We would like to thank three 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 **reviewer 1**, **reviewer 2**, **reviewer 3**,**reviewer 4** and **reviewer 5**.

●●●●● **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.

●● **Originality of the Contribution:**: We agree with the reviewer that our contribution stands as a combination of variance reduction ([Chen+, 2018], [Johnson+, 2013]), EM methods ([Karimi+, 2019], [Kuhn+, 2019]) and Stochastic Approximation ([Delyon+, 1999], [Robbins and Monro, 1951]). The diversity of all those contributions into a single framework constitues what we believe to be the originality of this paper both on the algorithmic and theoretical plans. Adding a layer of noise, due to MC approximation, and a second stepsize to reduce its variance present some added technicalities that need careful consideration.

●● **Importance of the Assumptions**:

Reviewer 1: We thank the reviewer for valuable comments. We would like to clarify the following points:

Potential Applications: We admit it is a challenging task to present all technical results and obvious applications 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. For instance, the deformable template analysis or the pharmacokinetics example (which can be found in the Appendix) will be presented throughout the paper with clear motivation for using our scheme.

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. The M-step is thus in general expressed in *closed-form* and not as a black-box optimization ($\arg\max$ operation). Yet, we would like to clarify to the reviewer that exponential family does not imply tractable posterior. The intractability of this posterior sampling step is, in our case, due to the nonconvexity of the loss function. 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.

Reviewer 2: We thank the reviewer for the comments and typos. We add the following remarks:

Comparison with [Karimi+, 2019]: We would like to clarify to the reviewer that the work in [Karimi+, 2019] can not be directly compared to ours since the problems and models tackled are different. While both of these papers are dealing with nonconvex objective functions, the added layer of randomness, due to the sampling step in our method, makes it practically and theoretically different approach. Yet, as pointed by the reviewer, somme lemmas (Lemma 1 and 2) are recalled in our paper and are needed to characterize the deterministic part of those models. The stochastic part (sampling from the posterior distribution) 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 they remain out of the scope of this paper as they tackle the high-dimensionality issue. Gradient-EM are also relevant when the M-step can only be solved through a gradient descent method. In our case, the exponential family assumption allows us to leverage the sufficient statistics and the maximization functions $\overline{\theta}(\overline{\mathbf{s}}(\theta))$ to update the parameters without an inner iterative process.

Reviewer 3: We thank the reviewer for insightful comments and typos. Our point-to-point response is as follows:

Compacity assumption: We agree with the reviewer on the need for random projections in order to stay in a compact set. 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 Theorems 2 and 3). The vrTTEM requires the tuning of the epoch length m but only stores one vector of n parameter and a control variate term while the fiTTEM requires storing two vector of parameters (for the two randomly drawn indices) but does not require any hyper-parameter tuning.

Reviewer 4: 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. The only algorithm we are aware of are the SAEM and the MCEM and none of them have been