We thank the reviewers for their insightful comments. Below we prioritize what we view as the most important points.

Motivation and Contributions

EM is the quintessential approach for mixture problems. Despite its long history and popularity, theoretical understanding of EM is disappointingly limited and largely lags its empirical application. Even in the simplest setting of 2 Gaussians (2GMM), global convergence was only established recently. We establish such a guarantee for a much more general class of distributions. In particular, *we identify the crucial role of log-concavity—rather than Gaussianity—in ensuring global convergence*, while recent related work considers specific distributions (e.g. Laplace or regression with Gaussian noise) on a case-by-case basis.

Compared to prior work, we overcame two main technical challenges: 1) We need to establish the angle shrinkage property of the LS-EM algorithm. This is contrastingly different from the working mechanism (shrinkage in distance) of classical EM for 2GMM; existing work (e.g. Balakrishnan, and Daskalakis) heavily relies on this mechanism, which does *not* work for general log-concave mixtures. 2) Unlike Gaussian, for general log-concave distributions in the high dimension, the coordinates are dependent of each other even when the covariance matrix is identity; therefore, a more sophisticated sensitivity analysis (Lemma D.1) is needed to establish angle shrinkage.

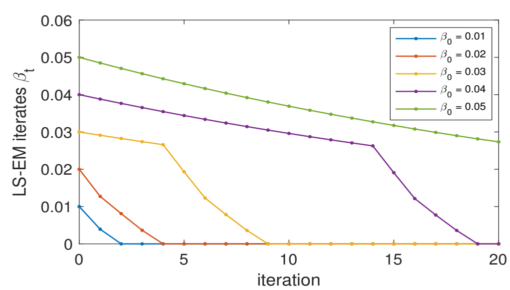
Compared to tensor methods, EM is much simpler, and manifests *linear* dependency on the dimension d in terms of time and sample complexities (as opposed to polynomial for tensor methods).

Assumptions

As Reviewer 1 pointed out, our analysis is built upon several assumptions (a balanced mixture of 2 distributions with same covariance). We note that these assumptions are common among recent literature on global convergence of EM. Moreover, there exist examples of more general mixtures where global convergence is impossible, including mixture of 3 Gaussians (see arXiv 1609.00978) or 2 unbalanced Gaussians (see arXiv 1810.11344). Additional comments below:

(i) Log-concavity is crucial. In our analysis, it guarantees that certain derivatives of the LS-EM operator are non-decreasing.

In our experiment, if the ground-truth distribution is not log-concave, with symmetric density function $\frac{1}{2}1_{0<x\le 0.5}+\frac{1}{8}1_{0.5<x<1.5}+\frac{1}{32}1_{1.5\le x<3.5}+\frac{1}{128}1_{3.5\le x<7.5}+\frac{1}{512}1_{7.5\le x<15.5}$ (defined symmetrically on the negative side), then LS-EM *incorrectly* converges to 0 when the initial solution β_0 is close to 0 (see figure on the right).



(ii) For model misspecification, we focus on fitting with Gaussian, as it is a typical and reasonable choice in practice. Experiments show that if the fitted distribution is not Gaussian, LS-EM may fail completely (Fig. 6 in Appendix of original paper).

(iii) The assumption of symmetric parameters $\pm\beta^*$ is just a form of centering and hence *not* essential to our results.

It is definitely an intriguing problem to figure out the exact setting in which global convergence can be achieved, or how one can modify the standard EM algorithm to avoid spurious fixed points. Recent work in arXiv 1810.11344 demonstrated that over-paramerized EM[1] converges globally for unbalanced (but still symmetric) 2GMM. By considering general log concave distributions, we view our work as another step towards a more complete theory.

Experiments

We totally agree on the importance of experiments. We did present numerical experiments for i) robustness under model misspecification, and ii) connection between LS-EM and classical EM, in the appendix due to page limit. Additional experiments (including the one above), which will be included in the final version, corroborate our theoretical findings and further explore the issues of model misspecification and sensitivity to the assumptions, as reviewer 1 suggested.

Other Comments

Comparison with Balakrishnan's paper [BWY17]: When specialized to 2GMM in 1-D, our results match those in BWY17. For higher dimensions, we establish global convergence (the analysis in BWY17 is local); while we currently do not have explicit bounds on the convergence rate and sample complexity, we expect they would again match BWY17.

Regularity Conditions: These conditions are explicitly explained in Section E of the appendix. They are indeed about differentiation under integral sign (as pointed out by Reviewer 2), and are satisfied by common log-concave mixtures.

SNR Condition in Proposition 6.1: Reviewer 3 mentioned that this condition is strong. Note that Proposition 6.1 is a simplified version of the more general result established in the proof of the proposition (pp. 32). There we derived the following bound for the distance between the misspecified solution and the true solution: $|\overline{\beta}-\beta^*|\le\frac{9\sigma}{1-\exp(-0.125\eta^2)}$. This bound holds for any SNR $\eta>0$; when the SNR is smaller, the bound becomes worse, as natural.

[1]It considers the weight parameter as a variable in the EM algorithm, even though the weight is known apriori