Reviewer 1:

- The potential applications of the method is non-obvious from the current introduction. A motivating example in the introduction where the expectation is intractable but the conditions of the method are satisfied, like a Markov random field, would help make it clear. - On the side of the theory, the submission is well executed but maybe on the incremental side. it combines variants of EM which have been analyzed separately before (the combination of Monte-Carlo and stochastic EM of Kuhn et al. [2019], with the variance-reduced stochastic EM of Chen et al. [2018] and Karimi et al. [2019]) and the convergence results do not provide significant new insight.

-Notations issue

-Explain better exponential family

Reviewer 2:

- The proposed algorithm seems to be incremental. Compared with the previous work [20], the proposed algorithm introduces an extra stochastic update for the statistics.

- The contributions of the current paper are not strong enough. The theoretical analyses of the current paper seem to be based on previous work [20]. For example, Lemmas 1 and 2 are both established in [20]. Theorem 1 looks to be a straight forward extension of the results in [20] by directly factor out the dependence of the noise. For Theorems 2 and 3, it is unclear what is the key challenges in the analyses compared with the analyses in [20] when introducing the extra stochastic update.

- The presentation of the current paper needs to be improved. There are lots of definitions are missing. For example, what is J in A3? In addition, the connections between the key lemmas and the main results are unclear in sections 3.2 and 3.3. The authors should include a proof sketch in these sections to clarify the role of these key lemmas.

- The proposed algorithm and the theoretical analyses are based on previous work [20], but the authors are failed to clarify the key challenges in the analyses by introducing the extra stochastic updates in the proposed method.

-To further show the scalability of the proposed algorithm, the authors should consider comparing the proposed method with the methods proposed in [20]. In addition, there are some related works[1,2,3] studying the gradient-based EM algorithms, the authors may want to discuss them in the prior work.

Reviewer 3:

- A lot of notations in the theorems are not defined nor explained. - In paragraph 2.2, the notations of $\tilde{S}$ are really confusing. They sometimes refer to the $\tilde{S}$ defined equation 6 (lines 104 and 105) and sometimes to the $\tilde{S}$ defined in equation 8 making the reading difficult.

- The theorems are proved under a compacity hypothesis (A1) that is usually removed by projection on increasing compact subsets.

- In addition, while proving the convergence under (A1), the last example does not satisfy the compacity hypothesis required in (A1).

- The authors present 3 different proxies for the incremental step in the table 1. However, they do not talk of the pros and cons of each method.

- In hypothesis (A5) the inequalities should be < and not <=.

- All theorems are proved with rho constant. What happens if rho is no longer a constant ?

- L243: lowercase phi should be replaced by uppercase phi.

Reviewer 4:

- The idea of the incremental-step is clear to scale to a large dataset, but there are some major concerns on the proposed two-timescale stochastic EM algorithms: (1). It is possible that t_i^k is empty, how is that event to influence the algorithms and the theoretical bounds derived and obtained in the paper; (2). How are these obtained convergence bounds compared with those of the previous algorithms under the same conditions? (3). How fast do the proposed algorithms? (4). How to determine $K - m$, the total number of iterations or when the algorithm stops?

-Notations issue too -The fitted Gaussian mixture models in subsection 4.1 are too simple and cannot represent the general cases.

Reviewer 5: -As mentioned, one limitation could be the theoretical guarantee in terms of the stationary point. This is loose and does not provide insights on how EM performs in many real problems, e.g., examples in Section 4. I recommend looking into particular problem structures and look at the landscape, hopefully can establish convergence in terms of some objective with concrete convergence rates. One starting example is the Gaussian mixture model.

-The authors list some prior works. As I see, it is hard to tell which standard performance we can compare with. For EM algorithm design, we could compare from sample scheme and computational burden.