We would like to thank the five reviewers for their feedback. Upon acceptance, we will include in the final version (a) *improved notations*, (b) *an improved presentation of related work* and (c) *missing references*. We first discuss a few common concerns shared by **R1**, **R2**, **R3**,**R4** and **R5**.

●●●●● **Notations Issue**: We acknowledge the cumbersome notations of our paper and will modify them in order to reflect the reviewers remarks. Deterministic and Stochastic quantities will be clearly identified in their notations and some less important abstractions will be dismissed in the revised paper.

●● **Novelty:**: We agree with the reviewers that our contribution stands as a combination of variance reduction ([11,19]), EM ([20,22]) and SA ([12,32]). The synthesis of these contributions into a single framework constitues the originality of this paper on the algorithmic and theoretical plans. Adding a layer of Monte Carlo (MC) approximation and the stepsize γ_k to reduce its variance introduce some new technicalities and challenges that need careful considerations.

R1: We thank the reviewer for valuable comments. We would like to clarify the following points:

Exponential Family: The curved exponential family is a classical one in the EM-related literature and holds for most models where EM is useful [McLachlan&Krishnan 2007] . While remaining general, the advantage of such family is to write the algorithm updates only with respect to the sufficient statistics and not in the space of parameters θ. Yet, we would like to clarify that due to Bayes rule and the intractable normalizing constant, a complete likelihood that belongs to the exponential family does not imply a tractable posterior distribution.

R2: We thank the reviewer for the comments and typos. We add the following remarks:

Comparison with [Karimi+, 2019]: While both of these papers are dealing with nonconvex functions, the added layer of randomness, due to the sampling step in our method, makes it a practically and theoretically different approach. Yet, as pointed by the reviewer, Lemmas 1 and 2 are needed to characterize the deterministic part of those models. The stochastic part (posterior sampling) is new and is the object of our paper.

Comparison with gradient-based EM algorithms: Gradient-based methods have been developed and analyzed in [Zhu+, 2017] but remain out of the scope as they tackle the high-dimensionality issue. The exponential family allows to leverage the sufficient statistics and a max. function $\overline{\theta}(\overline{\mathbf{s}}(\theta))$ updating θ without an inner iterative process (eg. GD).

R3: We thank the reviewer for insightful comments and typos. Our point-to-point response is as follows:

Compacity assumption: For our analysis, we assume that the statistics always remain in a defined compact subset of $\mathbb{R}^d$. While this assumption holds for the GMM example, it is not the case for the deformable template analysis one. We implemented the *Truncation on random boundaries* techniques found in [Allassonniere+, 2010] based on restart.

Comparison of proxies (Table 1): The advantage between the incremental proxy and the two variance reduction yields from their sublinear convergence rate (see Th. 2 and 3). The vrTTEM requires the tuning of the epoch length m but stores one vector of $n+1$ quantities while the fiTTEM requires storing two vectors of parameters without any tuning.

R4: We thank the reviewer for valuable comments and references. Our point-to-point response is as follows:

Various questions: • t_i^k is not empty by construction since it stores the iteration at which index i was last drawn. They are initialized after a single pass over all indices. • We are not aware of similar algorithms mixing optimization and sampling techniques. Neither SAEM nor MCEM have been studied non asymptotically. • The random stopping criterion K_m is common in non-convex optimization, see [Ghadimi & Lan,2013] and is needed for theoretical purposes.

R5: We thank the reviewer for valuable comments and references. We make the following precision:

Comparison to EM theory papers: After careful consideration, the listed references are either related to deterministic EM methods, where no sampling is required since the expectations are always tractable, or to gradient EM method which has been dealt with above (see **R2:**). We agree that more specific studies depending on the model (such as mixture) would lead to different analysis, yet we would lose the generality of our paper.

Response to Additional Feedbacks : • The two-timescale update is crucial for the following reason: the *noise induced by sampling a single index* is tempered by ρ_k (Eq. (9)) while the *noise induced by sampling the latent variables* is tempered by γ_k. Initial runs without γ_k showed poor convergence properties due to the large variance of the posterior sampling. Of course the downside will be the tuning of two stepsizes. In practice, using a decreasing stepsize as $\gamma_k = 1/k^\alpha$ and constant $\rho \propto n^{2/3}$ works well. Using a non constant ρ_k is an interesting open question. • Notations and paragraph on stepsizes will be clarified in the revised version. • Assumptions: **A1-A4** are necessary for the non asymptotic analysis and are easily satisfied for exponential family model such as GMM as thoroughly described in [Karimi+, 2019]. **A5** is a classical result which we realized that it needs to be developed in greater detail. For the case of i.i.d. samples, using Example 19.7, Lemma 19.36 from 'Asymptotic Statistics' by van der Vaart (2000), it can be shown that the MC noise η_i is is sublinear in p. Meanwhile, the cases for Markov samples are not as obvious, though comparable results can be found for β-mixing processes [Thm.2, Doukhan+1995]. • The term *Global* is employed in the sense that it does not restrict the initialization, a common assumption for the analysis of EM. • All the methods converge to a reasonable precision (y axis from 1 to 10^{-3}). The GMM example illustrates how the proposed framework achieves to reduce the variance of baseline method in order to reach a higher precision (10^{-3} in this example).