Report on "A Class of Two-Timescale Stochastic EM Algorithms for Nonconvex Latent Variable Models ."

The review consists of three parts : summary of the paper, main comments and questions, conclusion.

Summary of the paper

The paper presented a new class of stochastic expectation maximization algorithms dedicated to estimation in latent variable models. The authors proposed, studied and applied stochastic estimation algorithms which are able to deal with both large datasets and complex models leading to intractable expectation calculation in the E-step of the classical expectation maximization (EM) algorithm.
There were three main contributions. First the authors introduced a novel class of two-timescale algorithms based on stochastic approximation dynamic and Monte Carlo approximation to handle the intractable expectation calculation and on incremental updates to scale with large dataset. They proposed a unified setting gathering the new two-timescale algorithms and an existing incremental stochastic approximation EM algorithm as a particular case. Then they focused on two particular algorithms including variance reduction technics at two different levels. Second the authors established theoretical results for the three algorithms mentioned above. More precisely they presented "global", meaning that it is "independent of the initialization", and finite-time upper bounds on the second order moment of the gradient of the objective function. Finally the authors provided numerical experiments to illustrate the performances of the two proposed algorithms. They carried out simulation studies in a gaussian mixture model and in a pharmacokinetic model. They also applied their methods on real data for a deformable template model in image analysis. The paper ended with a short conclusion. Proofs of theoretical results are detailed in the appendix.

Main comments and questions

1. Writing of the paper
   - Many notations are not introduced in the paper or introduced after being used or imprecisely defined. For examples, the set S introduced in assumption A1 is not defined at all (page 6), the notation τ_i^k is introduced later (page 6), and notations used in the three main theoretical results (e.g. $\tilde{\alpha}_k$ and $\tilde{\Gamma}_k$ in Theorem 1) are introduced only in the appendix. Many cross-references are inexact : for examples, it is written that "the iSAEM update (1) is equivalent to ..." (page 9 line -5) but there is no equation (1) ; it is written that "$\bar{s}^{(k)}$ is defined by (1.4) " (page 9 line -1) but equation (1.4) defined the quantity $\bar{s}_i(\theta)$.
   - There are also many imprecisions, resulting in confusion for the reader. For example, the authors defined $\eta_i^{(k)} = \tilde{S}_i^{(k)} - \bar{s}_i(\theta^{(k)})$ (page 7) and claimed $\tilde{S}_{i_k}^{(k+1)} = \bar{s}_{i_k}(\theta^{(k)}) + \eta_{i_k}^{(k+1)}$ (page 10) shifting some indices of the definition. Another example, the authors claimed "we also establish global (independent of the initialization) ... upper bounds ..." (page 3 line 12). But they wrote after their first result " the convergence bound is composed of an initialization term ..."

(page 8 line 1). So the result is obviously not independent of the initialization. May be the authors would mean "for any initialization" ?

- Several statements are inaccurate : for example, the authors wrote " the integral and finite sum structure ... have been addressed separately in the literature" (page 2 line 16) and they cited few lines after a reference to "an incremental variant of the SAEM" which addressed both jointly.
- Several references are missing : for examples, the reference to Fort and Moulines (2003) for theoretical convergence results on the MCEM algorithm (page 4), a reference for the pharmacokinetic model considered in the numerical section.

2. Presentation of the new class of algorithms
   - Describing the dynamic of the EM algorithm at iteration k at the beginning of section 1 would help the reader (page 2).
   - The authors proposed three different possible proxies for the incremental step of their class of two-timescale algorithms (page 5 Table 1). What is the intuition of each of these three proxies ? In the definition of fiTTEM update, is it $t_{i_k}^k$ or $\tau_{i_k}^k$ in the first equation ? Indeed choosing the first would correspond to considering the last iteration l before k such that $j_l = i_k$ whereas the second would refer to the last iteration l before k such that $i_l = i_k$. Moreover the authors fixed for the proxy called iSAEM the stepsize ρ_{k+1} in (2.4) equal to 1 which remains to skip one of the two levels of their method. The advantage is that it allowed in particular to recover the incremental SAEM algorithm. However this algorithm did not include the variance reduction part and it is quite surprising to conduct the theoretical study and the numerical experiment of this algorithm omitting one of the timescale. Why did the authors not study the property of the proxy proposed in line 1 of Table 1 coupled with the incremental step (2.4) ? Can they comment on this choice ?
   - The authors defined the two timescale property as the stepsizes satisfy $\lim \gamma_k/\rho_k$ is less than 1 (page 5 line -6). However in the theoretical analysis and in the numerical experiments the stepsizes (ρ_k) are chosen constant. Can the stepsizes (γ_k) also be chosen constant such that $\gamma_k/\rho_k < 1$? What is exactly the sense of the two timescale property ? Would it be possible to illlustrate this through experiments ?
   - The initial value for the parameters θ is set to zero (page 6). Is it required that zero belongs to the parameter set ?

3. Theoretical results
   - Assumptions required for theoretical results are not commented (page 6). In particular, are these assumptions restrictive ? For example, the matrix $B(s)$ is introduced (page 7 line 2). Assumption 4 required regularity conditions on the function B which are not obvious to establish. Are the assumptions fulfilled by the pharmakokinetic model and the deformable template model considered in the numerical section ? For example if S (not defined in the paper) is the set of values taken by the sufficient statistic, then Assumption 1 is obviously not satisfied by the deformable template model and the pharmakokinetic model.

- Some assumptions are redundant, others are missing : for examples Assumption 5 page 7 required two expectations to be bounded ; however assuming the second one is bounded implies the first one is just by applying Jensen inequality. The functions ψ and ϕ should be at least twice differentiable to ensure the existence of the Hessian H_L^θ of the function L (page 7).
- The filtration $(\mathcal{F}_k)$, which is crucial in the theoretical analysis, is introduced (page 7 line 14) but is not precisely defined. In particular there are two sources of stochasticity, one related to the sampling of the latent variable (z_k) and one to the sampling of the random indices (i_k, j_k). How is $(\mathcal{F}_k)$ defined ?
- The authors did not detail the practical interest of their theoretical results. Many constants involved in the established upper bounds are not tractable. So what can be done practically with these upper bounds ? For example, could guidelines on how to fixed the tuning parameters such as the stepsizes (γ_k) be deduced, as proposed by Ghadimi and Lan (2013) cited by the authors ?

4. Numerical experiments
   - The authors compared the performances of the two proposed methods vrTTEM and fiTTEM including variance reduction technics with four others methods (EM, iEM, SAEM, iSAEM) of the literature which did not include variance reduction technics. They highlighted through experiment and concluded that their methods have better performances than these four methods. However there exists in the literature at least one method including variance reduction technics, called variance reduced stochastic EM algorithm proposed by Chen et al (2018) and cited by the authors (page 4 line -2). Why did the authors not compare their methods to this one which should be more competitive than methods without variance reduction ? Moreover as already written above the authors could have considered at least for numerical experiment the update proposed in line 1 of Table 1 integrated in their two timescale algorithm ($\rho_k \neq 1$).
   - The tuning parameters of the algorithms are chosen arbitrary in the experiment. Their choice is not discussed. At least it would be of interest to discuss this choice and also to vary the values of the sequences of stepsizes (γ_k) and (ρ_k) to illustrate their numerical effects, in particular the two timescale property mentioned by the authors (page 5). What would happen if it is not satisfied ?
   - There are many imprecisions in the numerical section : in the deformable template model, what is the geometric covariance indexed by g ? Also the photometric hyperparameter indexed by p ? In the pharmakokinetic model, what is $\theta^\star$?
   - The authors wrote that a MCMC procedure is used in the deformable template model to sample the latent variables. Is it a Metropolis Hastings algorithm as in the pharmakokinetic model later ? If so why not detailing this earlier ? Moreover there is no comment on the burn-in period chosen.
   - The code corresponding to the numerical experiments are not available. It is important targeting open access science to make the code accessible to guarantee the reproductibility of numerical experiments.

5. Proofs details in the appendix
   - There are also many inaccuracy in the proofs : for example the authors claimed "we observe that $\mathbb{E}(\bar{s}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) = \bar{s}^{(k)} - \bar{\mathcal{S}}^{(k)}$" (page 15). The left hand side of the equality is a real vector, whereas the right hand side is a random vector since $\bar{s}^{(k)}$ is a real vector and $\bar{\mathcal{S}}^{(k)}$ a random vector depending on the sequence of random indices (j_k). How can one observe this equality ? Is $\bar{\mathcal{S}}^{(k)}$ constant ? It is difficult to check if the proofs are correct.
   - The proof of Lemma 3 is missing.
   - The proofs of Lemma 4 and 5 appeared exactly in the same form twice in the section proof sketches and in the appendix.

Conclusion

The paper addressed a very interesting and actual topic. The proposed class of algorithms integrated a novelty regarding variance reduction technics at two different levels. However the paper is very bad written and difficult to read. Many notations are not introduced or introduced after being used or non rigourously defined. The authors gave no intuitions on the proposed two timescale methodology. The assumptions required for theoretical results were not commented, some were missing, some were redundant. The theoretical results were not rigourously stated, in particular notations of quantities involved in the statement of these results were defined in the appendix. The proofs, containing many mistakes and inaccuracies, could not be checked. The way theoretical results could be usefull in practice was not developed. The numerical study contained many inaccuracies, several references and definitions were missing and the results were not deeply discussed and commented. No comparisons with others existing algorithms which also integrated some variance reduction technics were carried out. The codes are not accessible.

To conclude the paper dealt with an interesting actual subject but is not suitable for publication in its actual form. Proofs have to be verified carefully. Numerical experiments have to be enriched to get fair comparison. Codes have to done accessible in an open access science context. Finally a rigourous and deep rewriting of the whole paper has to be carry out.