)}+\boldsymbol{\lambda}_2^{(t+1)}}{2}-\boldsymbol{\mu}+\frac{\boldsymbol{\delta}}{2}-\frac{\boldsymbol{\delta}}{2}\right\|_\Sigma\\ &\le \frac{1}{2}\left\|\boldsymbol{\lambda}_1^{(t+1)}-\boldsymbol{\mu}-\boldsymbol{\delta}\right\|_\Sigma+\frac{1}{2}\left\|\boldsymbol{\lambda}_2^{(t+1)}-\boldsymbol{\mu}+\boldsymbol{\delta}\right\|_\Sigma\\ &\overset{(6.9),(6.10)}{\le} \frac{1}{2}\kappa'^{(t)}\left(\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}-\boldsymbol{\delta}\right\|_\Sigma+\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}+\boldsymbol{\delta}\right\|_\Sigma\right)\\ &\le \kappa'^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}\right\|_\Sigma+\kappa'^{(t)}\left\|\boldsymbol{\delta}\right\|_\Sigma \implies\end{aligned}$$

$$\left\|\boldsymbol{\lambda}^{(t+1)}-\boldsymbol{\mu}\right\|_\Sigma \le \kappa'^{(t)}\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}\right\|_\Sigma+\kappa'^{(t)}\left\|\boldsymbol{\delta}\right\|_\Sigma \tag{6.11}$$

We can use the analysis of Section 4 to see that $\kappa'^{(t)} \leq \kappa'^{(0)}$. Therefore we also know that the population iteration satisfies:

$$\left\|\boldsymbol{\lambda}^{(t+1)} - \boldsymbol{\mu}\right\|_{\Sigma} \leq \kappa \left\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\mu}\right\|_{\Sigma} + \kappa \left\|\boldsymbol{\delta}\right\|_{\Sigma} \tag{6.12}$$

where for simplicity we let $\kappa = \kappa'^{(0)}$. Now we have $\boldsymbol{\lambda}^{(0)} = M\boldsymbol{\lambda}$. Also M is larger than μ and because of Lemma 9 $\langle \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}} \rangle_{\Sigma} \geq 1/2$ with high probability, where $\mu = \|\boldsymbol{\mu}\|_{\Sigma}$. Therefore

$$\kappa \leq \exp\left(-\frac{\mu^2}{4} + \frac{\varepsilon^2}{2}\right)$$

and since $\varepsilon \leq \mu/10$ we have that

$$\kappa \leq \exp\left(-\frac{\mu^2}{6}\right).$$

Also we have that $\|\boldsymbol{\delta}\|_{\Sigma} \leq \varepsilon$ and therefore (6.11) becomes

$$\left\|\boldsymbol{\lambda}^{(t+1)} - \boldsymbol{\mu}\right\|_{\Sigma} \leq e^{-\frac{\mu^2}{6}} \left\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\mu}\right\|_{\Sigma} + e^{-\frac{\mu^2}{6}} \varepsilon. \tag{6.13}$$

Step 2: Our next goal is to show that the sample based iteration (6.1) satisfies an equation similar to (6.13). We prove so by proving the concentration of $\tilde{\boldsymbol{\lambda}}^{(t)}$ around its mean $\boldsymbol{\lambda}^{(t)}$.

Lemma 10. *Let $\tilde{\boldsymbol{\lambda}}^{(t)} = \boldsymbol{\lambda}^{(t)}$ then if we use $n = \tilde{O}\left(\frac{d}{\varepsilon^2}\right)$ fresh samples at time step $t+1$ we have that*

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{\lambda}}^{(t+1)} - \boldsymbol{\lambda}\right\|_{\Sigma} > \varepsilon + \varepsilon \cdot \min\left\{1, \|\boldsymbol{\mu}\|_{\Sigma}\right\} \cdot \left\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\mu}\right\|_{\Sigma}\right) \leq \text{poly}\left(\frac{\varepsilon^2}{d}\right).$$

Proof. Once again we assume that $\Sigma = I$ and for simplicity we set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$, $\boldsymbol{\lambda}' = \boldsymbol{\lambda}^{(t+1)}$ and $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}^{(t)}$, $\tilde{\boldsymbol{\lambda}}' = \tilde{\boldsymbol{\lambda}}^{(t+1)}$. Also we assume that are working on a basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d\}$ such that all $\boldsymbol{v}_i$ for $i > 3$ are perpedicular to bot $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ and also $\boldsymbol{v}_2$ is perpedicular to $\boldsymbol{\lambda}$ and $\boldsymbol{v}_1$ parallel to $\boldsymbol{\lambda}$.

We first consider $i = 1$. In this case

$$\tilde{\lambda}_1' = \frac{1}{n}\sum_{i=1}^{n} \tanh(\langle \boldsymbol{\lambda}, \boldsymbol{x}_i \rangle)\langle \boldsymbol{x}_i, \boldsymbol{v}_1 \rangle,$$

$$\lambda_1' = \mathbb{E}[\tilde{\lambda}_1'] = \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\mu}+\boldsymbol{\delta}, \boldsymbol{\mu}-\boldsymbol{\delta}}}\left[\tanh(\langle \boldsymbol{\lambda}, \boldsymbol{x} \rangle)\langle \boldsymbol{x}, \boldsymbol{v}_1 \rangle\right].$$

Now we define $\mu_1 = \langle \boldsymbol{\mu}, \boldsymbol{v}_1 \rangle$, $\delta_1 = \langle \boldsymbol{\delta}, \boldsymbol{v}_1 \rangle$ and we have

$$\tilde{\lambda}_1' = \frac{1}{n}\sum_{i=1}^{n} \tanh(\lambda x_i) x_i,$$

$$\lambda_1' = \mathbb{E}[\tilde{\lambda}_1'] = \mathbb{E}_{x \sim p_{\mu_1+\delta_1, \mu_1-\delta_1}}\left[\tanh(\lambda x) x\right]$$

where x_i is distributed as $p_{\mu_1+\delta_1, \mu_1-\delta_1}$ and $\lambda = \|\boldsymbol{\lambda}\|$. For simplicity we refer to $p_{\mu_1+\delta_1, \mu_1-\delta_1}$ as $\mathcal{D}_1$. Our goal is to bound the following probability

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \tanh(\lambda x_i) x_i - \mathbb{E}_{x \sim \mathcal{D}_1}[\tanh(\lambda x) x]\right| > \kappa\right) \tag{6.14}$$

to do so we use the general large deviation technique. Because of symmetry of $\mathcal{D}_1$, we have that the above probability is equal twice the probability

$$\mathbb{P}\left(\exp\left(\theta\sum_{i=1}^{n}\left(\tanh(\lambda x_i)x_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)x]\right)\right)>\exp\left(\theta n\kappa\right)\right).$$

Using Markov's inequality we get that

$$\begin{aligned}\mathbb{P}\left(\exp\left(\theta\sum_{i=1}^{n}\left(\tanh(\lambda x_i)x_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)x]\right)\right)>\exp\left(\theta n\kappa\right)\right)\leq &\qquad (6.15)\\ \leq\left(\frac{\mathbb{E}_{x\sim\mathcal{D}_1}\left[\exp\left(\theta\left(\tanh(\lambda x)x-\mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)x\right]\right)\right)\right]}{\exp(\theta\kappa)}\right)^n &\qquad (6.16)\end{aligned}$$

We therefore have to bound the quantity

$$\begin{aligned}&\mathbb{E}_{x\sim\mathcal{D}_1}\left[\exp\left(\theta\left(\tanh(\lambda x)x-\mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)x\right]\right)\right)\right]\leq\\ &\leq\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\exp\left(\theta\left(\tanh(\lambda x)x-\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)x\right]\right)\right)\right]\cdot\\ &\cdot\exp\left(\frac{\theta}{2}\left(\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)x\right]-\mathbb{E}_{x\sim\mathcal{N}(\mu_1-\delta_1,1)}\left[\tanh(\lambda x)x\right]\right)\right)\end{aligned}\qquad (6.17)$$

The first term of (6.17) is equal to

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\exp\left(\theta\left(\tanh(\lambda(x+\mu_1+\delta_1))(x+\mu_1+\delta_1)-\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\tanh(\lambda(x+\mu_1+\delta_1))(x+\mu_1+\delta_1)\right]\right)\right)\right].$$

Now we use the following Lemma 2.1 of [Wai15].

Lemma 2.1 of Wainright 2015. *Suppose that $f:\mathbb{R}^d\to\mathbb{R}^d$ is differentiable. Then for any convex function $\phi:\mathbb{R}\to\mathbb{R}$, we have*

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

where $\boldsymbol{x},\boldsymbol{y}\sim\mathcal{N}(0,I)$ are standard multivariate Gaussian, and independent.

Combining this lemma with the fact that

$$\frac{\partial\tanh(\lambda(x+\mu_1+\delta_1))(x+\mu_1+\delta_1)}{\partial x}=\tanh'(\lambda(x+\mu_1+\delta_1))\lambda(x+\mu_1+\delta_1)+\tanh(\lambda(x+\mu_1+\delta_1))\leq 2$$

we get that

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\exp\left(\theta\left(\tanh(\lambda x)x-\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)x\right]\right)\right)\right]\leq\exp\left(5\theta^2\right).$$

Now for the second term of (6.17) we notice that as we present in Section 3

$$\frac{\partial}{\partial\mu}\mathbb{E}_{x\sim\mathcal{N}(\mu,1)}\left[\tanh(\lambda x)x\right]=\mathbb{E}_{x\sim\mathcal{N}(\mu,1)}\left[\tanh'(\lambda x)\lambda x+\tanh(\lambda x)\right]\leq 2$$

which imlies using the mean value theorem that

$$\exp\left(\frac{\theta}{2}\left(\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)x\right]-\mathbb{E}_{x\sim\mathcal{N}(\mu_1-\delta_1,1)}\left[\tanh(\lambda x)x\right]\right)\right)\leq\exp\left(\theta 2\left|\delta_1\right|\right).$$

Putting all together to (6.17) we have that

$$\mathbb{E}_{x\sim\mathcal{D}_1}\left[\exp\left(\theta\left(\tanh(\lambda x)x-\mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)x\right]\right)\right)\right]\leq\exp\left(5\theta^2+2\theta\left|\delta_1\right|\right)$$

which implies that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)x_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)x]\right|>\kappa\right)\leq 2\exp\left(n\left(5\theta^2+2\theta\left|\delta_1\right|-\theta\kappa\right)\right)\implies$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)x_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)x]\right|>\left|\delta_1\right|+\tau\right)\leq 2\exp\left(-\frac{n\tau^2}{20}\right)\implies$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)x_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)x]\right|>\left|\delta_1\right|+\frac{\varepsilon}{\sqrt{d}}\right)\leq 2\exp\left(-\frac{n\varepsilon^2}{20d}\right)$$

Therefore with $n=O(\frac{d}{\varepsilon^2}\log(d/\varepsilon^2))=\tilde{O}\left(\frac{d}{\varepsilon^2}\right)$ we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)x_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)x]\right|>\left|\delta_1\right|+\frac{\varepsilon}{\sqrt{d}}\right)\leq\text{poly}\left(\frac{\varepsilon^2}{d}\right)\tag{6.18}$$

We now consider $i=2$. In this case

$$\tilde{\lambda}_2'=\frac{1}{n}\sum_{i=1}^{n}\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}_i\rangle)\langle\boldsymbol{x}_i,\boldsymbol{v}_2\rangle,$$

$$\lambda_2'=\mathbb{E}[\tilde{\lambda}_2']=\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}+\boldsymbol{\delta},\boldsymbol{\mu}-\boldsymbol{\delta}}}\left[\tanh(\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle)\langle\boldsymbol{x},\boldsymbol{v}_2\rangle\right].$$

As before we define $\mu_2=\langle\boldsymbol{\mu},\boldsymbol{v}_2\rangle$, $\delta_2=\langle\boldsymbol{\delta},\boldsymbol{v}_2\rangle$ and we have

$$\tilde{\lambda}_2'=\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i,$$

$$\lambda_2'=\mathbb{E}[\tilde{\lambda}_2']=\mathbb{E}_{x\sim p_{\mu_1+\delta_1,\mu_1-\delta_1}}\left[\tanh(\lambda x)\right]\mathbb{E}_{y\sim p_{\mu_2+\delta_2,\mu_2-\delta_2}}\left[y\right]=\mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)\right]\mu_2$$

where x_i is distributed as $p_{\mu_1+\delta_1,\mu_1-\delta_1}$, y_i is distributed as $p_{\mu_2+\delta_2,\mu_2-\delta_2}$ and $\lambda=\|\boldsymbol{\lambda}\|$. For simplicity we refer to $p_{\mu_2+\delta_2,\mu_2-\delta_2}$ as $\mathcal{D}_2$. Our goal is to bound the following probability

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)]\mu_2\right|>\kappa\right)\tag{6.19}$$

to do so we use the general large deviation technique. Using the symmetry of $\mathcal{D}_1$ and $\mathcal{D}_2$ we have that the above probability is equal twice the

$$\mathbb{P}\left(\exp\left(\theta\sum_{i=1}^{n}\left(\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)]\mu_2\right)\right)>\exp\left(\theta n\kappa\right)\right).$$

Using Markov's inequality we get that

$$\mathbb{P}\left(\exp\left(\theta\sum_{i=1}^{n}\left(\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)]\mu_2\right)\right)>\exp\left(\theta n\kappa\right)\right)\leq\tag{6.20}$$

$$\leq\left(\frac{\mathbb{E}_{x\sim\mathcal{D}_1,y\sim\mathcal{D}_2}\left[\exp\left(\theta\left(\tanh(\lambda x)y-\mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)\right]\mu_2\right)\right)\right]}{\exp(\theta\kappa)}\right)^n\tag{6.21}$$

Using the fact that because of the initialization of EM $\lambda \geq \mu_1$ and by assumption $\delta_1 \leq \mu_1$ and also let $\alpha = \min(1, \mu_1)$ we have to bound the quantity

$$\mathbb{E}_{x\sim\mathcal{D}_1, y\sim\mathcal{D}_2}\left[\exp\left(\theta\left(\tanh(\lambda x)y - \mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)\right]\mu_2\right)\right)\right] \leq$$

$$\leq \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{D}_2}\left[\exp\left(\theta\left(\tanh(\lambda x)y - \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right]\mu_2\right)\right)\right]\cdot$$

$$\cdot\exp\left(\frac{\theta\mu_2}{2}\left(\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right] - \mathbb{E}_{x\sim\mathcal{N}(\mu_1-\delta_1,1)}\left[\tanh(\lambda x)\right]\right)\right) \leq$$

$$\leq \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{D}_2}\left[\exp\left(\theta\left(\tanh(\lambda x)y - \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right]\mu_2\right)\right)\right]\cdot\exp\left(\theta\left|\mu_2\right|\left|\delta_1\right|\right) \leq$$

$$\leq \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{N}(\mu_2+\delta_2,1)}\left[\exp\left(\theta\left(\tanh(\lambda x)y - \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right](\mu_2+\delta_2)\right)\right)\right]\cdot$$

$$\cdot\exp\left(\theta\left|\mu_2\right|\alpha\left|\delta_1\right|\right)\cdot\exp\left(\theta\left|\delta_2\right|\right) =$$

$$= \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{N}(0,1)}\left[\exp\left(\theta\left(\tanh(\lambda x)(y+\mu_2+\delta_2) - \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right](\mu_2+\delta_2)\right)\right)\right]\cdot$$

$$\cdot\exp\left(\theta\left|\mu_2\right|\alpha\left|\delta_1\right|\right)\cdot\exp\left(\theta\left|\delta_2\right|\right) =$$

$$= \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{N}(0,1)}\left[\exp\left(\theta(\mu_2+\delta_2)\left(\tanh(\lambda x) - \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right]\right)\right)\right]\cdot$$

$$\cdot\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{N}(0,1)}\left[\exp\left(\theta\tanh(\lambda x)y\right)\right]\cdot\exp\left(\theta\left|\mu_2\right|\alpha\left|\delta_1\right|\right)\cdot\exp\left(\theta\left|\delta_2\right|\right). \tag{6.22}$$

The first term of (6.22) is equal to

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{N}(0,1)}\left[\exp\left(\theta(\mu_2+\delta_2)\left(\tanh(\lambda x) - \mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right]\right)\right)\right].$$

Observe now that because of the initial conditions of EM at this step we have that $\mu_1 \geq \mu_2$ and $\lambda \geq \|\mu\| \geq \mu_1$. Also it is not hard to prove that $\tanh(y^2) \geq 1 - \frac{1}{y}$. This means that if $\mu_2 \geq \operatorname{poly}\log(d/\varepsilon^2)$ then with probability at least $\exp(-2\operatorname{poly}\log(d/\varepsilon^2))$ we will have that $\tanh(\lambda x) \geq 1 - \frac{1}{\mu_2}$. Now using the convexity of $\exp(\cdot)$ we have that the above term is less than

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\frac{\tanh(\lambda x) - \mathbb{E}[\tanh(\lambda x)] - 1 + \frac{1}{\mu_2}}{2}\exp(\theta(\mu_2+\delta_2)) + \right.$$

$$\left. + \frac{-\tanh(\lambda x) + \mathbb{E}[\tanh(\lambda x)] + 1}{2}\exp(-\theta(\mu_2+\delta_2))\right].$$

Now using a simple Taylor expansion used in the proof of the Hoeffding bound we get that the first term of (6.22) is less than or equal to

$$\exp\left(\frac{\delta_2^2\operatorname{poly}\log(d/\varepsilon^2)\theta^2}{2}\right) \leq \exp\left(\frac{\operatorname{poly}\log(d/\varepsilon^2)\theta^2}{2}\right)$$

and this holds with high probability at least $\operatorname{poly}\left(\frac{\varepsilon^2}{d}\right)$.
Now for the second term of (6.22) we have that

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1), y\sim\mathcal{N}(0,1)}\left[\exp\left(\theta\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right]y\right)\right] =$$

$$= \exp\left(\frac{\theta^2}{2}\left(\mathbb{E}_{x\sim\mathcal{N}(\mu_1+\delta_1,1)}\left[\tanh(\lambda x)\right]\right)^2\right) \leq \exp\left(\frac{\theta^2}{2}\right)$$

Putting all together to (6.22) we have that

$$\mathbb{E}_{x\sim\mathcal{D}_1,y\sim\mathcal{D}_2}\left[\exp\left(\theta\left(\tanh(\lambda x)y-\mathbb{E}_{x\sim\mathcal{D}_1}\left[\tanh(\lambda x)\right]\mu_2\right)\right)\right]\leq\exp\left(\frac{1+\operatorname{poly}\log(d/\varepsilon^2)}{2}\theta^2+\left(|\mu_2|\,\alpha\,|\delta_1|+|\delta_2|\right)\theta\right)$$

which implies that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)]\mu_2\right|>\kappa\right)\leq$$

$$\leq 2\exp\left(n\left(\frac{1+\operatorname{poly}\log(d/\varepsilon^2)}{2}\theta^2+\left(|\mu_2|\,\alpha\,|\delta_1|+|\delta_2|-\kappa\right)\theta\right)\right)\implies$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)]\mu_2\right|>\varepsilon\,|\mu_2|\,\alpha+|\delta_2|+\tau\right)\leq 2\exp\left(-\frac{n\tau^2}{4\operatorname{poly}\log(d/\varepsilon^2)}\right)\implies$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1}[\tanh(\lambda x)]\mu_2\right|>\varepsilon\,|\mu_2|\,\alpha+|\delta_2|+\frac{\varepsilon}{\sqrt{d}}\right)\leq 2\exp\left(-\frac{n\varepsilon^2}{4d\operatorname{poly}\log(d/\varepsilon^2)}\right)$$

Therefore with $n=4\frac{d}{\varepsilon^2}\operatorname{poly}\log(d/\varepsilon^2)=\tilde{O}\left(\frac{d}{\varepsilon^2}\right)$ we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1,y\sim\mathcal{D}_2}[\tanh(\lambda x)y]\right|>\varepsilon\alpha\,|\mu_2|+|\delta_2|+\frac{\varepsilon}{\sqrt{d}}\right)\leq\operatorname{poly}\left(\frac{\varepsilon^2}{d}\right)\tag{6.23}$$

For any $i\geq 3$ we follow the same analysis as for the bound (6.23) but because of the definition of the basis $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d\}$ we have that $\mu_i=\langle\boldsymbol{\mu},\boldsymbol{v}_i\rangle=0$ and therefore

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tanh(\lambda x_i)y_i-\mathbb{E}_{x\sim\mathcal{D}_1,y\sim\mathcal{D}_3}[\tanh(\lambda x)y]\right|>|\delta_3|+\frac{\varepsilon}{\sqrt{d}}\right)\leq\operatorname{poly}\left(\frac{\varepsilon^2}{d}\right)\tag{6.24}$$

Finally if we combine (6.18), (6.23) and (6.24) using the observation that $\left\|\boldsymbol{\lambda}^{(t+1)}-\boldsymbol{\mu}\right\|\geq|\mu_2|$ we get that

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{\lambda}}^{(t+1)}-\boldsymbol{\lambda}\right\|>\varepsilon+\varepsilon\min(\|\boldsymbol{\mu}\|,1)\left\|\boldsymbol{\lambda}^{(t)}-\boldsymbol{\mu}\right\|\right)\leq\operatorname{poly}\left(\frac{\varepsilon^2}{d}\right).$$

□

Proof of Theorem 3: Lemma 10 and Equation (6.13) imply that, using $\tilde{O}(d/\varepsilon^2\mu^4)$ samples, we have:

$$\left\|\tilde{\boldsymbol{\lambda}}^{(t+1)}-\boldsymbol{\mu}\right\|_\Sigma\leq\left(e^{-\frac{\mu^2}{6}}+\varepsilon\min(\mu,1)\right)\left\|\tilde{\boldsymbol{\lambda}}^{(t)}-\boldsymbol{\mu}\right\|_\Sigma+2\varepsilon\mu^2.$$

If $\mu>1$ then

$$\left(e^{-\frac{\mu^2}{6}}+\varepsilon\right)\leq\frac{9}{10}.$$

If $\mu \leq 1$ then

$$\left(e^{-\frac{\mu^2}{6}} + \varepsilon\mu\right) \leq \left(e^{-\frac{\mu^2}{6}} + \frac{\mu^2}{20}\right) \leq e^{-\frac{\mu^2}{10}}.$$

These imply that

$$\left\|\tilde{\boldsymbol{\lambda}}^{(t+1)} - \boldsymbol{\mu}\right\|_{\Sigma} \leq \max\left(e^{-\frac{\mu^2}{10}}, \frac{9}{10}\right)\left\|\tilde{\boldsymbol{\lambda}}^{(t)} - \boldsymbol{\mu}\right\|_{\Sigma} + 2\varepsilon\mu^2. \tag{6.25}$$

Therefore we need $\tilde{O}\left(\max\left(\frac{1}{\mu^2}, 1\right)\log(1/\varepsilon)\right)$ steps in order to get error 3ε. Since each step requires $\tilde{O}\left(\frac{d}{\varepsilon^2\mu^4}\right)$ samples, Theorem 3 follows.

Acknowledgements

We thank Sham Kakade for suggesting the problem to us, and for initial discussions. The authors were supported by NSF Awards CCF-0953960 (CAREER), CCF-1551875, CCF-1617730, and CCF-1650733, ONR Grant N00014-12-1-0999, and a Microsoft Faculty Fellowship.

References

[AK01] Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[BS10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[BV08] S Charles Brubaker and Santosh S Vempala. Isotropic PCA and affine-invariant clustering. In *the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008.

[BWY17] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[CDV09] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*, 2009.

[CH08] Stéphane Chrétien and Alfred O Hero. On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

[CR08] Kamalika Chaudhuri and Satish Rao. Learning Mixtures of Product Distributions Using Correlations and Independence. In *the 21st International Conference on Computational Learning Theory (COLT)*, 2008.

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory*, pages 1183–1213, 2014.

[DKT15] Constantinos Daskalakis, Gautam Kamath, and Christos Tzamos. On the structure, covering, and learning of poisson multinomial distributions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1203–1217. IEEE, 2015.

[DL12] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[DLR77] 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.

[DS07] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

[GHK15] Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *the 47th Annual ACM on Symposium on Theory of Computing (STOC)*, 2015.

[HK13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *the 4th conference on Innovations in Theoretical Computer Science (ITCS)*, 2013.

[HP15] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 753–760. ACM, 2015.

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

[KSV05] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *the 18th International Conference on Computational Learning Theory (COLT)*, 2005.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[RW84] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Tse04] Paul Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[Wai15] JM Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley*, 2015.

[Wu83] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[XHM16] Ji Xu, Daniel Hsu, and Arian Maleki. Global analysis of Expectation Maximization for mixtures of two Gaussians. In *the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.