Submitted to the Annals of Statistics

SINGULARITY, MISSPECIFICATION, AND THE CONVERGENCE RATE OF EM

By Raaz Dwivedi[*,¶], Nhat Ho[*,¶], Koulik Khamaru[*,¶], Martin J. Wainwright[†,¶,‖], Michael I. Jordan[‡,¶] and Bin Yu[§,¶]

University of California, Berkeley[¶] and Voleon Group[‖]

A line of recent work has analyzed the behavior of the Expectation-Maximization (EM) algorithm in the well-specified setting, in which the population likelihood is locally strongly concave around its maximizing argument. Examples include suitably separated Gaussian mixture models and mixtures of linear regressions. We consider over-specified settings in which the number of fitted components is larger than the number of components in the true distribution. Such misspecified settings can lead to singularity in the Fisher information matrix, and moreover, the maximum likelihood estimator based on n i.i.d. samples in d dimensions can have a non-standard $\mathcal{O}((d/n)^{\frac{1}{4}})$ rate of convergence. Focusing on the simple setting of two-component mixtures fit to a d-dimensional Gaussian distribution, we study the behavior of the EM algorithm both when the mixture weights are different (unbalanced case), and are equal (balanced case). Our analysis reveals a sharp distinction between these two cases: in the former, the EM algorithm converges geometrically to a point at Euclidean distance of $\mathcal{O}((d/n)^{\frac{1}{2}})$ from the true parameter, whereas in the latter case, the convergence rate is exponentially slower, and the fixed point has a much lower $\mathcal{O}((d/n)^{\frac{1}{4}})$ accuracy. Analysis of this singular case requires the introduction of some novel techniques: in particular, we make use of a careful form of localization in the associated empirical process, and develop a recursive argument to progressively sharpen the statistical rate.

1. Introduction. The growth in the size and scope of modern data sets has presented the field of statistics with a number of challenges, one of them being how to deal with various forms of heterogeneity. Mixture models provide a principled approach to modeling heterogeneous collections of data

[*]Raaz Dwivedi, Nhat Ho, and Koulik Khamaru contributed equally to this work.

[†]Supported by Office of Naval Research grant DOD ONR-N00014-18-1-2640 and National Science Foundation grant NSF-DMS-1612948

[‡]Supported by Army Research Office grant W911NF-17-1-0304

[§]Supported by National Science Foundation grant NSF-DMS-1613002

MSC 2010 subject classifications: Primary 62F15, 62G05; secondary 62G20

Keywords and phrases: Mixture models, Expectation-maximization, Fisher information matrix, Empirical process, Non-asymptotic convergence guarantees, Localization argument

(that are usually assumed i.i.d.). In practice, it is frequently the case that the number of mixture components in the fitted model does not match the number of mixture components in the data-generating mechanism. It is known that such mismatch can lead to substantially slower convergence rates for the maximum likelihood estimate (MLE) for the underlying parameters. In contrast, relatively less attention has been paid to the computational implications of this mismatch. In particular, the algorithm of choice for fitting finite mixture models is the Expectation-Maximization (EM) algorithm, a general framework that encompasses various types of divide-and-conquer computational strategies. The goal of this paper is to gain a fundamental understanding of the behavior of EM when used to fit over-specified mixture models.

Statistical issues with over-specification. While density estimation in finite mixture models is relatively well understood [12, 26], characterizing the behavior of maximum likelihood for parameter estimation has remained challenging. The main difficulty for analyzing the MLE in such settings arises from label switching between the mixtures [23, 25], and lack of strong concavity in the likelihood. Such issues do not interfere with density estimation, since the standard divergence measures like the Kullback-Leibler and Hellinger distances remain invariant under permutations of labels, and strong concavity is not required. An important contribution to the understanding of parameter estimation in finite mixture models was made by Chen [4]. He considered a class of over-specified finite mixture models; here the term "over-specified" means that the model to be fit has more mixture components than the distribution generating the data. In an interesting contrast to the usual $n^{-\frac{1}{2}}$ convergence rate for the MLE based on n samples, Chen showed that for estimating scalar location parameters in a certain class of over-specified finite mixture models, the corresponding rate slows down to $n^{-\frac{1}{4}}$. This theoretical result has practical significance, since methods that over-specify the number of mixtures are often more feasible than methods that first attempt to estimate the number of components, and then estimate the parameters using the estimated number of components [24]. In subsequent work, Nguyen [21] and Heinrich et al. [14] have characterized the (minimax) convergence rates of parameter estimation rates for mixture models in both exactly-fitted or over-specified settings in terms of the Wasserstein distance.

Computational concerns with mixture models. While the papers discussed above address the statistical behavior of a global maximum of the log-likelihood, they do not consider the associated computational issues of ob-

taining such a maximum. In general settings, non-convexity of the log-likelihood makes it impossible to guarantee that the iterative algorithms used in practice converge to the global optimum, or equivalently the MLE. Perhaps the most widely used algorithm for computing the MLE is the expectation-maximization (EM) algorithm [8]. Early work on the EM algorithm [29] showed that its iterates converge asymptotically to a local maximum of the log-likelihood function for a broad class of incomplete data models; this general class includes the fitting of mixture models as a special case. The EM algorithm has also been studied in the specific setting of Gaussian mixture models; here we find results both for the population EM algorithm, which is the idealized version of EM based on an infinite sample size, as well as the usual sample-based EM algorithm that is used in practice. For Gaussian mixture models, the population EM algorithm is known to exhibit various convergence rates, ranging from linear to super-linear (quasi-Newton like) convergence if the overlap between the mixture components tends to zero [20, 31]. It has also been noted in several papers [20, 22] that the convergence of EM can be prohibitively slow when the mixtures are not well separated.

Prior work on EM. Balakrishnan et al. [1] laid out a general theoretical framework for analysis of the EM algorithm, and in particular how to prove non-asymptotic bounds on the Euclidean distance between sample-based EM iterates and the true parameter. When applied to the special case of two-component Gaussian location mixtures, assumed to be well-specified and suitably separated, their theory guarantees that (1) population EM updates enjoy a geometric rate of convergence to the true parameter when initialized in a sufficiently small neighborhood around the truth, and (2) sample-based EM updates converge to an estimate at Euclidean distance of order $(d/n)^{\frac{1}{2}}$, based on n i.i.d. draws from a finite mixture model in $\mathbb{R}^d$. Further work in this vein has characterized the behavior of EM in a variety of settings for two Gaussian mixtures, including convergence analysis with additional sparsity constraints [13, 28, 33], global convergence of population EM [30], guarantees of geometric convergence under less restrictive conditions on the two mixture components [7, 16], analysis of EM with unknown mixture weights, means and covariances for two mixtures [3], and the analysis of EM to more than two Gaussian components [13, 32]. Other related work has provided optimization-theoretic guarantees for EM by viewing it in a generalized surrogate function framework [18], and analyzed the statistical properties of confidence intervals based on an EM estimator [5].

An assumption common to all of this previous work is that there is no misspecification in the fitting of the Gaussian mixtures; in particular, it is

assumed that the data is generated from a mixture model with the same number of components as the fitted model. A portion of our recent work [9] has shown that EM retains its fast convergence behavior—albeit to a biased estimate—in *under-specified* settings where the number of components in the fitted model are less than that in the true model. However, as noted above, in practice, it is most common to use over-specified mixture models. For these reasons, it is desirable to understand how the EM algorithm behaves in the over-specified settings.

Our contributions. The goal of this paper is to shed some light on the non-asymptotic performance of the EM algorithm for over-specified mixtures. We provide a comprehensive study of over-specified mixture models when fit to a particularly simple (non-mixture) data-generating mechanism; a multivariate normal distribution $\mathcal{N}(0, \sigma^2 I_d)$ in d dimensions with known scale parameter $\sigma > 0$. This setting, despite its simplicity, suffices to reveal some rather interesting properties of EM in the over-specified context. In particular, we obtain the following results.

- **Two-mixture unbalanced fit:** For our first model class, we study a mixture of two location-Gaussian distributions with unknown location, known variance and known unequal weights for the two components. For this case, we establish that the population EM updates converge at a geometric rate to the true parameter; as an immediate consequence, the sample-based EM algorithm converges in $\mathcal{O}(\log(n/d))$ steps to a ball of radius $(d/n)^{\frac{1}{2}}$. The fast convergence rate of EM under the unbalanced setting provides an antidote to the pessimistic belief that statistical estimators generically exhibit slow convergence for over-specified mixtures.
- **Two-mixture balanced fit:** In the balanced version of the problem in which the mixture weights are equal to $\frac{1}{2}$ for both components, we find that the EM algorithm behaves very differently. Beginning with the population version of the EM algorithm, we show that it converges to the true parameter from an arbitrary initialization. However, the rate of convergence varies as a function of the distance of the current iterate from the true parameter value, becoming exponentially slower as the iterates approach the true parameter. This behavior is in sharp contrast to well-specified settings [1, 7, 32], where the population updates converge at a geometric rate. We also show that our rates for population EM are tight. By combining the slow convergence of population EM with a novel localization argument, one involving the empirical process restricted to an annulus, we show that the sample-

based EM iterates converge to a ball of radius $(d/n)^{\frac{1}{4}}$ around the true parameter after $\mathcal{O}((n/d)^{\frac{1}{2}})$ steps. The $n^{-\frac{1}{4}}$ component of the Euclidean error matches known guarantees for the global maximum of the MLE [4]. The localization argument in our analysis is of independent interest, because such techniques are not required in analyzing the EM algorithm in well-specified settings when the population updates are globally contractive. We note that ball-based localization methods are known to be essential in deriving sharp statistical rates for M-estimators (e.g., [2, 17, 26]); to the best of our knowledge, the use of an annulus-based localization argument in analyzing an algorithm is novel.

Moreover, we show via extensive numerical experiments that the fast convergence of EM for the unbalanced fit is a special case; and that the slow behavior of EM proven for the balanced fit (in particular the rate of order $n^{-\frac{1}{4}}$) arises in several general (including more than two components) over-specified Gaussian mixtures with known variance, known or unknown weights, and unknown location parameters.

Organization. The remainder of the paper is organized as follows. In Section 2 we provide illustrative simulations of EM in different settings in order to motivate the settings analyzed later in the paper. We then provide a thorough analysis of the convergence rates of EM when over-fitting Gaussian data with two components in Section 3 and the key ideas of the novel proof techniques in Section 4. We provide a thorough discussion of our results in Section 5, exploring their general applicability, and presenting further simulations that substantiate the value of our theoretical framework. Detailed proofs of our results and discussion of certain additional technical aspects of our results are provided in the supplementary material [10].

Notation. For any two sequences a_n and b_n, the notation $a_n \precsim b_n$ or $a_n = \mathcal{O}(b_n)$ means that $a_n \leq C b_n$ for some universal constant C. Similarly, the notation $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ denotes that both the conditions, $a_n \precsim b_n$ and $b_n \precsim a_n$, hold. Throughout this paper, π denotes a variable and π denotes the mathematical constant "pi".

Experimental settings. We summarize a few common aspects of the numerical experiments presented in the paper. Population-level computations were done using numerical integration on a sufficiently fine grid. With finite samples, the stopping criteria for the convergence of EM were: (1) the change in the iterates was small enough, or (2) the number of iterations was too large (greater than $100,000$). Experiments were averaged over several repetitions

(ranging from 25 to 400). In majority of the runs, for each case, criteria (1) led to convergence. In our plots for sample EM, we report $\widehat{m}_e + 2\widehat{s}_e$ on the y-axis, where $\widehat{m}_e, \widehat{s}_e$ respectively denote the mean and standard deviation across the experiments for the metric under consideration, e.g., the parameter estimation error. Furthermore, whenever a slope is provided, it is the slope for the least-squares fit on the log-log scale for the quantity on y-axis when fitted with the quantity reported on the x-axis. For instance, in Figure 1(b), we plot $|\widehat{\theta}_n - \theta^*|$ on the y-axis value versus the sample size n on the x-axis, averaged over 400 experiments, accounting for the deviation across these experiments. Furthermore, the green dotted line with legend $\pi = 0.3$ and the corresponding slope -0.48 denote the least-squares fit and the respective slope for $\log|\widehat{\theta}_n - \theta^*|$ (green solid dots) with $\log n$ for the experiments corresponding to the setting $\pi = 0.3$.

2. Motivating simulations and problem set-up. In this section, we explore a wide range of behavior demonstrated by EM for certain settings of over-specified location Gaussian mixtures. We begin with several simulations that illustrate fast and slow convergence of EM for various settings, and serve as a motivation for the theoretical results derived later in the paper. We provide basic background on EM in Section 2.3, and describe the problems to be tackled.

2.1. *Problem set-up.* Let $\phi(\cdot; \mu, \Sigma)$ denote the density of a Gaussian random vector with mean μ and covariance Σ. Consider the two component Gaussian mixture model with density

$$f(x; \theta^*, \sigma, \pi) := \pi\phi(x; \theta^*, \sigma^2 I_d) + (1-\pi)\phi(x; -\theta^*, \sigma^2 I_d). \tag{1}$$

Given n samples from the distribution (1), suppose that we use the EM algorithm to fit a two-component location Gaussian mixture with fixed weights and variance[1] and special structure on the location parameters—more precisely, we fit the model with density

$$f(x; \theta, \sigma, \pi) := \pi\phi(x; \theta, \sigma^2 I_d) + (1-\pi)\phi(x; -\theta, \sigma^2 I_d) \tag{2}$$

using the EM algorithm, and take the solution[2] as an estimate of θ^*. An important aspect of the problem at hand is the signal strength, which is

[1]Refer to Section 5 for a discussion for the case of unknown weights and variances.

[2]Strictly speaking, different initialization of EM may converge to different estimates. For the settings analyzed theoretically in this work, the EM always converges towards the same estimate in the limit of infinite steps, and we use a stopping criterion to determine the final estimate. See the discussion on experimental settings in Section 1 for more details.

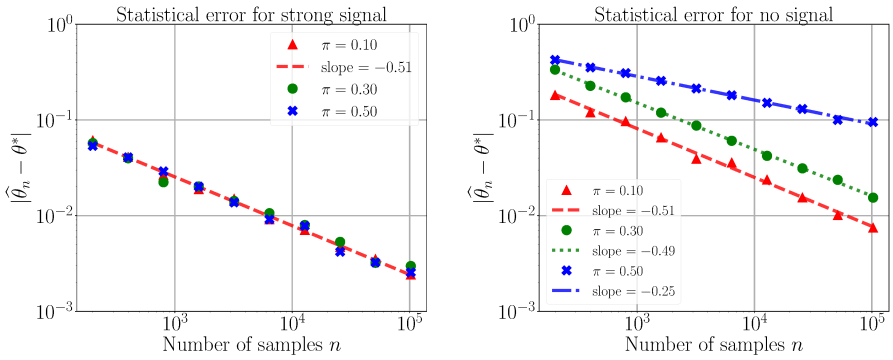
measured as the separation between the means of mixture components relative to the spread in the components. For the model (1), the signal strength is given by the ratio $\|\theta^*\|_2/\sigma$. When this ratio is large, we refer to it as the *strong signal* case; otherwise, it corresponds to the *weak signal* case. Of particular interest to us is the behavior of EM in the limit of weak signal when there is no separation; i.e., $\|\theta^*\|_2 = 0$. For such cases, we call the fit (2) an *unbalanced* fit when $\pi \neq \frac{1}{2}$ and a *balanced* fit when $\pi = \frac{1}{2}$. Note that the setting of $\theta^* = 0$ corresponds to the simplest case of over-specified fit, since the true model has just one component (standard normal distribution irrespective of the parameter π) but the fitted model has two (one extra) component (unless the EM estimate is also 0). We now present the empirical behavior of EM for these models and defer the derivation of EM updates to Section 2.3.

2.2. *Numerical Experiments: Fast to slow convergence of EM.* We begin with a numerical study of the effect of separation among the mixtures on the statistical behavior of the estimates returned by EM. Our main observation is that weak or no separation leads to relatively low accuracy estimates. Additional simulations for more general mixtures, including more than two components, are provided in Section 5.3. Next, via numerical integration on a grid with sufficiently small discretization width, we simulate the behavior of the population EM algorithm width—an idealized version of EM in the limit of infinite samples—in order to understand the effect of signal strength on EM's algorithmic rate of convergence, i.e., the number of steps needed for population EM to converge to a desired accuracy. We observe a slow down of EM on the algorithmic front when the signal strength approaches zero.

2.2.1. *Effect of signal strength on sample EM.* In Figure 1, we show simulation results for data generated from the model (1) in dimension $d = 1$ and noise variance $\sigma^2 = 1$, and for three different values of the weight $\pi \in \{0.1, 0.3, 0.5\}$. In all cases, we fit a two-location Gaussian mixture with fixed weights and variance as specified by equation (2). The two panels show the estimation error of the EM solution as a function of n for two distinct cases of the data-generating mechanism: (a) in the strong signal case, we set $\theta^* = 5$ so that the data has two well-separated mixture components, and (b) to obtain the limiting case of no signal, we set $\theta^* = 0$, so that the two mixture components in the data-generating distribution collapse to one, and we are simply fitting the data from a standard normal distribution.

In the strong signal case, it is well known [1, 7] that EM solutions have an estimation error (measured by the Euclidean distance between the EM estimate and the true parameter θ^*) that achieves the classical (parametric)

rate $n^{-\frac{1}{2}}$; the empirical results in Figure 1(a) are simply a confirmation of these theoretical predictions. More interesting is the case of no signal (which is the limiting case with weak signal), where the simulation results shown in panel (b) of Figure 1 reveal a different story. In this case, whereas the EM solution (with random standard normal initialization) has an error that decays as $n^{-\frac{1}{2}}$ when $\pi \neq 1/2$, its error decays at the considerably slower rate $n^{-\frac{1}{4}}$ when $\pi = 1/2$. We return to these cases in further detail in Section 3.

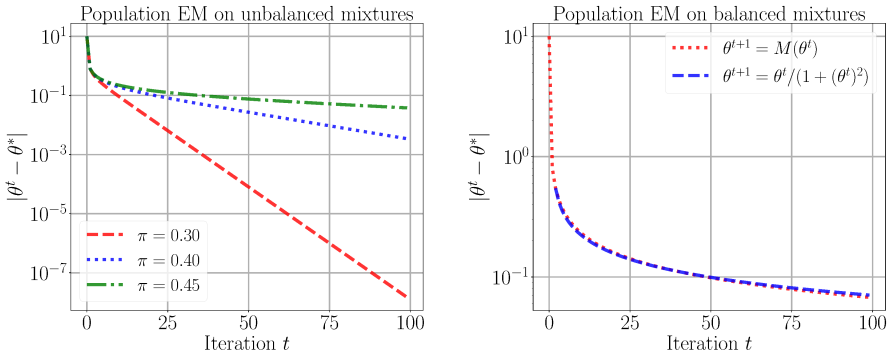


(a) $\theta^* = 5$ (b) $\theta^* = 0$

Fig 1. Plots of the error $|\widehat{\theta}_n - \theta^*|$ in the EM solution versus the sample size n, focusing on the effect of signal strength on EM solution accuracy. The true data distribution is given by $\pi\mathcal{N}(\theta^*, 1) + (1-\pi)\mathcal{N}(-\theta^*, 1)$ and we use EM to fit the model $\pi\mathcal{N}(\theta, 1)+(1-\pi)\mathcal{N}(-\theta, 1)$, generating the EM estimate $\widehat{\theta}_n$ based on n samples. (a) When the signal is strong, the estimation rate decays at the parametric rate $n^{-\frac{1}{2}}$, as revealed by the $-1/2$ slope in a least-square fit of the log error based on the log sample size $\log n$. (b) When there is no signal ($\theta^* = 0$), then depending on the choice of weight π in the fitted model, we observe two distinct scalings for the error: $n^{-\frac{1}{2}}$ when $\pi \neq 0.5$, and, $n^{-\frac{1}{4}}$ when $\pi = 0.5$, again as revealed by least-squares fits of the log error using the log sample size $\log n$.

2.2.2. *Interesting behavior of population EM.* The intriguing behavior of the sample EM algorithm in the "no signal" case motivated us to examine the behavior of population EM for this case. To be clear, while sample EM is the practical algorithm that can actually be applied, it can be insightful for theoretical purposes to first analyze the convergence of the population EM updates, and then leverage these findings to understand the behavior of sample EM [1]. Our analysis follows a similar road-map. Interestingly, for the case with $\theta^* = 0$, the population EM algorithm behaves significantly differently for the unbalanced fit ($\pi \neq \frac{1}{2}$) as compared to the balanced fit

($\pi = \frac{1}{2}$) (equation (2)). In Figure 2, we plot the distance of the population EM iterate θ^t to the true parameter value, $\theta^* = 0$, on the vertical axis, versus the iteration number t on the horizontal axis. With the vertical axis on a log scale, a geometric convergence rate of the algorithm shows up as a negatively sloped line (disregarding transient effects in the first few iterations).



(a) (b)

Fig 2. Behavior of the (numerically computed) population EM updates (8) when the underlying data distribution is $\mathcal{N}(0, 1)$. (a) Unbalanced mixture fits (2) with weights $(\pi, 1-\pi)$: We observe geometric convergence towards $\theta^* = 0$ for all $\pi \neq 0.5$ although the rate of convergence gets slower as $\pi \to 0.5$. (b) Balanced mixture fits (2) with weights $(0.5, 0.5)$: We observe two phases of convergence. First, EM quickly converges to ball of constant radius and then it exhibits slow convergence towards $\theta^* = 0$. Indeed, we see that during the slow convergence, the population EM updates track the curve given by $\theta^{t+1} = \theta^t/(1+(\theta^t)^2)$ very closely, as predicted by our theory.

For the unbalanced mixtures in panel (a), we see that EM converges geometrically quickly, although the rate of convergence (corresponding to the slope of the line) tends towards zero as the mixture weight π tends towards $1/2$ from below. For $\pi = 1/2$, we obtain a balanced mixture, and, as shown in the plot in panel (b), the convergence rate is now sub-geometric. In fact, the behavior of the iterates is extremely well characterized by the recursion $\theta \mapsto \frac{\theta}{1+\theta^2}$.

The theory to follow provides a precise characterization of the behavior seen in Figures 1(b) and 2. Furthermore, in Section 5, we provide further support for relevannce of our theoretical results in explaining the behavior of EM for other classes of over-specified models, including Gaussian mixture models with unknown weights as well as mixtures of linear regressions.

2.3. *EM updates for the model fit* (2). In this section, we provide a quick introduction to the EM updates. Readers familiar with the literature can

skip directly to the main results in Section 3. Recall that the two-component model fit is based on the density

$$\pi\phi(x;\theta,\sigma^2 I_d) + (1-\pi)\phi(x;-\theta,\sigma^2 I_d). \tag{3}$$

From now on we assume that the data is drawn from the zero-mean Gaussian distribution $\mathcal{N}(0,\sigma^2 I_d)$. Note that the model fit described above contains the true model with $\theta^* = 0$ and it is referred to as an over-specified fit since for any non-zero θ, the fitted model has two components.

The maximum likelihood estimate is obtained by solving the following optimization problem

$$\widehat{\theta}_n^{\mathrm{MLE}} \in \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \left\{\log(\pi\phi(x_i;\theta,\sigma^2 I_d) + (1-\pi)\phi(x_i;-\theta,\sigma^2 I_d))\right\}. \tag{4}$$

In general, there is no closed-form expression for $\widehat{\theta}_n^{\mathrm{MLE}}$. The EM algorithm circumvents this problem via a minorization-maximization scheme. Indeed, population EM is a surrogate method to compute the maximizer of the population log-likelihood

$$\mathcal{L}(\theta) := \mathbb{E}_X\left[\log(\pi\phi(X;\theta,\sigma^2 I_d) + (1-\pi)\phi(X;-\theta,\sigma^2 I_d)\right], \tag{5}$$

where the expectation is taken over the true distribution. On the other hand, sample EM attempts to estimate $\widehat{\theta}_n^{\mathrm{MLE}}$. We now describe the expressions for both the sample and population EM updates for the model-fit (3).

Given any point θ, the EM algorithm proceeds in two steps: (1) compute a surrogate function $Q(\cdot;\theta)$ such that $Q(\theta';\theta) \le \mathcal{L}(\theta')$ and $Q(\theta;\theta) = \mathcal{L}(\theta)$; and (2) compute the maximizer of $Q(\theta';\theta)$ with respect to θ'. These steps are referred to as the E-step and the M-step, respectively. In the case of two-component location Gaussian mixtures, it is useful to describe a hidden variable representation of the mixture model. Consider a binary indicator variable $Z \in \{0,1\}$ with the marginal distribution $\mathbb{P}(Z=1) = \pi$ and $\mathbb{P}(Z=0) = 1-\pi$, and define the conditional distributions

$$(X \mid Z=0) \sim \mathcal{N}(-\theta,\sigma^2 I_d), \quad \text{and} \quad (X \mid Z=1) \sim \mathcal{N}(\theta,\sigma^2 I_d).$$

These marginal and conditional distributions define a joint distribution over the pair (X,Z), and by construction, the induced marginal distribution over X is a Gaussian mixture of the form (3). For EM, we first compute the conditional probability of $Z=1$ given $X=x$:

$$w_\theta(x) = w_\theta^\pi(x) := \frac{\pi\exp\left(-\frac{\|\theta-x\|_2^2}{2\sigma^2}\right)}{\pi\exp\left(-\frac{\|\theta-x\|_2^2}{2\sigma^2}\right) + (1-\pi)\exp\left(-\frac{\|\theta+x\|_2^2}{2\sigma^2}\right)}. \tag{6}$$

Then, given a vector θ, the E-step in the population EM algorithm involves computing the minorization function $\theta' \mapsto Q(\theta', \theta)$. Doing so is equivalent to computing the expectation

$$Q(\theta'; \theta) = -\frac{1}{2}\mathbb{E}\left[w_\theta(X)\left\|X - \theta'\right\|_2^2 + (1 - w_\theta(X))\left\|X + \theta'\right\|_2^2\right], \tag{7}$$

where the expectation is taken over the true distribution (here $\mathcal{N}(0, \sigma^2 I_d)$. In the M-step, we maximize the function $\theta' \mapsto Q(\theta'; \theta)$. Doing so defines a mapping $M : \mathbb{R}^d \to \mathbb{R}^d$, known as the *population EM operator*, given by

$$M(\theta) = \arg\max_{\theta' \in \mathbb{R}^d} Q(\theta', \theta) = \mathbb{E}\Big[(2w_\theta(X) - 1)X\Big]. \tag{8}$$

In this definition, the second equality follows by computing the gradient $\nabla_{\theta'} Q$, and setting it to zero. In summary, for the two-component location mixtures considered in this paper, the population EM algorithm is defined by the sequence $\theta^{t+1} = M(\theta^t)$, where the operator M is defined in equation (8).

We obtain the *sample EM update* by simply replacing the expectation $\mathbb{E}$ in equations (7) and (8) by the empirical average based on an observed set of samples. In particular, given a set of i.i.d. samples $\{X_i\}_{i=1}^n$, the sample EM operator $M_n : \mathbb{R}^d \mapsto \mathbb{R}^d$ takes the form

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^n (2w_\theta(X_i) - 1)X_i. \tag{9}$$

Overall, the sample EM algorithm generates the sequence of iterates given by $\theta^{t+1} = M_n(\theta^t)$.

In the sequel, we study the convergence of EM both for the population EM algorithm in which the updates are given by $\theta^{t+1} = M(\theta^t)$, and the sample-based EM sequence given by $\theta^{t+1} = M_n(\theta^t)$. With this notation in place, we now turn to the main results of this paper.

3. Main results. In this section, we state our main results for the convergence rates of the EM updates under the unbalanced and balanced mixture fit. We start with the easier case of unbalanced mixture fit in Section 3.1 followed by the more delicate (and interesting) case of the balanced fit in Section 3.2.

3.1. *Behavior of EM for unbalanced mixtures.* We begin with a characterization of both the population and sample-based EM updates in the setting of unbalanced mixtures. In particular, we assume that the fitted two-components mixture model (3) has known weights π and $1 - \pi$, where $\pi \in (0, 1/2)$. The following result characterizes the behavior of the EM updates for this set-up.

THEOREM 1. *Suppose that we fit an unbalanced instance (i.e., $\pi \neq \frac{1}{2}$) of the mixture model (3) to $\mathcal{N}(0, \sigma^2 I_d)$ data. Then:*

(a) The population EM operator (8) is globally strictly contractive, meaning that

$$\|M(\theta)\|_2 \leq \left(1 - \rho^2/2\right) \|\theta\|_2 \qquad \text{for all } \theta \in \mathbb{R}^d, \tag{10a}$$

where $\rho := |1 - 2\pi| \in (0, 1)$.

(b) There are universal constants c, c' such that given any $\delta \in (0, 1)$ and a sample size $n \geq c\frac{\sigma^2}{\rho^4}(d + \log(1/\delta))$, the sample EM sequence $\theta^{t+1} = M_n(\theta^t)$ generated by the update (9) satisfies the upper bound

$$\left\|\theta^t\right\|_2 \leq \left\|\theta^0\right\|_2 \left(1 - \frac{\rho^2}{2}\right)^t + \frac{c'(\left\|\theta^0\right\|_2 \sigma^2 + \rho\sigma)}{\rho^2} \sqrt{\frac{d + \log(1/\delta)}{n}}, \tag{10b}$$

with probability at least $1 - \delta$.

See Appendix A.1 in the supplement [10] for the proof of this theorem.

Fast convergence of population EM. The bulk of the effort in proving Theorem 1 lies in establishing the guarantee (10a) for the population EM iterates. Such a contraction bound immediately yields the exponential fast convergence of the population EM updates $\theta^{t+1} = M(\theta^t)$ to $\theta^* = 0$:

$$\left\|\theta^T\right\|_2 \leq \epsilon \quad \text{for} \quad T \geq \frac{1}{\log \frac{1}{(1-\rho^2/2)}} \cdot \log\left(\frac{\left\|\theta^0\right\|_2}{\epsilon}\right). \tag{11}$$

Since the mixture weights $(\pi, 1 - \pi)$ are bounded away from $1/2$, we have that $\rho = |1 - 2\pi|$ is bounded away from zero, and thus population EM iterates converge in $\mathcal{O}(\log(1/\epsilon))$ steps to an ϵ-ball around $\theta^* = 0$. This result is equivalent to showing that in the unbalanced instance ($\pi \neq 1/2$), the log-likelihood is strongly concave around the true parameter.

Statistical rate of sample EM. Once the bounds (10a) and (11)) have been established, the proof of the statistical rate (10b) for sample EM utilizes the scheme laid out by Balakrishnan et al. [1]. In particular, we prove a non-asymptotic uniform law of large numbers (Lemma 1 stated in Section 4.1) that allows for the translation from population to sample EM iterates. Roughly speaking, Lemma 1 guarantees that for any radius $r > 0$, tolerance $\delta \in (0, 1)$, and sufficiently large n, we have

$$\mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c_2\sigma(\sigma r + \rho)\sqrt{\frac{d + \log(1/\delta)}{n}}\right] \geq 1 - \delta. \tag{12}$$

This bound, when combined with the contractive behavior (10a) or equivalently the exponentially fast convergence (11) of the population EM iterates allows us to establish the stated bound (10b). (See, e.g., Theorem 2 in the paper [1].)

Putting the pieces together, we conclude that the sample EM updates converge to an estimate of θ^*—that has Euclidean error of the order $(d/n)^{\frac{1}{2}}$—after a relatively small number of steps that are of the order $\log(n/d)$. Note that this theoretical prediction is verified by the simulation study in Figure 1(b) for the univariate setting ($d = 1$) of the unbalanced mixture-fit. In Figure 3, we present the scaling of the radius of the final EM iterate[3] with respect to the sample size n and the dimension d, averaged over 400 runs of sample EM for various settings of (n, d). Linear fits on the log-log scale in these simulations suggest a rate close to $(d/n)^{\frac{1}{2}}$ as claimed in Theorem 1.

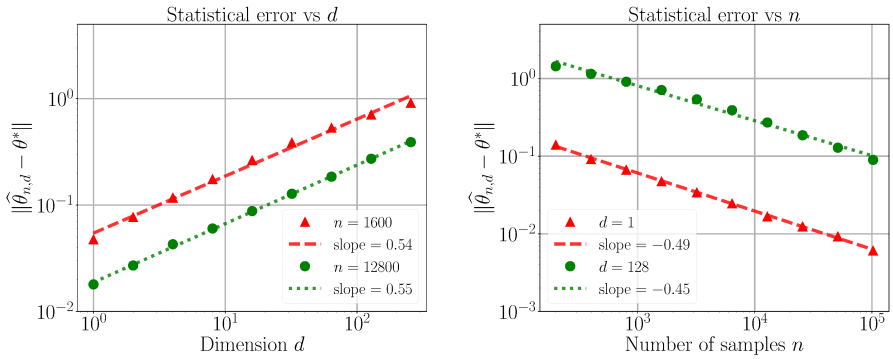


Fig 3. Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ for EM estimates $\widehat{\theta}_{n,d}$ computed using the unbalanced ($\pi \neq \frac{1}{2}$) mixture-fit (3). Here the true data distribution is $\mathcal{N}(0, I_d)$, i.e., $\theta^* = 0$, and $\widehat{\theta}_{n,d}$ denotes the EM iterate upon convergence when we fit a two-mixture model with mixture weights $(0.3, 0.7)$ using n samples in d dimensions. (a) Scaling with respect to d for $n \in \{1600, 12800\}$. (b) Scaling with respect to n for $d \in \{1, 128\}$. We ran experiments for several other pairs of (n, d) and the conclusions were the same. The empirical results here show that that our theoretical upper bound of the order $(d/n)^{\frac{1}{2}}$ on the EM solution is sharp in terms of n and d.

Remark. We make two comments in passing. First, the value of $\left\|\theta^0\right\|_2$ in the convergence rate of sample EM updates in Theorem 1 can be assumed to be of constant order; this assumption stems from the fact the population EM operator maps any θ^0 to a vector with norm smaller than $\sqrt{2/\pi}$ (cf. Lemma 5

[3]Refer to the discussion before Section 2 for details on the stopping rule for EM.

in Appendix C.1). Second, when the weight parameter π is assumed to be unknown in the model fit (3), the EM algorithm exhibits fast convergence when π is initialized sufficiently away from $\frac{1}{2}$; see Section 5.1 for more details.

From unbalanced to balanced fit. The bound (11) shows that the extent of unbalancedness in the mixture weights plays a crucial role in the geometric rate of convergence for the population EM. When the mixtures become more balanced, that is, weight π approaches $1/2$ or equivalently ρ approaches zero, the number of steps T required to achieve ϵ-accuracy scales as $\mathcal{O}\left(\log(\left\|\theta^0\right\|_2/\epsilon)/\rho^2\right)$ and in the limit $\rho \to 0$, this bound degenerates to ∞ for any finite ϵ. Indeed, the bound (10a) from Theorem 1 simply states that the population EM operator is non-expansive for balanced mixtures ($\rho = 0$), and does not provide any particular rate of convergence for this case. It turns out that the EM algorithm is worse in the balanced case, both in terms of the optimization speed and in terms of the statistical rate. This slower statistical rate is in accord with existing results for the MLE in over-specified mixture models [4]; the novel contribution here is the rigorous analysis of the analogous behavior for the EM algorithm.

3.2. *Behavior of EM for balanced mixtures.* In this section, we first provide a sharp characterization of the algorithmic rate of convergence of the population EM update for the balanced fit (see Section 3.2.1). We then provide sharp bound for the statistical rate for the sample EM updates (cf. Section 3.2.2).

3.2.1. *Slow convergence of population EM.* We now analyze the behavior of the population EM operator for the balanced fit. We show that it is globally convergent, albeit with a contraction parameter that depends on θ, and degrades towards 1 as $\|\theta\|_2 \to 0$. Our statement involves the constant $p := \mathbb{P}(|X| \leq 1) + \frac{1}{2}\mathbb{P}(|X| > 1)$, where $X \sim \mathcal{N}(0,1)$ denotes a standard normal variate. (Note that $p < 1$.)

THEOREM 2. *Suppose that we fit a balanced instance ($\pi = \frac{1}{2}$) of the mixture model (3) to $\mathcal{N}(0, \sigma^2 I_d)$ data. Then the population EM operator (8) $\theta \mapsto M(\theta)$ has the following properties:*

(a) For all non-zero θ, we have

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \leq \gamma_{up}(\theta) := 1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2\sigma^2}} < 1. \tag{13a}$$

(b) For all non-zero θ such that $\|\theta\|_2^2 \le \frac{5\sigma^2}{8}$, we have

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \ge \gamma_{low}(\theta) := \frac{1}{1+\frac{2\|\theta\|_2^2}{\sigma^2}}. \tag{13b}$$

See Appendix A.2 in the supplement [10] for the proof of Theorem 2.

The salient feature of Theorem 2 is that the contraction coefficient $\gamma_{\mathrm{up}}(\theta)$ is not globally bounded away from 1 and in fact satisfies $\lim_{\theta\to 0}\gamma_{\mathrm{up}}(\theta) = 1$. In conjunction with the lower bound (13b), we see that

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \asymp \left(1 - \frac{\|\theta\|_2^2}{\sigma^2}\right) \qquad \text{for small } \|\theta\|_2. \tag{14}$$

This precise contraction behavior of the population EM operator is in accord with that of the simulation study in Figure 2(b).

The preceding results show that the population EM updates should exhibit two phases of behavior. In the first phase, up to a relatively coarse accuracy of the order σ, the iterates exhibit geometric convergence. Concretely, we are guaranteed to have $\|\theta^{T_0}\|_2 \le \sqrt{2}\sigma$ after running the algorithm for $T_0 := \frac{\log(\|\theta^0\|_2^2/(2\sigma^2))}{\log(2/(2-p))}$ steps. In the second phase, as the error decreases from $\sqrt{2}\sigma$ to a given $\epsilon \in (0, \sqrt{2}\sigma)$, the convergence rate becomes sub-geometric; concretely, we have

$$\|\theta^{T_0+t}\|_2 \le \epsilon \quad \text{for} \quad t \ge \frac{c\sigma^2}{\epsilon^2}\log(\sigma/\epsilon). \tag{15}$$

Note that the conclusion (15) shows that for small enough ϵ, the population EM takes $\Theta(\log(1/\epsilon)/\epsilon^2)$ steps to find ϵ-accurate estimate of $\theta^* = 0$. This rate is extremely slow compared to the geometric rate $\mathcal{O}(\log(1/\epsilon))$ derived for the unbalanced mixtures in Theorem 1. Hence, the slow rate establishes a qualitative difference in the behavior of the EM algorithm between the balanced and unbalanced setting.

Moreover, the sub-geometric rate of EM in the balanced case is also in stark contrast with the favorable behavior of EM for the exact-fitted settings analyzed in past work. Balakrishnan et al. [1] showed that when the EM algorithm is used to fit a two-component Gaussian mixture with sufficiently large value of $\frac{\|\theta^\star\|_2}{\sigma}$ (known as the high signal-to-noise ratio, or high SNR for short), the population EM operator is contractive, and hence geometrically convergent, within a neighborhood of the true parameter θ^*. In a later work on the two-component balanced mixture fit model, Daskalakis et

al. [7] showed that the convergence is in fact geometric for any non-zero value of the SNR. The model considered in Theorem 2 can be seen as the limiting case of weak signal for a two mixture model—which degenerates to the Gaussian distribution when the SNR becomes exactly zero. For such a limit, we observe that the fast convergence of population EM sequence no longer holds.

3.2.2. *Upper and lower bounds on sample EM.* We now turn to the statements of upper and lower bounds on the rate of the sample EM iterates for the balanced fit on Gaussian data. We begin with an upper bound, which involves the previously defined function $\gamma_{\mathrm{up}}(\theta) := 1 - p + p/\big(1 + \frac{\|\theta\|_2^2}{2\sigma^2}\big)$.

THEOREM 3. *Consider the sample EM updates $\theta^t = M_n(\theta^{t-1})$ for the balanced instance ($\pi = \frac{1}{2}$) of the mixture model (3) based on n i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ samples. Then, there exist universal constants $\{c_k'\}_{k=1}^4$ such that for any scalars $\alpha \in (0, \frac{1}{4})$ and $\delta \in (0,1)$, any sample size $n \geq c_1'(d + \log(\log(1/\alpha)/\delta))$ and any iterate number $t \geq c_2' \log \frac{\|\theta^0\|^2 n}{\sigma^2 d} + c_3' \left(\frac{n}{d}\right)^{\frac{1}{2}-2\alpha} \log(\frac{n}{d}) \log(\frac{1}{\alpha})$, we have*

$$\|\theta^t\|_2 \leq \left[\big\|\theta^0\big\|_2 \cdot \prod_{j=0}^{t-1} \gamma_{up}(\theta^j) \right] + c_4' \sigma \left(\frac{\sigma^2 (d + \log \frac{\log(4/\epsilon)}{\delta})}{n} \right)^{\frac{1}{4}-\alpha}, \tag{16}$$

with probability at least $1 - \delta$.

See Section 4 for a discussion of the techniques employed to prove this theorem. The detailed proof is provided in Appendix A.3 in the supplementary material [10], where we also provide some more details on the definitions of these constants.

As we show in our proofs, once the iteration number t satisfies the lower bound stated in the theorem, the second term on the right-hand side of the bound (16) dominates the first term; therefore, from this point onwards, the the sample EM iterates have Euclidean norm of the order $(d/n)^{\frac{1}{4}-\alpha}$. Note that $\alpha \in (0, \frac{1}{4})$ can be chosen arbitrarily close to zero, so at the expense of increasing the lower bound on the number of iterations t by a logarithmic factor $\log(1/\alpha)$, we can obtain rates arbitrarily close to $(d/n)^{\frac{1}{4}}$.

We note that earlier studies of parameter estimation for over-specified mixtures, in both the frequentist [4] and Bayesian settings [15, 21], have derived a rate of $n^{-\frac{1}{4}}$ for the global maximum of the log likelihood. To the best of our knowledge, Theorem 3 is the first non-asymptotic algorithmic result that shows that such rates apply to the fixed points and dynamics of the EM algorithm, which need not converge to the global optima.

The preceding discussion was devoted to an upper bound on sample EM for the balanced fit. Let us now match this upper bound, at least in the univariate case $d = 1$, by showing that any non-zero fixed point of the sample EM updates has Euclidean norm of the order $n^{-\frac{1}{4}}$. In particular, we prove the following lower bound.

THEOREM 4. *There are universal positive constants c, c' such that for any non-zero solution $\widehat{\theta}_n$ to the sample EM fixed-point equation $\theta = M_n(\theta)$ for the balanced mixture fit, we have*

$$\mathbb{P}\left[|\widehat{\theta}_n| \geq c\, n^{-\frac{1}{4}}\right] \geq c'. \tag{17}$$

See Appendix A.4 in the supplement [10] for the proof of this theorem.

Since the iterative EM scheme converges only to one of its fixed points, the theorem shows that one cannot obtain a high-probability bound for any radius smaller than $n^{-\frac{1}{4}}$. As a consequence, with constant probability, the radius of convergence $n^{-\frac{1}{4}}$ for sample EM convergence in Theorem 3 for the univariate setting is tight.

4. New techniques for sharp analysis of sample EM. In this section, we highlight the new proof techniques introduced in this work that are required to obtain the sharp characterization of the sample EM updates in the balanced case (Theorem 3). We begin in Section 4.1 by elaborating that a direct application of the previous frameworks leads to sub-optimal statistical rates for sample EM in the balanced fit. This sub-optimality motivates the development of new methods for analyzing the behavior of the sample EM iterates, based on an *annulus-based localization argument* over a sequence of epochs, which we sketch out in Sections 4.2 and 4.3. We remark that our novel techniques, introduced here for analyzing EM with the balanced fit, are likely to be of independent interest. We believe that they can potentially be extended to derive sharp statistical rates in other settings when the algorithm under consideration does not exhibit an geometrically fast convergence.

4.1. *A sub-optimal guarantee.* Let us recall the set-up for the procedure suggested by Balakrishnan et al. [1], specializing to the case where the true parameter $\theta^* = 0$, as in our specific set-up. Using the triangle inequality, the norm of the sample EM iterates $\theta^{t+1} = M_n(\theta^t)$ can be upper bounded by a sum of two terms as follows:

$$\left\|\theta^{t+1}\right\|_2 = \left\|M_n(\theta^t)\right\|_2 \leq \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 + \left\|M(\theta^t)\right\|_2 \tag{18}$$

for all $t \geq 0$. The first term on the right-hand side corresponds to the deviations between the sample and population EM operators, and can be controlled via empirical process theory. The second term corresponds to the behavior of the (deterministic) population EM operator, as applied to the sample EM iterate θ^t, and needs to be controlled via a result on population EM.

Theorem 2 from Balakrishnan et al. [1] is based on imposing generic conditions on each of these two terms, and then using them to derive a generic bound on the sample EM iterates. In the current context, their theorem can be summarized as follows. For given tolerances $\delta \in (0,1)$, $\epsilon > 0$ and starting radius $r > 0$, suppose that there exists a function $\varepsilon(n,\delta) > 0$, decreasing in terms of the sample size n, and a contraction coefficient $\kappa \in (0,1)$ such that

$$\text{(19a)} \quad \sup_{\|\theta\|_2 \geq \epsilon} \frac{\|M(\theta)\|_2}{\|\theta\|_2} \leq \kappa \text{ and } \mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon(n,\delta)\right] \geq 1-\delta.$$

Then for a sample size n sufficiently large and ϵ sufficiently small to ensure that

$$\text{(19b)} \quad \epsilon \overset{(i)}{\leq} \frac{\varepsilon(n,\delta)}{1-\kappa} \overset{(ii)}{\leq} r,$$

the sample EM iterates are guaranteed to converge to a ball of radius $\varepsilon(n,\delta)/(1-\kappa)$ around the true parameter $\theta^* = 0$.

In order to apply this theorem to the current setting, we need to specify a choice of $\varepsilon(n,\delta)$ for which the bound on the empirical process holds. The following auxiliary result provides such control for us:

LEMMA 1. *There exists universal positive constants c_1 and c_2 such that for any positive radius r, any threshold $\delta \in (0,1)$, and any sample size $n \geq c_2 d \log(1/\delta)$, we have*

$$\text{(20)} \quad \mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c_1 \sigma(\sigma r + \rho)\sqrt{\frac{d + \log(1/\delta)}{n}}\right] \geq 1 - \delta,$$

where $\rho = |1 - 2\pi|$ denotes the imbalance in the mixture fit (3).

The proof of this lemma is based on Rademacher complexity arguments; see Appendix B.1 in the supplement [10] for the details.

With the choice $r = \|\theta^0\|_2$, Lemma 1 guarantees that the second inequality in line (19a) holds with $\varepsilon(n,\delta) \lesssim \sigma^2 \|\theta^0\|_2 \sqrt{d/n}$. On the other hand,

Theorem 2 implies that for any θ such that $\|\theta\|_2 \geq \epsilon$, we have that population EM is contractive with parameter bounded above by $\kappa(\epsilon) \asymp 1 - \epsilon^2$. In order to satisfy inequality (i) in equation (19b), we solve the equation $\varepsilon(n, \delta)/(1 - \kappa(\epsilon)) = \epsilon$. Tracking only the dependency on d and n, we obtain[4]

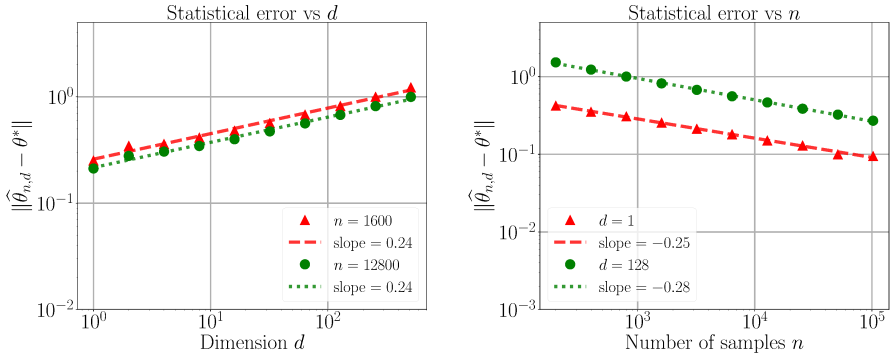
$$\frac{\sqrt{d/n}}{\epsilon^2} = \epsilon \quad \Longrightarrow \quad \epsilon = \mathcal{O}\left((d/n)^{\frac{1}{6}}\right), \tag{21}$$

which shows that the Euclidean norm of the sample EM iterate is bounded by a term of order $(d/n)^{\frac{1}{6}}$.

While this rate is much slower than the classical $(d/n)^{\frac{1}{2}}$ rate that we established in the unbalanced case, it does not coincide with the $n^{-\frac{1}{4}}$ rate that we obtained in Figure 1(b) for balanced setting with $d = 1$. Thus, the proof technique based on the framework of Balakrishnan et al. [1] appears to be non-optimal. The sub-optimality of this approach necessitates the development of a more refined technique. Before sketching this technique, we now quantify empirically the convergence rate of sample EM in terms of both dimension d and sample size n for the balanced mixture fit. In Figure 4, we summarize the results of these experiments. The two panels in the figure exhibit that the error in the sample EM estimate scales as $(d/n)^{\frac{1}{4}}$, thereby providing further numerical evidence that the preceding approach indeed led to a sub-optimal result.

4.2. *Annulus-based localization over epochs.* Let us try to understand why the preceding argument led to a sub-optimal bound. In brief, its "one-shot" nature contains two major deficiencies. First, the tolerance parameter ϵ is used both (a) for measuring the contractivity of the updates, as in the first inequality in equation (19a), *and* (b) for determining the final accuracy that we achieve. At earlier phases of the iteration, the algorithm will converge *more quickly* than suggested by the worst-case analysis based on the final accuracy. A second deficiency is that the argument uses the radius r only once, setting it to a constant to reflect the initialization θ^0 at the start of the algorithm. This means that we failed to "localize" our bound on the empirical process in Lemma 1. At later iterations of the algorithm, the norm $\|\theta^t\|_2$ will be smaller, meaning that the empirical process can be more tightly controlled. We note that ideas of localizing the radius r for an empirical process plays a crucial role in obtaining sharp bounds on the error of M-estimation procedures [2, 17, 26, 27].

[4]Moreover, with this choice of ϵ, inequality (ii) in equation (19b) is satisfied with a constant r, as long as n is sufficiently large relative to d.



(a) (b)

Fig 4. Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ for EM estimates $\widehat{\theta}_{n,d}$ computed using the balanced ($\pi = \frac{1}{2}$) mixture-fit (2). Here the true data distribution is $\mathcal{N}(0, I_d)$, i.e., $\theta^* = 0$, and $\widehat{\theta}_{n,d}$ denotes the EM iterate upon convergence when we fit a balanced mixture with n samples in d dimensions. (a) Scaling with respect to d for $n \in \{1600, 12800\}$. (b) Scaling with respect to n for $d \in \{1, 128\}$. We ran experiments for several other pairs of (n, d) and the conclusions were the same. Clearly, the empirical results suggest a scaling of order $(d/n)^{\frac{1}{4}}$ for the final iterate of sample-based EM.

A novel aspect of the localization argument in our setting is the use of an annulus instead of a ball. In particular, we analyze the iterates from the EM algorithm assuming that they lie within a pre-specicied annulus, defined by an inner and an outer radius. On one hand, the outer radius of the annulus helps to provide a sharp control on the perturbation bounds between the population and sample operators. On the other hand, the inner radius of the annulus is used to tightly control the algorithmic rate of convergence.

We now summarize our key arguments. The entire sequence of sample EM iterations is broken up into a sequence of different epochs. During each epoch, we localize the EM iterates to an annulus. In more detail:

- We index epochs by the integer $\ell = 0, 1, 2, \ldots$, and associate them with a sequence $\{\alpha_\ell\}_{\ell \geq 0}$ of scalars in the interval $[0, \frac{1}{4}]$. The input to epoch ℓ is the scalar α_ℓ, and the output from epoch ℓ is the scalar $\alpha_{\ell+1}$.
- The ℓ-th epoch is defined to be the set of all iterations t of the sample EM algorithm such that the sample EM iterate θ^t lies in the following annulus:

$$\left(\frac{d}{n}\right)^{\alpha_{\ell+1}} \leq \|\theta^t - \theta^*\|_2 \leq \left(\frac{d}{n}\right)^{\alpha_\ell}. \tag{22}$$

We establish that the sample-EM operator is non-expansive so that

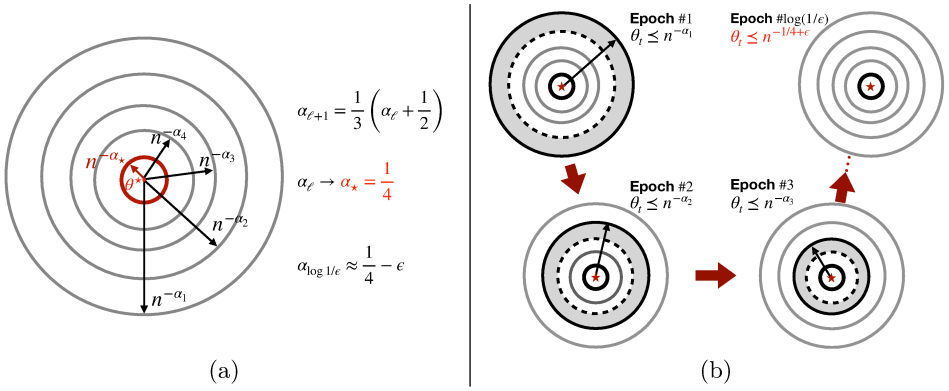


Fig 5. Illustration of the annulus-based-localization argument part (I): Defining the epochs or equivalently the annuli. (a) Outer radius for the ℓ-th epoch is given by $n^{-\alpha_\ell}$ (tracking dependency only on n). (b) For any given epoch ℓ, we analyze the behavior of the EM sequence $\theta^{t+1} = M_n(\theta^t)$, when θ^t lies in the annulus around θ^* with inner and outer radii given by $n^{-\alpha_{\ell+1}}$, and $n^{-\alpha_\ell}$, respectively. We prove that EM iterates move from one epoch to the next epoch (e.g. epoch ℓ to epoch $\ell+1$) after at most $\sqrt{n}$ iterations. Given the definition of α_ℓ, we see that the inner and outer radii of the aforementioned annulus converges linearly to $n^{-\frac{1}{4}}$. Consequently, after at most $\log(1/\alpha)$ epochs (or $\sqrt{n}\log(1/\alpha)$ iterations), the EM iterate lies in a ball of radius $n^{-1/4+\alpha}$ around θ^*. We illustrate the one-step dynamics in any given annulus in Figure 6.

each epoch is well-defined (and that subsequent iterations can only correspond to subsequent epochs).

- Upon completion of epoch ℓ at iteration T_ℓ, the EM algorithm returns an estimate θ^{T_ℓ} such that $\|\theta^{T_\ell}\|_2 \precsim (d/n)^{\alpha_{\ell+1}}$, where

$$\alpha_{\ell+1} = \frac{1}{3}\alpha_\ell + \frac{1}{6}. \tag{23}$$

Note that the new scalar $\alpha_{\ell+1}$ serves as the input to epoch $\ell+1$.

The recursion (23) is crucial in our analysis: it tracks the evolution of the exponent acting upon the ratio d/n, and the rate $(d/n)^{\alpha_{\ell+1}}$ is the bound on the Euclidean norm of the sample EM iterates achieved at the end of epoch ℓ.

A few properties of the recursion (23) are worth noting. First, given our initialization $\alpha_0 = 0$, we see that $\alpha_1 = \frac{1}{6}$, which agrees with the outcome of our one-step analysis from above. Second, as the recursion is iterated, it converges from below to the fixed point $\alpha^* = \frac{1}{4}$. Thus, our argument will allow us to prove a bound arbitrarily close to $(d/n)^{\frac{1}{4}}$, as stated formally

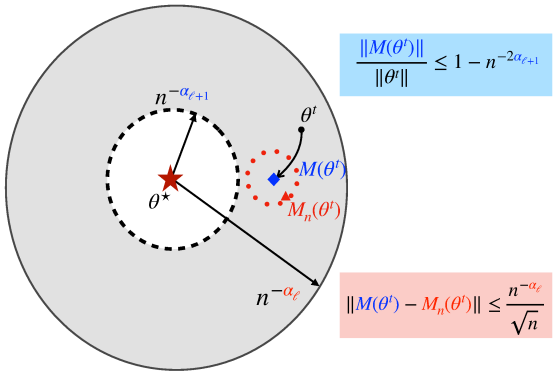


Fig 6. Illustration of the annulus-based-localization argument part (II): Dynamics of EM in the ℓ-th epoch or equivalently the annulus $n^{-\alpha_{\ell+1}} \le \|\theta^t - \theta^*\|_2 \le n^{-\alpha_\ell}$. For a given epoch ℓ, we analyze the behavior of the EM sequence $\theta^{t+1} = M_n(\theta^t)$, when θ^t lies in the annulus with inner and outer radii given by $n^{-\alpha_{\ell+1}}$, and $n^{-\alpha_\ell}$, respectively. In this epoch, the population EM operator $M(\theta^t)$ contracts with a contraction coefficient that depends on $n^{-\alpha_{\ell+1}}$, which is the inner radius of the disc, while the perturbation error $\|M_n(\theta^t) - M(\theta)\|_2$ between the sample and population EM operators depends on $n^{-\alpha_\ell}$, which is the outer radius of the disc. Overall, we prove that M_n is non-expansive and after at most $\sqrt{n}$ steps, the sample EM updates move from epoch ℓ to epoch $\ell + 1$.

in Theorem 3 to follow. Refer to Figures 5 and 6 for an illustration of the definition of these annuli, epochs and the associated conclusions.

4.3. *How does the key recursion* (23) *arise?.* Let us now sketch out how the key recursion (23) arises. Consider epoch ℓ specified by input $\alpha_\ell < \frac{1}{4}$, and consider an iterate θ^t in the following annulus: $\left\|\theta^t\right\|_2 \in [(d/n)^{\alpha_{\ell+1}}, (d/n)^{\alpha_\ell}]$. We begin by proving that this initial condition ensures that $\left\|\theta^t\right\|_2$ is less than level $(d/n)^{\alpha_\ell}$ for all future iterations; for details, see Lemma 4 stated in the supplement. Given this guarantee, our second step is to make use of the inner radius of the considered annulus to apply Theorem 2 for the population EM operator, for all iterations t such that $\left\|\theta^t\right\|_2 \ge (d/n)^{\alpha_{\ell+1}}$. Consequently, for these iterations, we have

$$\begin{aligned} \left\|M(\theta^t)\right\|_2 &\le \left(1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2\sigma^2}}\right) \left\|\theta^t\right\|_2 \\ &\precsim (1 - (d/n)^{2\alpha_{\ell+1}})(d/n)^{\alpha_\ell} \le \widetilde{\gamma} \left(\frac{d}{n}\right)^{\alpha_\ell}, \end{aligned} \tag{24a}$$

where $\widetilde{\gamma} := e^{-(d/n)^{2\alpha_{\ell+1}}}$. On the other hand, using the outer radii of the annulus and applying Lemma 1 for this epoch, we obtain that

$$\big\|M_n(\theta^t) - M(\theta^t)\big\|_2 \precsim \left(\frac{d}{n}\right)^{\alpha_\ell} \sqrt{\frac{d}{n}} = \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2}, \tag{24b}$$

for all t in the epoch. Unfolding the basic triangle inequality (18) for T steps, we find that

$$\begin{aligned}\big\|\theta^{t+T}\big\|_2 &\leq \big\|M_n(\theta^t) - M(\theta^t)\big\|_2 \,(1 + \widetilde{\gamma} + \ldots + \widetilde{\gamma}^{T-1}) + \widetilde{\gamma}^T \,\|\theta_t\|_2 \\ &\leq \frac{1}{1-\widetilde{\gamma}} \big\|M_n(\theta^t) - M(\theta^t)\big\|_2 + e^{-T(d/n)^{2\alpha_{\ell+1}}} (d/n)^{\alpha_\ell}.\end{aligned}$$

The second term decays exponentially in T, and our analysis shows that it is dominated by the first term in the relevant regime of analysis. Examining the first term, we find that θ^{t+T} has Euclidean norm of the order

$$\big\|\theta^{t+T}\big\|_2 \precsim \frac{1}{1-\widetilde{\gamma}} \big\|M_n(\theta^t) - M(\theta^t)\big\|_2 \approx \underbrace{\left(\frac{d}{n}\right)^{-2\alpha_{\ell+1}} \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2}}_{=:\, r}. \tag{25}$$

The epoch is said to be complete once $\big\|\theta^{t+T}\big\|_2 \precsim \big(\frac{d}{n}\big)^{\alpha_{\ell+1}}$. Disregarding constants, this condition is satisfied when $r = \big(\frac{d}{n}\big)^{\alpha_{\ell+1}}$, or equivalently when

$$\left(\frac{d}{n}\right)^{-2\alpha_{\ell+1}} \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2} = \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}.$$

Viewing this equation as a function of the pair $(\alpha_{\ell+1}, \alpha_\ell)$ and solving for $\alpha_{\ell+1}$ in terms of α_ℓ yields the recursion (23). Refer to Figure 6 for a visual illustration of the localization argument summarized above for a given epoch.

Of course, the preceding discussion is informal, and there remain many details to be addressed in order to obtain a formal proof. We refer the reader to Appendix A.3 for the complete argument.

5. Generality of results and future work. Thus far, we have characterized the behavior of the EM algorithm for different settings of over-specified location Gaussian mixtures. We established rigorous statistical guarantees of EM under two particular but representative settings of over-specified location Gaussian mixtures: the balanced and unbalanced mixture-fit. The log-likelihood for the unbalanced fit remains strongly log-concave[5]

[5]Moreover, in Appendix D we differentiate the unbalanced and balanced fit based on the log-likelihood and the Fisher matrix and provide a heuristic justification for the different rates between the two cases.

(due to the fixed weights and location parameters being sign flips) and hence the Euclidean error of the final iterate of EM decays at the usual rate $(d/n)^{\frac{1}{2}}$ with n samples in d dimensions. However, in the balanced case, the log-likelihood is no longer strongly log-concave and the error decays at the slower rate $(d/n)^{\frac{1}{4}}$. We view our results as the first step in understanding and possibly improving the EM algorithm in non-regular settings. We now provide a detailed discussion that sheds light on the general applicability of our results. In particular, we discuss the behavior of EM under the following settings: (i) over-specified mixture models with unknown weight parameters (Section 5.1), (ii) over-specified mixture of linear regression (Section 5.2), and (iii) more general settings with over-specified mixture models (Section 5.3). We conclude the paper with a discussion of several future directions that arise from the previous settings in Section 5.4.

5.1. *When the weights are unknown.* Our theoretical analysis so far assumed that the weights were fixed, an assumption common to a number of previous papers in the area [1, 7, 18]. In Appendix C.2, we consider the case of unknown weights for the model fit (3). In this context, our main contribution is to show that if the weights are initialized far away from $\frac{1}{2}$—meaning that the initial mixture is highly unbalanced—then the EM algorithm converges quickly, and the results from Theorem 1 are valid. (See Lemma 6 in Appendix C.2 for the details.) On the other hand, if the initial mixture is not heavily imbalanced, we observe the slow convergence of EM consistent with Theorems 2 and 3.

5.2. *Slow rates for mixture of regressions.* Thus far, we have considered the behavior of the EM algorithm in application to parameter estimation in mixture models. Our findings turn out to hold somewhat more generally, with Theorems 2 and 3 having analogues when the EM algorithm is used to fit a mixture of linear regressions in over-specified settings. Concretely, suppose that $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ are i.i.d. samples generated from the model

$$Y_i = X_i^\top \theta^* + \sigma \xi_i, \qquad \text{for } i = 1, \ldots, n, \tag{26}$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. standard Gaussian variates, and the covariate vectors $X_i \in \mathbb{R}^d$ are also i.i.d. samples from the standard multivariate Gaussian $\mathcal{N}(0, I_d)$. Of interest is to estimate the parameter θ^* using these samples and EM is a popular method for doing so. When θ^* has sufficiently large Euclidean norm, a setting referred to as the strong signal case, Balakrishnan et al. [1] showed that the estimate returned by EM is at a distance $(d/n)^{\frac{1}{2}}$ from

the true parameter θ^* with high probability. On the other hand, our analysis shows that when $\|\theta^*\|_2$ decays to zero—leading to an over-specified setting—the convergence of EM becomes slow. In particular, the EM algorithm takes significantly more steps and returns an estimate that is statistically worse, lying at Euclidean distance of the order $(d/n)^{\frac{1}{4}}$ from the true parameter. While the EM operators in this case are slightly different when compared to the over-specified Gaussian mixture analyzed before, the proof techniques remain similar. More concretely, we first show that the convergence of population EM is slow (similar to Theorem 2) and then use the annulus-based localization argument (similar to the proof of Theorem 3 from Section 4) to derive a sharp rate. For completeness, we present these results formally in Lemma 7 and Corollary 2 in Appendix E.

5.3. *Slow rates for general mixtures.* We now present several experiments that provide numerical backing to the claim that the slow rate of order $n^{-\frac{1}{4}}$ is not merely an artifact of the special balanced fit ((3) with $\pi = \frac{1}{2}$). We demonstrate that the slow convergence of EM is very likely to arise while fitting general over-specified location Gaussian mixtures with unknown weights (and known covariance). We consider three settings: (A) fitting several general over-specified location Gaussian mixture fits to Gaussian data (Figure 7), (B) fitting a special three-component mixture fit to a two mixture of Gaussians (Figure 8), and (C) fitting mixtures with unknown weights and location parameters when the number of components in the fitted model is over-specified by two (Figure 9). We now turn to the details of these settings.

General over-specified mixture fits on Gaussian data. First, we remark that the fast convergence in the unbalanced fit (Theorem 1) was a joint result of the facts that (a) the weights were fixed and unequal, and (b) the parameters were constrained to be a sign flip. If either of these conditions is violated, the EM algorithm exhibits slow convergence on both algorithmic and statistical fronts. Theorems 2, 3 and 4 provide rigorous details for the case of equal and fixed weights (balanced fit). When the weights are unknown, EM can exhibit slow rate (see Section 5.1 and Appendix C.2 for further details). When the weights are fixed and unequal, but the location parameters are estimated freely—that is, with the model $\pi\phi(x;\theta_1,1)+(1-\pi)\phi(x;\theta_2,1)$, as illustrated in Figure 7(a)—then the EM estimates have error[6] of order $n^{-\frac{1}{4}}$.

[6]For more general cases, we measure the error of parameter estimation using the Wasserstein metric of second order $\widehat{W}_{2,n}$ to account for label-switching between the components. When the true model is standard Gaussian this metric is simply the weighted Euclidean error: $(\sum \pi_k \widehat{\theta}_{k,n}^2)^{\frac{1}{2}}$, where π_k and $\widehat{\theta}_{k,n}$, respectively, denote the mixture weight

In such cases, the parameter estimates approximately satisfy the relation $\sum_k \pi_k \widehat{\theta}_{k,n} \approx 0$, since the mean of the data is close to zero; moreover, for a two-components mixture model, the location estimates become weighted sign flips of each other. The features are the intuitive reason underlying the similarity of behavior of EM between this fit and the balanced fit. Finally, when we fit a two mixture model with unknown weight parameter and free location parameters, the final error also has a scaling of order $n^{-\frac{1}{4}}$. Refer to Figure 7 for a numerical validation of these results.

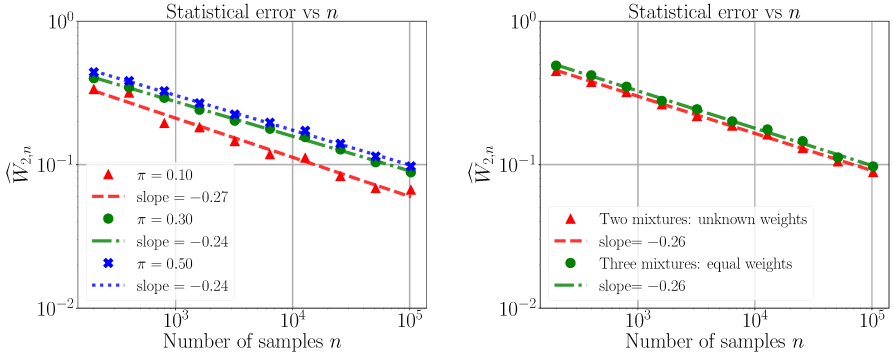


Fig 7. Plots of the Wasserstein error $\widehat{W}_{2,n}$ associated with EM fixed points versus the sample size for fitting various kinds of location mixture models to standard normal $\mathcal{N}(0,1)$ data. We fit mixture models with either two or three components, with all location parameters estimated in an unconstrained manner. The lines are obtained by a linear regression of the log error on the sample size n. (a) Fitting a two-mixture model $\pi\mathcal{N}(\theta_1,1) + (1-\pi)\mathcal{N}(\theta_2,1)$ with three different *fixed* values of weights $\pi \in \{0.1, 0.3, 0.5\}$ and two (unconstrained) location parameters, along with least-squares fits to the log errors. (b) Data plotted as red triangles is obtained by fitting a two-component model with *unknown* mixture weights and two location parameters $\pi\mathcal{N}(\theta_1,1) + (1-\pi)\mathcal{N}(\theta_2,1)$, whereas green circles correspond to results fitting a three-component mixture model $\sum_{i=1}^{3} \frac{1}{3}\mathcal{N}(\theta_i,1)$. In all cases, the EM solutions exhibit the slow $n^{-\frac{1}{4}}$ statistical rate for the error in parameter estimation. Also see Figure 9.

Over-specified fits for mixtures of Gaussian data. Using similar reasoning as above, let us sketch out how our theoretical results also yield usable predictions for more general over-specified models. Roughly speaking, whenever there are extra number of components to be estimated, parameters of some of them are likely to end up satisfying certain form of local constraint. More

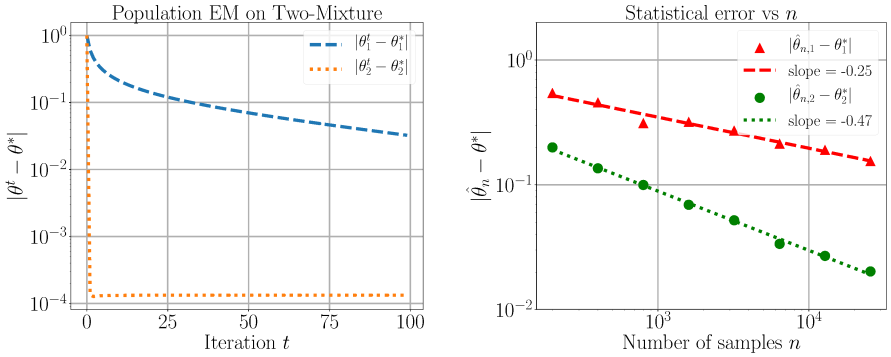
and the location parameter of the k-th component of the mixture.

concretely, suppose that we are given data generated from a k-component mixture, and we use the EM algorithm to fit the location parameters of a mixture model with $k+1$ components. Loosely speaking, the EM estimates corresponding to a set of $k-1$ components are likely to converge quickly, leaving the two remaining components to fit a single component in the true model. If the other components are far away, the EM updates for the parameters of these two components are unaffected by them and start to behave like the balanced case. See Figure 8 for a numerical illustration of this intuition in an idealized setting where we use $k+1=3$ components to fit data generated from a $k=2$ component model. In this idealized setting, the error for one of the parameter scales at the fast rate of order $n^{-\frac{1}{2}}$, and that of the parameter that is locally over-fitted exhibits a slow rate of order $n^{-\frac{1}{4}}$. Finally, we see that the statistical error of order $n^{-\frac{1}{4}}$ also arises when we over-specify the number of components by more than one. In particular, we observe in Figure 7(b) (green dashed dotted line with solid circles) and Figure 9 (both curves) that a similar scaling of order $n^{-\frac{1}{4}}$ arises when we over-specify the number of components by 2 and estimate the weight and location parameters.

Besides formally analyzing EM in these general cases, several other future directions arise from our work which we now discuss.

5.4. *Future directions.* In our current work, we assumed that only the location parameters were unknown and that the scale parameters of the underlying model are known. Nevertheless in practice, this assumption is rather restrictive and it is natural to ask what happens if the scale parameters were also unknown. We note that the MLE is known to have even slower statistical rates for the estimation error with such higher-order mixtures; therefore, it would be interesting to determine if the EM algorithm also suffers from a similar slow down when the scale parameters are unknown. We refer the readers to a recent preprint [11], where we establish that the EM algorithm can suffer from a further slow-down on the statistical and computational ends when over-specified mixtures are fitted with an unknown scale parameter.

Another important direction is to analyze the behavior of EM under different models for generating the data. While our analysis is focused on Gaussian mixtures, the non-standard statistical rate $n^{-\frac{1}{4}}$ also arises in other types of over-specified mixture models, such as those involving mixtures with other exponential family members, or Student-t distributions, suitable for heavy-tailed data. We believe that the analysis of our paper can be generalized to a broader class of finite mixture models that includes the aforementioned



(a) (b)

Fig 8. Behavior of EM for an over-specified Gaussian mixture. True model: $\frac{1}{2}\mathcal{N}(\theta_1^*, 1) + \frac{1}{2}\mathcal{N}(\theta_2^*, 1)$ where $\theta_1^* = 0$ and $\theta_2^* = 10$. We fit a model $\frac{1}{4}\mathcal{N}(-\theta_1, 1) + \frac{1}{4}\mathcal{N}(\theta_1, 1) + \frac{1}{2}\mathcal{N}(\theta_2, 1)$, where we initialize θ_1^0 close to θ_1^* and θ_2^0 close to θ_2^*. (a) Population EM updates: We observe that while θ_1^t converges slowly to $\theta_1^* = 0$, the iterates θ_2^t converge exponentially fast to $\theta_2^* = 10$. (b) We plot the statistical error for the two parameters. While the strong signal component has a parametric $n^{-\frac{1}{2}}$ rate, for the no signal component EM has the slower $n^{-\frac{1}{4}}$ rate, which is in good agreement with the theoretical results derived in the paper. (We remark that the error floor for θ_2^t in panel (a) arises from the finite precision inherent to numerical integration.)

models.

A final direction of interest is whether the behavior of EM—slow versus fast convergence—can be used as a statistic in a classical testing problem: testing the simple null of a standard multivariate Gaussian versus the compound alternative of a two-component Gaussian mixture. This problem is known to be challenging due to the break-down of the (generalized) likelihood ratio test, due the singularity of the Fisher information matrix; see the papers [6, 19] for some past work on the problem. The results of our paper suggest an alternative approach, which is based on monitoring the convergence rate of EM. If the EM algorithm converges slowly for a balanced fit, then we may accept the null, whereas the opposite behavior can be used as an evidence for rejecting the null. We leave for future work the analysis of such a testing procedure based on the convergence rates of EM.

Acknowledgments. This work was partially supported by Office of Naval Research grant DOD ONR-N00014-18-1-2640 and National Science Foundation grant NSF-DMS-1612948 to MJW, and National Science Foundation grant NSF-DMS-1613002 to BY, and by Army Research Office grant W911NF-17-1-0304 to MIJ.

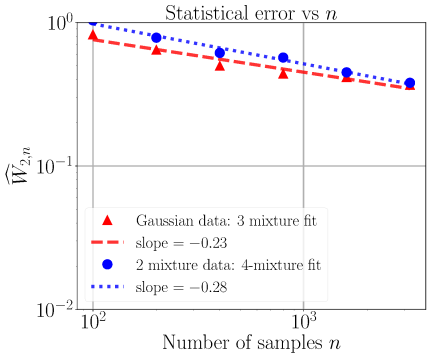


Fig 9. Plots of Wasserstein error when both weights and location parameters are unknown and estimated using EM and the fitted multivariate mixture model is over-specified. (a) True model: $\mathcal{N}([0,0]^\top, I_2)$, and fitted model $\sum_{i=1}^{3} w_i \mathcal{N}(\theta_i, I_2)$ and (b) True model: $\frac{2}{5}\mathcal{N}([0,0]^\top, I_2) + \frac{3}{5}\mathcal{N}([4,4]^\top, I_2)$ and fitted model: $\sum_{i=1}^{4} w_i \mathcal{N}(\theta_i, I_2)$. In both cases, once again we see the scaling of order $n^{-\frac{1}{4}}$ for the final error (similar to results in Figure 7 and 8).

SUPPLEMENTARY MATERIAL

Supplement to "Singularity, Misspecification, and the Convergence Rate of EM"
(doi: COMPLETED BY THE TYPESETTER; .pdf). In this supplemental material we present the proofs for all the technical results in the paper.

References.

[1] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* **45** 77-120. (Cited on pages 3, 4, 7, 8, 12, 13, 15, 17, 18, 19, and 24.)

[2] Bartlett, P. L., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics* **33** 1497–1537. (Cited on pages 5 and 19.)

[3] Cai, T. T., Ma, J., Zhang, L. et al. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics* **47** 1234–1267. (Cited on page 3.)

[4] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics* **23** 221-233. (Cited on pages 2, 5, 14, and 16.)

[5] Chen, Y. C. (2018). Statistical inference with local optima. *arXiv preprint arXiv:1807.04431*. (Cited on page 3.)

[6] Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *The Annals of Statistics* **37** 2523-2542. (Cited on page 28.)

[7] Daskalakis, C., Tzamos, C. and Zampetakis, M. (2017). Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*. (Cited on pages 3, 4, 7, 16, and 24.)

[8] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39** 1-38. (Cited on page 3.)

[9] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J. and Jordan, M. I. (2018a). Theoretical guarantees for EM under misspecified Gaussian mixture models. In *NeurIPS 31*. (Cited on page 4.)

[10] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I. and Yu, B. (2018b). Supplement to "Singularity, misspecification and the convergence rate of EM". (Cited on pages 5, 12, 15, 16, 17, and 18.)

[11] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I. and Yu, B. (2019). Challenges with EM in application to weakly identifiable mixture models. *arXiv preprint arXiv:1902.00194*. (Cited on page 27.)

[12] Ghosal, S. and van der Vaart, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29** 1233-1263. (Cited on page 2.)

[13] Hao, B., Sun, W., Liu, Y. and Cheng, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research* **18** 1 - 58. (Cited on page 3.)

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

[15] Ishwaran, H., James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association* **96** 1316-1332. (Cited on page 16.)

[16] Klusowski, J. M., Yang, D. and Brinda, W. (2019). Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory* **65** 3515–3524. (Cited on page 3.)

[17] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* **34** 2593–2656. (Cited on pages 5 and 19.)

[18] Kumar, R. and Schmidt, M. (2017). Convergence rate of expectation-maximization. *10th NIPS Workshop on Optimization for Machine Learning*. (Cited on pages 3 and 24.)

[19] Li, P., Chen, J. and Marriott, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika* **96** 411-426. (Cited on page 28.)

[20] Ma, J., Xu, L. and Jordan, M. I. (2000). Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation* **12** 2881-2907. (Cited on page 3.)

[21] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **4** 370-400. (Cited on pages 2 and 16.)

[22] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review* **26** 195–239. (Cited on page 3.)

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

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

[25] Stephens, M. (2002). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** 795-809. (Cited on page 2.)

[26] VAN DE GEER, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press. (Cited on pages 2, 5, and 19.)
[27] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, UK. (Cited on page 19.)
[28] WANG, Z., GU, Q., NING, Y. and LIU, H. (2015). High-dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems 28*. (Cited on page 3.)
[29] WU, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11** 95-103. (Cited on page 3.)
[30] XU, J., HSU, D. and MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*. (Cited on page 3.)
[31] XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM Algorithm for Gaussian mixtures. *Neural Computation* **8** 129-151. (Cited on page 3.)
[32] YAN, B., YIN, M. and SARKAR, P. (2017). Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems 30*. (Cited on pages 3 and 4.)
[33] YI, X. and CARAMANIS, C. (2015). Regularized EM algorithms: a unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems 28*. (Cited on page 3.)

RAAZ DWIVEDI
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES
E-MAIL: raaz.rsk@berkeley.edu

KOULIK KHAMARU
DEPARTMENT OF STATISTICS
E-MAIL: koulik@berkeley.edu

MICHAEL I. JORDAN
DEPARTMENT OF STATISTICS
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES
E-MAIL: jordan@cs.berkeley.edu

NHAT HO
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES
E-MAIL: minhnhat@berkeley.edu

MARTIN J. WAINWRIGHT
DEPARTMENT OF STATISTICS
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES
E-MAIL: wainwrig@berkeley.edu

BIN YU
DEPARTMENT OF STATISTICS
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES
E-MAIL: binyu@berkeley.edu