Two-Timescale Stochastic EM Algorithms

author names withheld

Editor: Under Review for ALT 2021

1. Reviewer 1

Originality: 1. The result in Theorem 2 seems to be partly follow from standard results in non-convex optimization literature, where we can bound the sum of the "squared norm gradients" of the objective function V using the initial sub-optimality gap ΔV. There are some additional terms due to Monte Carlo averaging. Does these extra terms also arise by applying some well known results in MCMC literature? It would be nice if the authors could point out the novelty in the Theorem proofs. A similar comment applies to Theorems 1 and 3.

2. The paper should include a clear discussion comparing Theorems 1, 2 and 3 with convergence guarantees of similar algorithms in the literature (SAEM, for example). This will help us understand the novelty of the paper. This may also highlight scenarios where the proposed algorithms are clearly preferable.

Questions: 1. The usefulness of the integer K_m in the theorem statements is not clear to me. Is it necessary?

2. Given a dataset, a natural question for a practitioner is which algorithm to choose among the proposed algorithms in the paper. It will be nice to have a clear discussion on – *scenarios where one algorithm is preferable than the others?*

3. The hyper-parameters in the algorithms depends on a few quantities which may not be known to the practitioners; e.g. L_s, L_v, v_{min} etc. I presume that the theorem statements hold with appropriate upper or lower bounds on these quantities. A clear discussion on how to obtain bounds on these quantities will increase the appeal of the methods to a practitioner.

Our Reply.

We thank the reviewer for the insightful comments and provide as many clarifications as we can in the following:

Novelty:

Our contribution stands as a combination of variance reduction, EM and Stochastic Approximation. 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.

Does these extra terms also arise by applying some well known results in MCMC literature?

We thank the reviewer for pointing out where our main contribution actually is. Indeed, those extra noise terms represent the main challenge in our theoretical analysis (and appear to be a practical challenge as well, that we resolve by adding the Robbins Monro type of update, thus reducing the noise as much as possible).

The main assumption needed for making this analysis (with a MC noise) possible, is Assumption A5 that bounds the noise induced by the MC approximation. This noise is defined in Eq. (12) and is simply put the gap between the expectation and its stochastic counterpart. Current MCMC literature allows us to make sure that this gap is bounded, in other words that the approximation can be controlled.

Why is bounding this control term possible?

It is common in statistical and optimization problems, to deal with the manipulation and the control of random variables indexed by sets with an infinite number of elements. Here, the controlled random variable is an image of a continuous function defined as η_i (Eq. 12). To characterize such control, we have recourse to the notion of metric entropy (or bracketing number) as developed in [Van der Vaart, 2000, Vershynin, 2018, Wainwright, 2019]. A collection of results from those references gives intuition behind our assumption A5, classical in empirical processes. In [Vershynin, 2018, Theorem 8.2.3], the authors recall the uniform law of large numbers:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{M}\sum_{i=1}^{M} f\left(z_{i,m}\right)-\mathbb{E}[f(z_i)]\right|\right]\leq\frac{CL}{\sqrt{M}}\quad\text{for all}\quad z_{i,m}, i\in[1,M],$$

where $\mathcal{F}$ is a class of L-Lipschitz functions. Moreover, in [Vershynin, 2018, Theorem 8.1.3] and [Wainwright, 2019, Theorem 5.22], the application of the Dudley inequality yields:

$$\mathbb{E}[\sup_{f\in\mathcal{F}}|X_f-X_0|]\leq\frac{1}{\sqrt{M}}\int_0^1\sqrt{\log\mathcal{N}\left(\mathcal{F},\|\cdot\|_\infty,\varepsilon\right)}d\varepsilon,$$

where $\mathcal{N}\left(\mathcal{F},\|\cdot\|_\infty,\varepsilon\right)$ is the bracketing number and ϵ denotes the level of approximation (the bracketing number goes to infinity when $\epsilon\to 0$).

The paper should include a clear discussion comparing Theorems 1, 2 and 3 with convergence guarantees of similar algorithms in the literature (SAEM, for example)

Non asymptotic bounds, for either convex or nonconvex objective functions, do not exist in the literature. Neither MCEM or SAEM have been studied in a finite time horizon making our contribution the first to tackle this joint problem theoretically.

Many papers related to deterministic EM methods or to gradient EM method have been published with great theoretical analysis but are not in the scope of our study since no sampling is required as the expectations are always tractable (models are mainly convex), such as: [Xu J, Hsu DJ, Maleki A. Global analysis of expectation maximization for mixtures of two gaussians], [Yan B, Yin M, Sarkar P. Convergence of gradient EM on multi-component mixture of Gaussians], [Kwon J, Ho N, Caramanis C. On the Minimax Optimality of the EM Algorithm for Learning Two-Component Mixed Linear Regression], [Kwon J, Caramanis C. EM converges for a mixture of many linear regressions] or [Wu Y, Zhou HH. Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations].

1. The usefulness of the integer K_m *in the theorem statements is not clear to me. Is it necessary?*

The random termination number K_m is inspired by [Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming, Ghadimi & Lan,2013] which enables one to show non-asymptotic convergence to stationary point for non-convex optimization. Involving a randomly drawn stopping criterion has been widely used since then in the stochastic nonconvex optimization to provide novel finit-time analysis (in expectation), see [On the Global Convergence of (Fast)

Incremental Expectation Maximization Methods, Karimi, Wai, Moulines and Lavielle, 2019], [A Simple Convergence Proof of Adam and Adagrad, Defossez, Bottou, Bach and Usunier, 2020] or [On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization, Zhou, Chen, Cao, Tang, Yang and Gu, 2020] to name a few.

2. It will be nice to have a clear discussion on – scenarios where one algorithm is preferable than the others?

Comparison of proxies (Table 1): The advantage of the variance reduced proxies over the incremental proxy 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. In terms of performance, our numerical experiments show that the two variance reduced methods (vrTTEM and fiTTEM) are always the best options. The choice between both for practitioners will depend on how they perform on their specific problems and we would advise running both (as the implementation are very close).

3. The hyper-parameters in the algorithms depends on a few quantities which may not be known to the practitioners; e.g. L_s, L_v, v_{min} etc. I presume that the theorem statements hold with appropriate upper or lower bounds on these quantities. A clear discussion on how to obtain bounds on these quantities will increase the appeal of the methods to a practitioner.

2. Reviewer 4

This paper cobbles together two earlier pieces of work. As such, there isn't a whole lot of novelty in either the algorithm or analysis. But it is nicely done, and demonstrates that it is in fact possible to address both sets of problems (described above) at the same time.

Our Reply.

We would like to thank the reviewer for the feedback on our paper. We would like to stress on the originality of our contribution both on the practical and theoretical aspects.

As the reviewer mentioned, the aim of our paper is to be able to tackle the two problems that he/she presented, i.e. large dataset (heavy pass over the dataset) and intractable expectation in the E-step (commonly an issue when the models involved are nonconvex, see image analysis and pharmacokinetics examples in our numerical experiments section).

However, the combination of existing methods tackling each problem separately is not obvious and constitutes in our opinion the originality of our paper. The only paper deriving *non asymptotic bounds* for *nonconvex* objectives of EM algorithms is Karimi-Wai-Moulines-Lavielle (2019) as rightly cited by the reviewer. Yet, when the expectations in the E-step is not tractable, the results on their paper do not hold anymore.

Comparison with [Karimi-Wai-Moulines-Lavielle, 2019]: While both papers are dealing with nonconvex functions, the added layer of randomness, due to the sampling step (direct or MCMC) in our method, makes it a different approach in practice and theory. Lemmas 1 and 2 are needed to characterize the deterministic part of the model, common to our paper and theirs, though the stochastic part (posterior sampling), and its variance reduction, is new and is the object of our paper (eg. Lemma 6). Besides, due to the high noise that Monte Carlo (MC) approximation can involve, our framework adds a second stepsize γ_k (compared to the only stepsize in Karimi-Wai-Moulines-Lavielle (2019)) making the algorithm two-timescale and hence involves both algorithmic

and theoretical challenges. For completeness on this additional stepsize γ_k: 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.

3. Respecting the 6000 characters limit

Reviewer 1: We thank the reviewer for the insightful comments and provide some clarifications: Novelty: Our contribution stands as a combination of variance reduction, EM, and Stochastic Approximation. The synthesis of these contributions into a single framework constitutes 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 introduces some new technicalities and challenges that need careful considerations.

"...some well-known results in MCMC literature?": Indeed, those extra noise terms represent the main challenge in our theoretical analysis (and appear to be a practical challenge as well, that we resolve by adding the Robbins Monro type of update, thus reducing the noise as much as possible). The main assumption needed for making this analysis (with a MC noise) possible, is Assumption A5 that bounds the noise induced by the MC approximation. This noise is defined in Eq. (12) and is simply put the gap between the expectation and its stochastic counterpart. Current MCMC literature allows us to make sure that this gap is bounded, in other words, that the approximation can be controlled.

It is common in statistical and optimization problems, to deal with the manipulation and the control of random variables indexed by sets with an infinite number of elements. Here, the controlled random variable is an image of a continuous function defined as η_i (Eq. 12). To characterize such control, we have recourse to the notion of metric entropy (or bracketing number) as developed in [Van der Vaart, 2000, Vershynin, 2018, Wainwright, 2019]. A collection of results from those references gives intuition behind our assumption A5, classical in empirical processes. In [Vershynin, 2018, Th. 8.2.3]:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{M}\sum_{i=1}^{M}\tilde{S}_i - \bar{s}_i(\theta)\right|\right] \leq \frac{CL}{\sqrt{M}} \quad \text{for all} \quad i \in [1, M]$$

where $\mathcal{F}$ is a class of L-Lipschitz functions. Moreover, in [Vershynin, 2018, Th. 8.1.3] and [Wainwright, 2019, Th. 5.22], Dudley inequality yields:

$$\mathbb{E}[\sup_{\tilde{S}\in\mathcal{S}}|X_f - X_0|] \leq \frac{1}{\sqrt{M}}\int_0^1 \sqrt{\log \mathcal{N}\left(\mathcal{F}, \|\cdot\|_\infty, \varepsilon\right)}d\varepsilon$$

where $\mathcal{N}\left(\mathcal{S}, \|\cdot\|_\infty, \varepsilon\right)$ is the bracketing number and ϵ denotes the level of approximation (the bracketing number goes to infinity when $\epsilon \to 0$).

Regarding existing theoretical results: Nonasymptotic bounds, for either convex or nonconvex objective functions, do not exist in the literature. Neither MCEM or SAEM have been studied in a finite time horizon making our contribution the first to tackle this joint problem theoretically. Many papers related to deterministic or gradient EM have been published with theoretical analysis but are not in the scope of our study since no sampling is required (models are mainly convex), such

as: [XHM. Global analysis of expectation maximization for mixtures of two gaussians] or [YYS, Convergence of gradient EM on multi-component mixture of Gaussians]

The random termination K_m is inspired by [Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming, Ghadimi, Lan,2013] which enables one to show non-asymptotic convergence to stationary point for non-convex optimization. Involving a randomly drawn stopping criterion has been widely used since then in the literature to provide novel finite-time analysis (in expectation), see [On the Global Convergence of (Fast) Incremental EM Methods, KWML, 2019], [A Simple Convergence Proof of Adam and Adagrad, DBBU, 2020] or [On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization, ZCCTYG, 2020] to name a few.

The advantage of the variance reduced proxies over the incremental proxy yields from their sublinear convergence rate (see Th. 2-3). 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 tuning. Our numerical experiments show that the two variance reduced methods (vrTTEM and fiTTEM) are always the best options.The choice between both for practitioners will depend on how they perform on their specific problems and we would advise running both (as the implementation are very close).

Reviewer 4: We would like to thank the reviewer for the feedback on our paper. The combination of existing methods tackling each problem separately (large dataset and intractable expectation) is not obvious and constitutes in our opinion the originality of our paper. The only paper deriving *nonasymptotic bounds* for *nonconvex* objectives of EM algorithms is KWML (2019) as rightly cited by the reviewer. Yet, when the expectations in the E-step are not tractable, the results on their paper do not hold anymore. Comparison with [KWML, 2019]: While both papers are dealing with nonconvex functions, the added layer of randomness, due to the sampling step (direct or MCMC), makes it a different approach in practice and theory. Lemmas 1 and 2 characterize the deterministic part of the model, common to our paper and theirs, though the stochastic part (posterior sampling), and its variance reduction, is new and is the object of our paper (eg. Lemma 6). Besides, due to the high noise that MC approximation involves, our framework adds a second stepsize γ_k making the algorithm two-timescale and hence involves both algorithmic and theoretical challenges The two-timescale update (two stepsizes) is crucial so that the noise induced by sampling a single index is tempered by ρ_k and 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 variance