We would like to thank three reviewers for their feedback. Upon acceptance, we will include in the final version (a) *new local linear convergence results for fiEM method*, (b) *an improved presentation of main results* and (c) *missing references*. We first discuss a few common concerns shared by **reviewer 1**, **reviewer 2**, **reviewer 3**.

● ● **Local Linear Convergence of fiEM**: As observed by the reviewers, empirically fiEM shows a local linear convergence similar to sEM-VR. We found that fiEM has local linear convergence **in theory**, too. The new analysis requires same assumptions as [Thm.1,Chen+2018] and adopts proof of [Defazio+2014]. We show $\mathbb{E}\|\hat{\mathrm{s}}^{(k)} - \mathrm{s}^\star\|^2 \leq (1-\delta)^k \|\hat{\mathrm{s}}^{(0)} - \mathrm{s}^\star\|^2$ for $k \geq 0$ with $\delta = \Theta(1/n)$, where $\mathrm{s}^\star$ is a stationary point to (19).

● ● **Satisfaction of Assumptions**: *All* assumptions H1-H5 are verified rigorously in the GMM, pLSA applications presented, as proven in Appendix G. They are mild even though should be checked on a case-by-case basis. Reviewers are referred to [McLachlan&Krishnan 2007] which shows satisfaction of similar assumptions on a variety of applications.

● ● **Clarity:** We admit it is a challenging task to present all technical results within the page limit, but we will try our best to improve in the final version, viz. using a running example to illustrate the assumptions used and implementation of algorithms. We will also clarify about the expectation operators in theorems and correct typos.

Reviewer 1: We thank the reviewer for valuable comments and references. Our point-to-point response is as follows:

Related work: The paper [Karimi+2019] is relevant and will be included. Thank you for bringing it to our attention. Karimi+[2019] focused on a biased stochastic approximation scheme and gave a global convergence rate for sEM. In this case, their analysis shares similar scaled gradient interpretation as fiEM and sEM-VR, yet w/o variance reduction.

iEM's Rate: You are right as the rate of iEM is comparable to GD. Yet, iEM is a popular method without a previously known global rate. Indeed, the comparison of iEM to fiEM, sEM-VR (theoretical & empirical) is our main contribution.

Comparison to [Chen+2018]: Our assumptions are more practical and less restrictive. Global convergence to stationary point for sEM-VR in [Thm.2, Chen+2018] assumes i) the sufficient statistics $\mathrm{s}_i(\overline{\boldsymbol{\theta}}(\boldsymbol{s}'))$ is L_s-Lipschitz continuous in $\boldsymbol{s}'$, $\forall i$ – this is implied by our H1-H5 via Lemma 4; ii) the complete log-likelihood is strongly concave – this is slightly relaxed in our H4 which only requires a unique global minimizer for the complete log-likelihood. Besides, H1-H5 are **directly verifiable** (as explained above) and we provide the rate towards a stationary point. Lastly, local convergence in [Thm.1, Chen+2018] requires $\|\hat{\mathrm{s}}^{(k)} - \mathrm{s}^\star\|$ to be in a ball of radius $\mathcal{O}(1/L_s)$ for **any** $k \geq 1$. This is a strong assumption that is not directly verifiable even if $\hat{\mathrm{s}}^{(0)} \approx \mathrm{s}^\star$ is known a-priori.

Reviewer 2: We thank the reviewer for useful comments. Please find the comparison of fiEM, sEM-VR below:

Comparing fiEM to sEM-VR: This comparison is analogous to comparing SAGA to SVRG for finite sum optimization, and there is no clear ordering. In short, it depends on the trade-off of memory imprint and computation complexity. sEM-VR requires $\mathcal{O}(\dim(\mathsf{S}))$ space to store $\overline{\mathrm{s}}^{(\ell(k))}$, yet a *full pass* on the data set is needed at each epoch, resulting in higher complexity; meanwhile, fiEM only processes the data set *incrementally*, but it requires $\mathcal{O}(n\dim(\mathsf{S}))$ to store the variables involved. We remark that the global rate for sEM-VR is **not proven** in [Chen+2018]. In Fig. 2 we show fiEM outperformed sEM-VR in one dataset (a bigger one) but was outperformed by sEM-VR in the other (a smaller one).

Reviewer 3: We thank the reviewer for the comments. We clarify that *in addition* to analyzing iEM using the MISO framework (which will be mentioned explicitly), we analyzed fiEM, sEM-VR with a **completely different** framework w/ **scaled gradient**, the latter constitutes our main contribution of fast global convergence rates; see p.2 of our paper.

Global Convergence & H4: We emphasize H4 **does not** imply that every stationary point of (1) is global minima, as having a unique global minimizer *does not* imply *any first order critical point is global minima*. Also, H4 refers to *complete log-likelihood* $L(\boldsymbol{s}, \boldsymbol{\theta})$ with fixed $\boldsymbol{s}$, instead of incomplete log-likelihood $\overline{\mathcal{L}}(\boldsymbol{\theta})$ in (1). It holds for most exponential family models where EM is useful [McLachlan&Krishnan 2007]. Mind that $\overline{\mathcal{L}}(\boldsymbol{\theta})$ is non-convex and our convergence is *global* in the sense that it does not restrict the initialization, a common assumption for analysis of EM.

Bounds in theorems: The current presentation style of theorems, which evaluates the gradient norm of a randomly terminated stochastic EM solution, is common in **stochastic non-convex** optimization e.g., [Ghadimi&Lan 2013,Reddi+2016a/b]. Part of the reason is that it results in a practical solution. While picking the best iterate leads to the same sublinear rate as ours, doing so involves a full pass on the data ($\nabla\overline{\mathcal{L}}$) and computing the incomplete likelihood, both are **difficult** tasks avoided in stochastic EM methods. Besides, as the reviewer mentioned, both random termination and best iterate schemes lead to a quantity upper bounded by $\sum_k \mathbb{E}\|\nabla\overline{\mathcal{L}}(\boldsymbol{\theta}^{(k)})\|^2/K_{\mathsf{max}}$. This quantity is not equal to the *averaged iterate*, and upper bounding it by $\mathcal{O}(1/K_{\mathsf{max}})$ is a non-trivial task – and is precisely our main contribution.

Theorem 1 of Paper 1613: Indeed, Theorem 1 for iEM is a special case of [Thm. 1, 1613]. We will cite the latter properly. Our main contribution here lies on *fast convergence* of fiEM, sEM-VR shown by a different framework, see Theorem 2. Detailed comments about the difference between this paper and 1613 has been sent to the AC.