We would like to thank the four reviewers for their feedback. We first discuss the common concern about our contributions and novelty shared by **reviewer 2, reviewer 3** as follows.

•• **Key Contributions & Novelty**: While the algorithm design idea of combining SPIDER and EM can be natural, this paper involves a significant work in the convergence analysis technique and gives new insights to deriving future stochastic EM algorithms. We highlight four particular challenges and contributions: (A) **EM is not a gradient-based algorithm**, while SPIDER is designed for accelerating gradient algorithm. The convergence analysis of EM-based algorithms can not be deduced from that of gradient-based procedures: they require a specific study. As a key observation to outline this difference, note that EM can be equivalently studied in the parameter space and in the expectation space - which is not the case for the gradient methods. (B) Among the incremental-EM techniques, we provide **state-of-the art** complexity bounds that overpass all the previous ones. (C) Contrary to previous incremental EM algorithms [6,15], the approximate mean field H_{k+1} is **biased** except at the beginning of the epoch (Lemma 4, Corollary 5). A main novelty of the proof is therefore to show that the squared error when approximating the mean field, is of the same order as the squared of the approximation (Proposition 6) - contrary to previous EM-based proofs, the usual martingale increment property no longer holds, thus requiring a novel approach for the control of this error. (D) We provide **a new perspective** to interpret SPIDER-EM by showing its equivalence to a perturbed Online-EM where the perturbation acts as a control variate in order to reduce the variance (see Section 5).

Reviewer 1. We thank the reviewer for the careful review of our paper. Please find our response as follows.

Linear Convergence: The reviewer has made a mindful observation. Indeed, in Fig. 3, SPIDER-EM appears to converge linearly, i.e., faster than our theory's prediction. We speculate that this is due to a local strong convexity condition of $\mathrm{W}(\cdot)$ around a local minimum statistics $\hat{s}$, after SPIDER-EM reaches a neighborhood of $\hat{s}$. This condition was never made in our analysis, however it has been used as an assumption in sEM-vr *without verification*. Nevertheless, under this condition, we can show local linear convergence for SPIDER-EM. The proof idea will be included in the final version. Lastly, Balakrishnan et al. consider a non-incremental, first order EM that optimizes θ via gradient ascent, this is different from SPIDER-EM that operates on the statistics ($\hat{S}$), this reference will be included in the final version.

Complexity Guarantee with $\mathsf{b} = \mathcal{O}(1)$: This can be easily derived from Theorem 2. If $\mathsf{b} \asymp n^{\mathsf{a}}$, $k_{in} \asymp n^{\mathsf{c}}$ in the case $\mathsf{c} \geq \mathsf{a} \geq 0$, then the overall complexity will be $K_{CE}(n,\epsilon) = n + \mathcal{O}(n^{\max\{\frac{\mathsf{a}+\mathsf{c}}{2}, 1-\frac{\mathsf{a}+\mathsf{c}}{2}\}}\epsilon^{-1})$, $K_{\mathrm{Opt}}(n,\epsilon) = \mathcal{O}(n^{\frac{\mathsf{c}-\mathsf{a}}{2}}\epsilon^{-1})$. Setting $\mathsf{b} = \mathcal{O}(1)$ or equivalently $\mathsf{a} = 0$ will lead to a tradeoff in complexity in terms of the number of parameter updates and of conditional expectations computed. We will include a discussion for these general settings.

Comparison to Karimi et al.: The quasi-gradient interpretation of the E-step updates (see Proposition 1 and H5) is similar to those from Karimi et al. which analyzed the online EM, and the same technique has also been used in [8,15]. Our analysis are different as there are additional complexity due to the *biased, variance reduced* estimator, note that this improves the convergence rate from $\mathcal{O}(1/\epsilon^2)$ (online EM) to $\mathcal{O}(\sqrt{n}/\epsilon)$ (SPIDER-EM). That said, Karimi et al. is relevant to this work, which we will discuss in the final version.

Second Order Stationary (SOS) point: This is an interesting question. The difficulty in applying the analysis for SOS point from [10] lies on the non-gradient update nature of SPIDER-EM. In fact, the procedures in Sec. 3.3. from [10] hinge on the availability of approximate Hessian. To our best knowledge, the convergence to SOS point **has not been studied** in the context of incremental EM. A possible solution is to apply the Louis's missing info. principle [8] to estimate the Hessian of the log-likelihood, and thus diagnosing if the stationary point is SOS. We will include a discussion about the challenge in the final version.

Reviewer 2: Thank you for the careful review of our paper. As per your suggestion, we will include a table in the final paper that compares the complexity results to existing works [also see the response on **Complexity with** $\mathsf{b} = \mathcal{O}(1)$].

We emphasize that our contributions are beyond *an application of SPIDER on EM algorithm*. Instead, as explained in the beginning of this rebuttal, our analysis involve non-trivial proof techniques to bound the error of approximation to the mean field (non-gradient) update of SPIDER-EM. This analysis is different from sEM-vr, FIEM in [6,15] (inspired by SVRG, SAGA) while the approximate mean field for SPIDER-EM is **biased**. With such challenges, it is not obvious that the analysis of SPIDER can be directly applied to SPIDER-EM. In fact, our algorithm design/analysis is only **inspired** by SPIDER, for it involves a brand new analysis on incremental-EM algorithms.

Reviewer 3: Thank you for the careful review and positive feedback. We have addressed your comments regarding the contributions and novelty in the beginning of this rebuttal. In the final version, we will include a highlight about them.

Reviewer 4: Thank you for the careful review and positive feedback. Here are the specific answers to the comments.

Assumptions H4.3 and H5: These assumptions are actually classical in the analysis of incremental/stochastic EM algorithms - a relevant reference is [15], where explicit examples satisfying H4.3, H5 are provided. In fact, $B(s)$ can be written in terms of the Hessian of the function in H3, which is positive definite if the map $\mathsf{T}(s)$ is unique. As shown in [15], this holds for a number of distributions such as GMM when $\mathsf{R}(\theta)$ is a properly designed barrier function.

Parameter/Statistics Space: We emphasize that it is essential for the EM algorithm to work alternatively in the expectation space (E-step) and the parameter space (M-step). However, we agree that the present writing is not optimal and we will try to streamline the discussions in the final version.