A Class of Two-Timescale Stochastic EM Algorithms for Nonconvex Latent Variable Models

Belhal Karimi and Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{belhalkarimi, liping11}@baidu.comm

July 26, 2021

Abstract

The Expectation-Maximization (EM) algorithm is a popular choice for learning latent variable models. Variants of the EM have been initially introduced by Neal and Hinton (1998), using incremental updates to scale to large datasets, and by Wei and Tanner (1990); Delyon et al. (1999), using Monte Carlo (MC) approximations to bypass the intractable conditional expectation of the latent data for most nonconvex models. In this paper, we propose a general class of methods called Two-Timescale EM Methods based on a two-stage approach of stochastic updates to tackle an essential nonconvex optimization task for latent variable models. We motivate the choice of a double dynamic by invoking the variance reduction virtue of each stage of the method on both sources of noise: the index sampling for the incremental update and the MC approximation. We establish finite-time and global convergence bounds for nonconvex objective functions. Numerical applications on various models such as deformable template for image analysis or nonlinear mixed-effects models for pharmacokinetics are also presented to illustrate our findings.

Keywords: two-timescale, stochastic, em, nonconvex, mcmc, monte carlo, latent variable

1 Introduction

Learning latent variable models is critical for many important modern machine learning problems, see for instance McLachlan and Krishnan (2007) for references. We formulate the training of this type of model as the following empirical risk minimization problem:

$$\min_{\boldsymbol{\theta}\in\Theta} \overline{\mathsf{L}}(\boldsymbol{\theta}) := \mathsf{L}(\boldsymbol{\theta}) + \mathrm{r}(\boldsymbol{\theta}) \quad \text{with } \mathsf{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{L}_i(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \big\{ -\log g(y_i;\boldsymbol{\theta}) \big\} \,, \tag{1}$$

where $\{y_i\}_{i=1}^n$ are observations, $\Theta \subset \mathbb{R}^d$ is the parameters set and $\mathrm{r} : \Theta \to \mathbb{R}$ is a smooth regularizer. The objective $\overline{\mathsf{L}}(\boldsymbol{\theta})$ is possibly nonconvex and is assumed to be lower bounded. In the latent data model, the likelihood $g(y_i;\boldsymbol{\theta})$, is the marginal distribution of the complete data likelihood, noted $f(z_i, y_i;\boldsymbol{\theta})$, such that

$$g(y_i;\boldsymbol{\theta}) = \int_{\mathsf{Z}} f(z_i, y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i), \tag{2}$$

where $\{z_i\}_{i=1}^n$ are the vectors of latent variables associated to the observations $\{y_i\}_{i=1}^n$. In this paper, we assume that the complete data likelihood belongs to the curved exponential family (Efron, 1975), i.e.,

$$f(z_i, y_i;\boldsymbol{\theta}) = h(z_i, y_i)\exp(\big\langle S(z_i, y_i)\,,\, \phi(\boldsymbol{\theta})\big\rangle - \psi(\boldsymbol{\theta}))\,, \tag{3}$$

where $\psi(\boldsymbol{\theta})$, $h(z_i, y_i)$ are scalar functions, $\phi(\boldsymbol{\theta}) \in \mathbb{R}^k$ is a vector function, and $\{S(z_i, y_i) \in \mathbb{R}^k\}_{i=1}^n$ is the vector of sufficient statistics. Batch EM (Dempster et al., 1977; Wu, 1983), the method of reference for (1), is comprised of two steps. The E-step computes the conditional expectation of the sufficient statistics of (3), noted $\overline{\mathrm{s}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \overline{\mathrm{s}}_i(\boldsymbol{\theta})$, where for all $\boldsymbol{\theta} \in \Theta$ and $i \in [n]$, where $[n] := \{1, \cdots, n\}$:

$$\overline{\mathrm{s}}_i(\boldsymbol{\theta}) := \int_{\mathsf{Z}} S(z_i, y_i)p(z_i|y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i)\,, \tag{4}$$

and the M-step is given by

$$\overline{\boldsymbol{\theta}}(\overline{\mathrm{s}}(\boldsymbol{\theta})) := \arg\min_{\vartheta\in\Theta} \big\{ \mathrm{r}(\vartheta) + \psi(\vartheta) - \big\langle \overline{\mathrm{s}}(\boldsymbol{\theta})\,,\, \phi(\vartheta)\big\rangle \big\}\,. \tag{5}$$

There are two main caveats of such a method: (a) with the explosion of data, the first step of the EM is computationally inefficient as it requires, at each iteration, a full pass over the dataset; and (b) the complexity of modern models makes the expectation in (4) intractable. Both of these constraints occur in the E-step of the EM algorithm, see the integral and finite sum structure of (4), and to the best of our knowledge, have been addressed separately in the literature. In this work, we tackle them jointly.

1.1 Prior Work

Inspired by stochastic optimization procedures, Neal and Hinton (1998); Cappé and Moulines (2009) developed respectively an incremental and an online variant of the E-step in models where the expectation is computable, and were then extensively used and studied in Nguyen et al. (2020); Liang and Klein (2009); Cappé (2011). Some improvements of those methods have been provided and analyzed, globally and in finite-time, in Karimi et al. (2019) where variance reduction techniques taken from the optimization literature have been efficiently applied to scale the EM algorithm to large datasets. Follow-up studies on variance reduced stochastic EM include Fort et al. (2020; 2021). Regarding the computation of the expectation under the posterior distribution, the Monte Carlo EM (MCEM) has been introduced in Wei and Tanner (1990) where a Monte Carlo (MC) approximation for this expectation is computed. A variant of that algorithm is the Stochastic Approximation of the EM (SAEM) in Delyon et al. (1999) leveraging the power of Robbins-Monro update (Robbins and Monro, 1951) to ensure pointwise convergence of the vector of estimated parameters using a decreasing stepsize rather than increasing the number of MC samples. The MCEM and the SAEM have been successfully applied in mixed effects models (McCulloch, 1997; Hughes, 1999; Baey et al., 2016) or to do inference for joint modeling of time-to-event data coming from clinical trials in Chakraborty and Das (2010), unsupervised clustering in Ng and McLachlan (2003), variational inference of graphical models in Blei et al. (2017) among other applications. An incremental variant of the SAEM was proposed in Kuhn et al. (2020) but its analysis is limited to asymptotic consideration. Gradient-based methods have been developed and analyzed in Zhu et al. (2017) but remain out of the scope of this paper as they tackle the high-dimensionality issue.

1.2 Contributions

This paper introduces and analyzes a new class of methods which purpose is to update two proxies for the target expected quantities in a two-timescale manner. Those approximated quantities are then used to optimize the objective function (1) for challenging examples (nonlinear) and settings (large-scale) using the M-step of the EM algorithm. Our main contributions can be summarized as follows:

- We propose a two-timescale method based on (i) stochastic approximation (SA), to alleviate the burden of computing MC approximations, and on (ii) incremental updates, scaling to large datasets. We describe the edges of each level of our method based on variance reduction argu-

ments. Such class of algorithms has two advantages. First, it naturally leverages variance reduction and Robbins-Monro type of updates to tackle large-scale and highly nonconvex learning tasks. Then, it gives a simple formulation as a scaled-gradient method which makes the analysis and implementation accessible.

- We also establish global (independent of the initialization) and finite-time (true at each iteration) upper bounds on a classical sub-optimality condition (Jain and Kar, 2017; Ghadimi and Lan, 2013), *i.e.,* the second order moment of the gradient of the objective function. We discuss the double dynamic of those bounds due to the two-timescale property of our algorithm update and we theoretically show the advantages of introducing variance reduction in a stochastic approximation (Robbins and Monro, 1951) scheme.

- Our theoretical findings include MC sampling noise contrary to existing studies related to the EM where the expectations are computed exactly. Adding a layer of MC approximation and the SA step to reduce its variance introduce some new challenges that need careful considerations and account for the originality of our research paper, both on the algorithmic and theoretical plans.

- Numerical experiments are presented in this contribution on a variety of models and datasets. In particular, we provide empirical insights on the edges of our method for learning latent variable models in image analysis and pharmacokinetics.

In Section 2 we formalize both incremental and Monte Carlo variants of the EM. We introduce our two-timescale class of EM (TTSEM) algorithms for which we derive several statistical guarantees in Section 3 for possibly nonconvex functions. Section 4 corresponds to the sketches of the proofs for our main results. Section 5 is devoted to the numerical experiments showing the benefits of our methods on several tasks and datasets. Proofs and additional experimental details are deferred to the Appendix.

2 Two-Timescale Stochastic EM Algorithms

We recall and formalize in this section the different methods found in the literature that aim at solving the intractable expectation problem and the large-scale problem. We then introduce our class of stochastic methods that efficiently tackles the optimization problem in (1).

2.1 Monte Carlo Integration and Stochastic Approximation

As mentioned in the introduction, for complex and possibly nonconvex models, the expectation under the posterior distribution defined in (4) is not tractable. In that case, the first solution involves computing a Monte Carlo integration of that expectation. For all $i \in [n]$, draw M samples, noted $\{z_{i,m} \sim p(z_i|y_i;\boldsymbol{\theta})\}_{m=1}^M$, and compute the MC integration, noted $\tilde{S}$, of $\bar{\mathbf{s}}(\boldsymbol{\theta})$ defined by (4):

$$\text{MC-step}: \quad \tilde{S} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}, y_i) \; . \tag{6}$$

Then, update the parameter via the maximization function $\overline{\boldsymbol{\theta}}(\tilde{S})$. This algorithm, called the MCEM (Wei and Tanner, 1990), bypasses the intractable expectation issue but is rather computationally expensive. Indeed, in order to reach pointwise convergence, the number of samples M needs to be increasingly large. An alternative to the MCEM is to use a Robbins-Monro (RM) type of update, see Robbins and Monro (1951). We denote, for $k > 0$, the number of samples M_k and the approximation by $\tilde{S}^{(k+1)}$:

$$\tilde{S}^{(k+1)} := \frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_k}\sum_{m=1}^{M_k} S(z_{i,m}^{(k)}, y_i) \; , \tag{7}$$

where for $m \in [M_k]$, $z_{i,m}^{(k)} \sim p(z_i|y_i;\boldsymbol{\theta}^{(k)})$. Then, the RM update of the sufficient statistics $\hat{\mathbf{s}}^{(k+1)}$ reads:

$$\text{SA-step}: \quad \hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \; , \tag{8}$$

where $\{\gamma_k\}_{k>1} \in (0,1)$ is a sequence of decreasing stepsizes to ensure asymptotic convergence. The combination of (7) and (8) is called the Stochastic Approximation of the EM (SAEM) and has been shown to converge to a maximum likelihood of the observations under very general conditions, see Delyon et al. (1999) for a proof of convergence. In simple scenarios, the samples $\{z_{i,m}\}_{m=1}^M$ are conditionally independent and identically distributed with distribution $p(z_i, \boldsymbol{\theta})$. Nevertheless, in most cases, since the loss function between the observed data y_i and the latent variable z_i can be nonconvex, sampling exactly from this distribution is not an option and the MC batch is sampled by Markov Chain Monte Carlo (MCMC) algorithm (Meyn and Tweedie, 2012; Brooks et al., 2011). It has been proved in Kuhn and Lavielle (2004) that (8) converges almost surely when coupled with an MCMC procedure.

Role of the stepsize γ_k: The sequence of decreasing positive integers $\{\gamma_k\}_{k>1}$ controls the convergence of the algorithm. It is inefficient to start with small values for the stepsize γ_k and large values for the number of simulations M_k. Rather, it is recommended that one decreases γ_k, as in $\gamma_k = 1/k^\alpha$, with

$\alpha \in (0,1)$, and keeps a constant and small number of samples M_k, hence bypassing the computationally involved sampling step in (6). In practice, γ_k is set equal to 1 during the first few iterations to let the iterates explore the parameter space without memory and converge quickly to a neighborhood of the target estimate. The Stochastic Approximation is performed during the remaining iterations ensuring the almost sure convergence of the vector of estimates. This Robbins-Monro type of update constitutes the first level of our algorithm, needed to temper the variance and noise introduced by the Monte Carlo integration. In the next section, we derive variants of this algorithm to adapt to the sheer size of data of modern applications and formalize the second level of our class of two-timescale EM methods.

2.2 Incremental and Two-Timescale Stochastic EM Methods

Efficient strategies to scale to large datasets include incremental (Neal and Hinton, 1998) and variance reduced (Johnson and Zhang, 2013; Chen et al., 2018) methods. We explicit a general update that covers those latter variants and that represents the second level of our algorithm, *i.e.,* the incremental update of the noisy statistics $\tilde{S}^{(k+1)}$ in (7). Instead of computing its full batch $\tilde{S}^{(k+1)}$ as in (7), the MC approximation is incrementally evaluated through $S_{\text{tts}}^{(k+1)}$ as:

$$\text{Inc-step}: \quad S_{\text{tts}}^{(k+1)} = S_{\text{tts}}^{(k)} + \rho_{k+1}(\boldsymbol{\mathcal{S}}^{(k+1)} - S_{\text{tts}}^{(k)}) \,. \tag{9}$$

Note that $\{\rho_k\}_{k>1} \in (0,1)$ is a sequence of stepsizes, $\boldsymbol{\mathcal{S}}^{(k)}$ is a proxy for $\tilde{S}^{(k)}$ defined in (7). If the stepsize ρ_k is equal to 1 and $\boldsymbol{\mathcal{S}}^{(k)} = \tilde{S}^{(k)}$, i.e., computed in a full batch manner as in (7), then we recover the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\boldsymbol{\mathcal{S}}^{(k)} = \tilde{S}^{(k)}$, then we recover the MCEM algorithm.

Two-Timescale Stochastic EM methods: We introduce the general method derived using the two variance reduction techniques described above. Beforehand, we list in Table 1, variants of the Inc-step, stated in (9), of Algorithm 1 for the quantity $\boldsymbol{\mathcal{S}}^{(k+1)}$, at iteration $k > 0$.

Table 1 Proxies for the Incremental-step (9)

1:	iSAEM	$\boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + n^{-1}(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)})$
2:	vrTTEM	$\boldsymbol{\mathcal{S}}^{(k+1)} = S_{\text{tts}}^{(\ell(k))} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))})$
3:	fiTTEM	$\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})$ and $\overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}(\tilde{S}_{j_k}^{(k)} - \tilde{S}_{j_k}^{(t_{j_k}^k)})$

Note that the proposed fiTTEM update, Line 3 in Table 1, draws two independent and uniform indices $(i_k, j_k) \in [n]$. Thus, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample

Algorithm 1 Two-Timescale Stochastic EM methods.

1: **Input:** $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathbf{s}}^{(0)} \leftarrow \tilde{S}^{(0)}$, $\{\gamma_k\}_{k>0}$, $\{\rho_k\}_{k>0}$ and $K_f \in \mathbb{N}^*$.

2: Set the terminating iteration number, $K \in \{0, \dots, K_f - 1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_f-1} \gamma_\ell} = \frac{\gamma_k}{P_m} \,. \tag{10}$$

3: **for** $k = 0, 1, 2, \dots, K_f - 1$ **do**

4: Draw index $i_k \in [n]$ uniformly (and $j_k \in [n]$ for fiTTEM).

5: Compute $\tilde{S}_{i_k}^{(k)}$ using the MC-step (6), for the drawn indices.

6: Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using Lines 1, 2 or 3 in Table 1.

7: Compute $\hat{\mathbf{s}}^{(k+1)}$ and $S_{\text{tts}}^{(k+1)}$ using resp. (8) and (9):

$$\begin{aligned} S_{\text{tts}}^{(k+1)} &= S_{\text{tts}}^{(k)} + \rho_{k+1}(\boldsymbol{\mathcal{S}}^{(k+1)} - S_{\text{tts}}^{(k)}) \,, \\ \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \,. \end{aligned} \tag{11}$$

8: Update $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$ via the M-step (5).

9: **end for**

$j \in [n]$ is last drawn as j_k prior to iteration k in addition to τ_i^k which was defined w.r.t. i_k. We recall that $\tilde{S}_{i_k}^{(k)} := \frac{1}{M_k} \sum_{m=1}^{M_k} S(z_{i_k,m}^{(k)}, y_{i_k})$ where $z_{i_k,m}^{(k)}$ are samples drawn from $p(z_{i_k}|y_{i_k}; \boldsymbol{\theta}^{(k)})$. The stepsize in (9) is set to $\rho_{k+1} = 1$ for the iSAEM method initializing with $\boldsymbol{\mathcal{S}}^{(0)} = \tilde{S}^{(0)}$; $\rho_{k+1} = \rho$ is constant for the vrTTEM and fiTTEM. Note that we initialize with $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \tilde{S}^{(0)}$ for the fiTTEM which can be seen as a slightly modified version of SAGA inspired by Reddi et al. (2016). For vrTTEM we set an epoch size of m and we define $\ell(k) := m\lfloor k/m \rfloor$ as the first iteration number in the epoch that iteration k is in.

Remarks on Table 1: For all methods, we define a random index noted $i_k \in [n]$ and drawn at iteration k, and $\tau_i^k = \max\{k' : i_{k'} = i,\ k' < k\}$ as the iteration index where $i \in [n]$ is last drawn prior to iteration k.

Then, our general class of methods, see Algorithm 1, leverages both levels (8) and (9) in order to output a vector of fitted parameters $\hat{\boldsymbol{\theta}}^{(K_f)}$ where K_f is the total number of iterations. The update in (11) is said to have a two-timescale property as the stepsizes satisfy $\lim_{k\to\infty} \gamma_k/\rho_k < 1$ such that $\tilde{S}^{(k+1)}$ is updated at a faster time-scale, determined by ρ_{k+1}, than $\hat{\mathbf{s}}^{(k+1)}$, determined by γ_{k+1}. The next section introduces the main results of this paper and establishes global and finite-time bounds for the three

different updates of our scheme.

3 Finite Time Analysis of Two-Timescale EMs

Notations reminder: $\tilde{S}$ represents the Monte Carlo approximation of its expected counterpart $\bar{\mathrm{s}}$ at index $i \in [n]$. S_{tts} denotes the variance-reduced quantity in (9), related to the stepsize ρ (assumed constant here), and leveraging the incrementally updated quantity $\boldsymbol{\mathcal{S}}$ via Table 1. The quantity noted $\hat{\mathrm{s}}$ stands for the sufficient statistics resulting from the RM procedure in (8) and is updated using the SA stepsize γ.

Following Cappé and Moulines (2009), it can be shown that stationary points of the objective function (1) corresponds to the stationary points of the following nonconvex Lyapunov function:

$$\min_{\mathbf{s}\in\mathsf{S}} V(\mathbf{s}) := \bar{\mathsf{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{L}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\ , \tag{12}$$

that we propose to study in this paper and where L_i is defined in (1).

3.1 Assumptions and Intermediate Lemmas

In order to derive the desired convergence guarantees, several important assumptions are given below:

A1. *The sets* Z, S *are compact. Besides, there exist constants* $C_{\mathsf{S}}, C_{\mathsf{Z}}$ *such that*

$$C_{\mathsf{S}} := \max_{\mathbf{s},\mathbf{s}'\in\mathsf{S}} \|\mathbf{s}-\mathbf{s}'\| < \infty \quad \textit{and} \quad C_{\mathsf{Z}} := \max_{i\in[n]} \int_{\mathsf{Z}} |S(z,y_i)|\mu(\mathrm{d}z) < \infty.$$

A2. *For any* $i \in [n]$, $z \in \mathsf{Z}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathrm{int}(\Theta)^2$, *where* $\mathrm{int}(\Theta)$ *denotes the interior of* Θ, *we have* $\left|p(z|y_i;\boldsymbol{\theta}) - p(z|y_i;\boldsymbol{\theta}')\right| \le \mathrm{L}_p \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$.

We recall that we consider curved exponential family models such that the objective function satisfies:

A3. *For any* $\boldsymbol{s} \in \mathsf{S}$, *the function* $\boldsymbol{\theta} \mapsto L(s,\boldsymbol{\theta}) := \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \mathbf{s}\,,\, \phi(\boldsymbol{\theta})\right\rangle$ *admits a unique global minimum* $\overline{\boldsymbol{\theta}}(\mathbf{s}) \in \mathrm{int}(\Theta)$. *In addition,* $\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))$, *the Jacobian of the function* ϕ *at* $\boldsymbol{\theta}$, *is full rank,* L_p*-Lipschitz and* $\overline{\boldsymbol{\theta}}(\mathbf{s})$ *is* L_t*-Lipschitz.*

We denote by $\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\theta})$ the Hessian (w.r.t to $\boldsymbol{\theta}$ for a given value of $\mathbf{s} \in \mathsf{S}$) of the function $\boldsymbol{\theta} \mapsto L(\mathbf{s},\boldsymbol{\theta}) = \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \mathbf{s}\,,\, \phi(\boldsymbol{\theta})\right\rangle$, and define $\mathrm{B}(\mathbf{s}) := \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\overline{\boldsymbol{\theta}}(\mathbf{s})))^{-1}\,\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top$.

A4. *It holds that* $\upsilon_{\max} := \sup_{\mathbf{s}\in\mathsf{S}} \|\,\mathrm{B}(\mathbf{s})\| < \infty$ *and* $0 < \upsilon_{\min} := \inf_{\mathbf{s}\in\mathsf{S}} \lambda_{\min}(\mathrm{B}(\mathbf{s}))$. *There exists a constant* L_b *such that for all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}^2$, *we have* $\|\,\mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\| \leq \mathrm{L}_b \,\|\mathbf{s} - \mathbf{s}'\|$.

The class of TTSEM methods, summarized in Algorithm 1, is composed of two levels where the second stage corresponds to the variance reduction trick used in Karimi et al. (2019) in order to accelerate incremental methods and reduce the variance introduced by the index sampling step. The first stage is the Robbins-Monro update that aims at reducing the Monte Carlo noise of $\tilde{S}^{(k+1)}$ at iteration k, defined as follows:

$$\eta_i^{(k)} := \tilde{S}_i^{(k)} - \bar{\mathbf{s}}_i(\vartheta^{(k)}) \quad \text{for all} \quad i \in [n] \quad \text{and} \quad k > 0\,. \tag{13}$$

We consider that the MC approximation is unbiased if for all $i \in [n]$ and $m \in [M]$, the samples $z_{i,m} \sim p(z_i|y_i;\boldsymbol{\theta})$ are i.i.d. under the posterior distribution, *i.e.,* $\mathbb{E}[\eta_i^{(k)}|\mathcal{F}_k] = 0$ where $\mathcal{F}_k$ is the filtration up to iteration k. The following results are derived under the assumption that the fluctuations implied by the approximation are bounded:

A5. *For all* $k > 0$, $i \in [n]$, *it holds that* $\mathbb{E}[\|\eta_i^{(k)}\|^2] < \infty$ *and* $\mathbb{E}[\|\mathbb{E}[\eta_i^{(k)}|\mathcal{F}_k]\|^2] < \infty$.

Note that typically, the controls exhibited above are vanishing when the number of MC samples M_k increases with k. We now state two important results on the Lyapunov function; its smoothness:

Lemma 1. *(Karimi et al., 2019) Assume A1-A4. For all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}$ *and* $i \in [n]$, *we have*

$$\|\bar{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \bar{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \,\|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \mathrm{L}_V \,\|\mathbf{s} - \mathbf{s}'\|\,, \tag{14}$$

where $\mathrm{L}_{\mathbf{s}} := C_{\mathsf{Z}}\,\mathrm{L}_p\,\mathrm{L}_t$ *and* $\mathrm{L}_V := \upsilon_{\max}(1 + \mathrm{L}_{\mathbf{s}}) + \mathrm{L}_b\,C_{\mathsf{S}}$.

We also establish a growth condition on the gradient of V related to the mean field of the algorithm:

Lemma 2. *Assume A3 and A4. For all* $\mathrm{s} \in \mathsf{S}$,

$$\upsilon_{\min}^{-1} \left\langle \nabla V(\mathrm{s}) \, , \, \mathrm{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) \right\rangle \geq \|\mathrm{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathrm{s})\|^2 \, . \tag{15}$$

We present in the following a finite-time and global, *i.e.,* independent of the initialization, analysis of both the incremental and two-timescale variants of our general method described in Algorithm 1.

3.2 Global Convergence of Incremental and Two-Timescale Stochastic EM

Then, the following non-asymptotic convergence rate can be derived for the iSAEM algorithm:

Theorem 1. *Assume A1-A5. Consider the iSAEM sequence* $\{\hat{\mathrm{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ *obtained with* $\rho_{k+1} = 1$ *for any* $k \leq K_f$ *where* $K_f > 0$. *Let* $\{\gamma_k = 1/(k^a \alpha c_1 \overline{L})\}_{k>0}$, *where* $a \in (0,1)$, *be a sequence of stepsizes,* $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = c_1 \overline{L}/n$, *then:*

$$\upsilon_{\max}^{-2} \sum_{k=0}^{K_f} \tilde{\alpha}_k \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(K_f)})] + \sum_{k=0}^{K_f - 1} \tilde{\Gamma}_k \mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \, .$$

Observe that, in Theorem 1, the convergence bound is composed of an initialization term $V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(K_f)})$ and suffers from the Monte Carlo noise introduced by the posterior sampling step, see the second term on the RHS of the inequality. We observe, in the next section, that when variance reduction is applied ($\rho_k < 1$), a second phase of convergence manifests.

We now deal with the analysis of Algorithm 1 when variance reduction is applied *i.e.,* $\rho < 1$. Let K be an independent discrete r.v. drawn from $\{1, \ldots, K_f\}$ with distribution $\{\gamma_{k+1}/P_m\}_{k=0}^{K_f - 1}$, then, for any $K_f > 0$, the convergence criterion used in our study reads

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2] = \frac{1}{P_m} \sum_{k=0}^{K_f - 1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \, ,$$

where $P_m := \sum_{\ell=0}^{K_f - 1} \gamma_\ell$ and the expectation above is taken over the overall randomness of the algorithm. Denote $\Delta V := V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(K_f)})$ and $\|\Delta S\|^2 := \|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k)}\|^2$. The vrTTEM method satisfies:

Theorem 2. *Assume A1-A5. Consider the vrTTEM sequence* $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ *for any* $k \leq K_f$ *where* K_f *is a positive integer. Let* $\{\gamma_{k+1} = 1/(k^a \overline{L})\}_{k>0}$*, where* $a \in (0,1)$*, be a sequence of stepsizes,* $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$*,* $\rho = \mu/(c_1 \overline{L} n^{2/3})$*,* $m = nc_1^2/(2\mu^2 + \mu c_1^2)$ *and a constant* $\mu \in (0,1)$*. Then:*

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2] \leq \frac{2n^{2/3}\overline{L}}{\mu P_m \upsilon_{\min}^2 \upsilon_{\max}^2}(\mathbb{E}[\Delta V] + \sum_{k=0}^{K_f-1} \tilde{\eta}^{(k+1)} + \chi^{(k+1)}\mathbb{E}[\|\Delta S\|^2]).$$

Furthermore, the fiTTEM method displays the following convergence rate:

Theorem 3. *Assume A1-A5. Consider the fiTTEM sequence* $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ *for any* $k \leq K_f$ *where* K_f *be a positive integer. Let* $\{\gamma_{k+1} = 1/(k^a \alpha c_1 \overline{L})\}_{k>0}$*, where* $a \in (0,1)$*, be a sequence of positive stepsizes,* $\alpha = \max\{2, 1 + 2\upsilon_{\min}\}$*,* $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$*,* $\beta = 1/(\alpha n)$*,* $\rho = 1/(\alpha c_1 \overline{L} n^{2/3})$ *and* $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$*. Then:*

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2] \leq \frac{4\alpha \overline{L} n^{2/3}}{P_m \upsilon_{\min}^2 \upsilon_{\max}^2} (\mathbb{E}[\Delta V] + \sum_{k=0}^{K_f - 1} \Xi^{(k+1)} + \Gamma^{(k+1)} \mathbb{E}[\|\Delta S\|^2]) \; .$$

Note that in those two bounds, $\tilde{\eta}^{(k+1)}$ and $\Xi^{(k+1)}$ depend only on the Monte Carlo noises $\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2]$, $\mathbb{E}[\|\mathbb{E}[\eta_i^{(r)}|\mathcal{F}_r]\|^2]$, bounded under assumption A5, and some constants.

Remarks: Theorem 2 and Theorem 3 exhibit in their convergence bounds two different phases. The upper bounds display a bias term due to the initial conditions, *i.e.,* the term ΔV, and a double dynamic burden exemplified by the term $\mathbb{E}[\|\Delta S\|^2$. Indeed, we remark the following: (i) This term is the price we pay for the two-timescale dynamic and corresponds to the gap between the two asynchronous updates (one on $\hat{\boldsymbol{s}}^{(k)}$ and the other on $\tilde{S}^{(k)}$). (ii) It is readily understood that if $\rho = 1$, *i.e.,* there is no variance reduction, then for any $k > 0$,

$$\mathbb{E}[\|\Delta S\|^2] = \mathbb{E}[\|\boldsymbol{\mathcal{S}}^{(k+1)} - S_{\mathrm{tts}}^{(k+1)}\|^2] = 0 \; ,$$

with $\hat{\boldsymbol{s}}^{(0)} = \tilde{S}^{(0)} = 0$, which strengthens the fact that this quantity characterizes the impact of the variance reduction technique introduced in our scheme. The following Lemma describes this gap:

Lemma 3. *Considering a decreasing stepsize* $\gamma_k \in (0,1)$ *and a constant* $\rho \in (0,1)$*, we have*

$$\mathbb{E}[\|\Delta S\|^2] \leq \frac{\rho}{1-\rho} \sum_{\ell=0}^{k} (1 - \gamma_\ell)^2 (\boldsymbol{\mathcal{S}}^{(\ell)} - S_{tts}^{(\ell)}) \; ,$$

where $\boldsymbol{\mathcal{S}}^{(\ell)}$ *is defined by Line 2 (vrTTEM) or 3 (fiTTEM).*

4 Proof Sketches

We provide in the sequel sketches of the proofs of our main Theorems along important auxiliary Lemmas used throughout the proofs.

4.1 Proof of Theorem 1

The main convergence result for the iSAEM algorithm, *i.e.,* Theorem 1, is derived under the control of the Monte Carlo fluctuations as described by assumption A5 and is built upon the following intermediary Lemma, characterizing the quantity of interest $S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}$ at each iteration index $k > 0$:

Lemma 4. *Assume A1. The iSAEM update (1) is equivalent to the following update on the statistics*

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \hat{\mathbf{s}}^{(k)}) \; .$$

Also:

$$\mathbb{E}[S_{tts}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E}[\overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + (1 - 1/n)\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}] + \frac{1}{n}\mathbb{E}[\eta_{i_k}^{(k+1)}] \; ,$$

where $\overline{\mathbf{s}}^{(k)}$ *is defined by (4) and* $\tau_i^k = \max\{k' : i_{k'} = i, \; k' < k\}$.

Proof. From update (1), we have:

$$S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = S_{\text{tts}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \frac{1}{n}(\tilde{S}_{i_k}^{(k+1)} - \tilde{S}_{i_k}^{(\tau_i^k)}) = \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + S_{\text{tts}}^{(k)} - \overline{\mathbf{s}}^{(k)} - \frac{1}{n}(\tilde{S}_{i_k}^{(\tau_i^k)} - \tilde{S}_{i_k}^{(k+1)}) \; .$$

Since $\tilde{S}_{i_k}^{(k+1)} = \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)}) + \eta_{i_k}^{(k+1)}$ we have

$$S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + S_{\text{tts}}^{(k)} - \overline{\mathbf{s}}^{(k)} - \frac{1}{n}(\tilde{S}_{i_k}^{(\tau_i^k)} - \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})) + \frac{1}{n}\eta_{i_k}^{(k+1)} \; .$$

Taking the full expectation of both side of the equation leads to:

$$\mathbb{E}[S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E}[\overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}] - \frac{1}{n}\mathbb{E}[\mathbb{E}[\tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})|\mathcal{F}_k]] + \frac{1}{n}\mathbb{E}[\eta_{i_k}^{(k+1)}] \; .$$

Since we have $\mathbb{E}[\tilde{S}_i^{(\tau_i^k)}|\mathcal{F}_k] = \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)}$ and $\mathbb{E}[\overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})|\mathcal{F}_k] = \overline{\mathbf{s}}^{(k)}$, we conclude the proof. □

We derive the following Lemma which establishes an upper bound of the quantity $\mathbb{E}[\|S_{\text{tts}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2]$, another important quantity in order to characterize the convergence of our incremental scheme.

Lemma 5. *For any* $k \geq 0$ *and consider the iSAEM update in (1), it holds that*

$$\mathbb{E}[\|S_{tts}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \leq 4\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{2\,\mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] + 2\frac{c_\eta}{M_k} + 4\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\|^2] \; .$$

Proof. Applying the iSAEM update yields:

$$\begin{aligned}\mathbb{E}[\|S_{\text{tts}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] =& \mathbb{E}[\|S_{\text{tts}}^{(k)} - \hat{\boldsymbol{s}}^{(k)} - \frac{1}{n}(\tilde{S}_{i_k}^{(\tau_i^k)} - \tilde{S}_{i_k}^{(k)})\|^2] \\ \leq& 4\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\|^2] + 4\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{2}{n^2}\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] + 2\frac{c_\eta}{M_k} \,.\end{aligned}$$

The last expectation can be further bounded by

$$\frac{2}{n^2}\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{2}{n^3}\sum_{i=1}^n \mathbb{E}[\|\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{(t_i^k)}\|^2] \overset{(a)}{\leq} \frac{2\,\mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \,,$$

where (a) is due to Lemma 1 and which concludes the proof of the Lemma. □

Proof of Theorem 1: Having established those two auxiliary results, we now give a proof sketch for Theorem 1. We consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ obtained with $\rho_{k+1} = 1$ via Algorithm 1 and Line 1 of Table 1. Under the classical smoothness assumption of the Lyapunov function V (cf. Lemma 1), Lemma 4 yields:

$$\begin{aligned}\mathbb{E}[\langle S_{\text{tts}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,,\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle] \leq& (\upsilon_{\max}^2\frac{\beta(n-1)+1}{2n} - \upsilon_{\min})\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\ &+ (1-\frac{1}{n})/(2\beta)\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\|^2] + \frac{1}{2n}\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \,,\end{aligned}$$

where the inequality is due to the growth condition (2) and Young's inequality (with $\beta \to 1$). Besides,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n}\sum_{i=1}^n (\frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2]) \,,$$

where the equality holds as i_k and j_k are drawn independently. For any $\beta > 0$, it holds

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \leq& \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}\|^2 \\ &+ \gamma_{k+1}\beta\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2] \,,\end{aligned}$$

where the last inequality is due to Young's inequality. Subsequently, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\tau_i^{k+1})}\|^2] \\ \leq& \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n^2}\sum_{i=1}^n \mathbb{E}[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}\|^2] \,.\end{aligned}$$

Applying Lemma 5 gives

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)}-\hat{\boldsymbol{s}}^{(\tau_i^{k+1})}\|^2] \leq 4(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta})\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\hat{\boldsymbol{s}}^{(k)}\|^2]+2(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta})\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2]$$
$$+4(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta})\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)}-\overline{\mathbf{s}}^{(k)}\|^2]$$
$$+\sum_{i=1}^{n}\mathbb{E}[\frac{1-\frac{1}{n}+\gamma_{k+1}\beta+\frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})}{n}\|\hat{\boldsymbol{s}}^{(k)}-\hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \ ,$$

and define the following quantity

$$\Delta^{(k)} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2] \ .$$

Setting $c_1=\upsilon_{\min}^{-1}$, $\alpha=\max\{8,1+6\upsilon_{\min}\}$, $\overline{L}=\max\{\mathrm{L}_{\mathbf{s}},\mathrm{L}_V\}$, $\gamma_{k+1}=\frac{1}{k\alpha c_1\overline{L}}$, $\beta=\frac{c_1\overline{L}}{n}$, we have that $c_1(k\alpha-1)\geq c_1(\alpha-1)\geq 6$ and observe that

$$1-\frac{1}{n}+\gamma_{k+1}\beta+\frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta}) \leq 1-\frac{c_1(k\alpha-1)-4}{k\alpha n c_1} \leq 1-\frac{2}{k\alpha n c_1} \ ,$$

which shows that $1-\frac{1}{n}+\gamma_{k+1}\beta+\frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})\in(0,1)$ for any $k>0$. Denote $\Lambda_{(k+1)}=\frac{1}{n}-\gamma_{k+1}\beta-\frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})$ and note that $\Delta^{(0)}=0$, thus the telescoping sum yields:

$$\Delta^{(k+1)}$$
$$\leq 4\sum_{\ell=0}^{k}\prod_{j=\ell+1}^{k}(1-\Lambda_{(j)})(\gamma_{\ell+1}^2+\frac{\gamma_{\ell+1}}{\beta})\mathbb{E}[\|\overline{\boldsymbol{s}}^{(\ell)}-\hat{\boldsymbol{s}}^{(\ell)}\|^2]+2\sum_{\ell=0}^{k}\prod_{j=\ell+1}^{k}(1-\Lambda_{(j)})(\gamma_{\ell+1}^2+\frac{\gamma_{\ell+1}}{\beta})\mathbb{E}[\|\eta_{i_\ell}^{(\ell)}\|^2]$$
$$+4\sum_{\ell=0}^{k}\prod_{j=\ell+1}^{k}(1-\Lambda_{(j)})(\gamma_{\ell+1}^2+\frac{\gamma_{\ell+1}}{\beta})\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^\ell)}-\overline{\mathbf{s}}^{(\ell)}\|^2] \ .$$

Note $\omega_{k,\ell}=\prod_{j=\ell+1}^{k}(1-\Lambda_{(j)})$ Summing on both sides over $k=0$ to $k=K_m-1$ and upper bounding the quantity $\sum_{k=0}^{K_m-1}\Delta^{(k+1)}$ leads to the combination of the above equations and yields:

$$\sum_{k=0}^{K_m-1}\tilde{\alpha}_k\mathbb{E}[\|\overline{\mathbf{s}}^{(k)}-\hat{\boldsymbol{s}}^{(k)}\|^2]+\sum_{k=0}^{K_m-1}\tilde{\beta}_k\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)}-\overline{\mathbf{s}}^{(k)}\|^2] \leq \mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)})-V(\hat{\boldsymbol{s}}^{(K)})]+\sum_{k=0}^{K_m-1}\tilde{\Gamma}_k\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \ ,$$

where the various quantities are provided in the Appendix for the sake of clarity. For any $k>0$, $\tilde{\alpha}_k\geq 0$, we have by Lemma 2 that:

$$\sum_{k=0}^{K_m}\tilde{\alpha}_k\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \upsilon_{\max}^2\sum_{k=0}^{K_m}\tilde{\alpha}_k\mathbb{E}[\|\overline{\mathbf{s}}^{(k)}-\hat{\boldsymbol{s}}^{(k)}\|^2] \ ,$$

which yields an upper bound of the gradient of the Lyapunov function V and concludes the proof.

4.2 Proof of Theorem 2

We first derive an identity for the drift term of the vrTTEM :

Lemma 6. *Consider the vrTTEM update (2) with $\rho_k = \rho$, it holds for all $k > 0$*

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - S_{tts}^{(k+1)}\|^2] \leq & 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2\,\mathrm{L}_{\mathbf{s}}^2\,\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] \\ & + 2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - S_{tts}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2]\,,\end{aligned}$$

where we recall that $\ell(k)$ is the first iteration number in the epoch that iteration k is in.

Proof. Beforehand, we provide a rewiriting of the quantity $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}$ that will be useful throughout this proof:

$$\begin{aligned}\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)S_{\text{tts}}^{(k)} - \rho\boldsymbol{\mathcal{S}}^{(k+1)}) \\ &= -\gamma_{k+1}((1-\rho)[\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k)}] + \rho[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}])\,.\end{aligned} \tag{16}$$

We observe, using the identity (16), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] + 2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - S_{\text{tts}}^{(k)}\|^2]. \tag{17}$$

For the latter term, we obtain its upper bound as

$$\begin{aligned}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] =& \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}(\overline{\boldsymbol{s}}_i^{(k)} - \tilde{S}_i^{\ell(k)}) - (\overline{\boldsymbol{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))})\|^2] \\ \overset{(a)}{\leq}& \mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \overset{(b)}{\leq} \mathrm{L}_{\mathbf{s}}^2\,\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2]\,,\end{aligned}$$

where (a) uses the variance inequality and (b) uses Lemma 1. Substituting into (17) proves the lemma. □

Proof of Theorem 2: Similar arguments using the smoothness of the Lyapunov function as above are used at the beginning of the following proof. The main different argument when dealing with two-timescale methods, rather than incremental ones, is in the construction of the following sequence:

$$R_k := \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)}) + b_k\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2]\,, \tag{18}$$

where for $k > 0$, $b_k := \overline{b}_{k \bmod m}$ is a periodic sequence where:

$$\overline{b}_i = \overline{b}_{i+1}(1 + \gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2) + \gamma_{k+1}^2\rho^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2,\quad i = 0, 1, \ldots, m-1 \;\text{ with }\; \overline{b}_m = 0\,.$$

Note that $\bar{b}_i$ is decreasing with i and this implies

$$\bar{b}_i \leq \bar{b}_0 = \gamma_{k+1}^2 \rho^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2 \, \frac{(1+\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2 \, \mathrm{L}_{\mathbf{s}}^2)^m - 1}{\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2 \, \mathrm{L}_{\mathbf{s}}^2}, \; i = 1, 2, \ldots, m \; .$$

For $k+1 \leq \ell(k) + m$, we have the following inequality

$$\begin{aligned} R_{k+1} \leq \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}(\rho \upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \, \mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2))\mathbb{E}[\|\mathsf{h}_k\|^2] \\ + \underbrace{(b_{k+1}(1+\gamma\beta + 2\gamma^2\rho^2 \, \mathrm{L}_{\mathbf{s}}^2) + \gamma^2\rho^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2)}_{= b_k \quad \text{since } k+1 \leq \ell(k)+m} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] + \tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)} \; , \end{aligned}$$

where we have used Lemma 6. Then, using Lemma 2, that for any γ_{k+1}, ρ and β such that $\rho \upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \, \mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2) > 0$,

$$\begin{aligned} \upsilon_{\max}^2 \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \leq & \frac{R_k - R_{k+1}}{\gamma_{k+1}(\rho \upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \, \mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2))} \\ & + \frac{\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}}{\gamma_{k+1}(\rho \upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \, \mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2))} \; . \end{aligned}$$

We first remark that

$$\gamma_{k+1}(\rho \upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \, \mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)) \geq \frac{\gamma_{k+1}\rho}{c_1}(1 - \gamma_{k+1}c_1\rho \, \mathrm{L}_V - b_{k+1}(\frac{c_1}{\beta} + 2\gamma_{k+1}\rho c_1)) \; ,$$

where $c_1 = \upsilon_{\min}^{-1}$. By setting $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = \frac{c_1\overline{L}}{n^{1/3}}$, $\rho = \frac{\mu}{c_1\overline{L}n^{2/3}}$, $m = \frac{nc_1^2}{2\mu^2 + \mu c_1^2}$ and $\{\gamma_{k+1}\}$ any sequence of decreasing stepsizes in $(0,1)$, it can be shown that there exists $\mu \in (0,1)$, such that the following lower bound holds

$$1 - \gamma_{k+1}c_1\rho \, \mathrm{L}_V - b_{k+1}(\frac{c_1}{\beta} + 2\gamma_{k+1}\rho c_1) \overset{(a)}{\geq} 1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{\mu}{c_1^2}(\mathrm{e}-1)(1 + \frac{2\mu}{n}) \geq 1 - \mu - \mu(1+2\mu)\frac{\mathrm{e}-1}{c_1^2} \overset{(b)}{\geq} \frac{1}{2} \; ,$$

where the simplification in (a) is due to

$$\frac{\mu}{n} \leq \gamma\beta + 2\gamma^2 \, \mathrm{L}_{\mathbf{s}}^2 \leq \frac{\mu}{n} + \frac{2\mu^2}{c_1^2 n^{\frac{4}{3}}} \leq \frac{\mu c_1^2 + 2\mu^2}{c_1^2}\frac{1}{n} \;\; \text{and} \;\; (1 + \gamma\beta + 2\gamma^2 \, \mathrm{L}_{\mathbf{s}}^2)^m \leq \mathrm{e} - 1 \; ,$$

where the required μ in (b) can be found by solving the quadratic equation. Noting that $R_0 = \mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)})]$ and if K_m is a multiple of m, then $R_{max} = \mathbb{E}[V(\hat{\boldsymbol{s}}^{(K_m)})]$, hence concluding our proof.

4.3 Proof of Theorem 3

We begin with the statement and proofs of two required Lemmas. First, an equivalent update of Line (3) is given below for the purpose of the proof.

Lemma 7. *At iteration $k+1$,the drift term of update (3), with $\rho_{k+1} = \rho$, is equivalent to the following :*

$$\begin{aligned}\hat{\boldsymbol{s}}^{(k)} - S_{tts}^{(k+1)} =& \rho(\hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}) + \rho\eta_{i_k}^{(k+1)} + \rho[(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}]] \\ &+ (1-\rho)\left(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right) ,\end{aligned}$$

where we recall that $\eta_{i_k}^{(k+1)}$, defined in (13), which is the gap between the MC approximation and the expected statistics.

Proof. Using the fiTTEM update $S_{\text{tts}}^{(k+1)} = (1-\rho)S_{\text{tts}}^{(k)} + \rho\boldsymbol{\mathcal{S}}^{(k+1)}$ where $\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})$ leads to the following decomposition:

$$S_{\text{tts}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = \rho(\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}) + \rho\eta_{i_k}^{(k+1)} - \rho[(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}]] + (1-\rho)\left(S_{\text{tts}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right) ,$$

where we observe that $\mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}] = \overline{\mathbf{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)}$ and which concludes the proof.

Important Note: Note that $\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not equal to $\eta_{i_k}^{(k+1)}$, defined in (13), which is the gap between the MC approximation and the expected statistics. Indeed $\tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not computed under the same model as $\overline{\mathbf{s}}_{i_k}^{(k)}$. □

Then, we derive an identity for the quantity $\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}\|^2]$ using the fiTTEM update:

Lemma 8. *Consider the fiTTEM update (3) with $\rho_k = \rho$. It holds for all $k > 0$ that*

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - S_{tts}^{(k+1)}\|^2] \leq& 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2\frac{\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\ &+ 2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - S_{tts}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] ,\end{aligned}$$

where $\mathrm{L}_{\mathbf{s}}$ is the smoothness constant defined in Lemma 1.

Proof. Beforehand, we provide a rewiriting of the quantity $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}$ as follows:

$$\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = -\gamma_{k+1}((1-\rho)[\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k)}] + \rho[\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)} - (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})]) . \tag{19}$$

We observe, using the identity (19), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] + 2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - S_{\text{tts}}^{(k)}\|^2] . \tag{20}$$

For the latter term, we obtain its upper bound as

$$\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] = \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}(\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{\mathcal{S}}}_i^{(k)}) - (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})\|^2] \overset{(a)}{\leq} \mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \,,$$

where (a) uses the variance inequality. We can further bound the last expectation using Lemma 1:

$$\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{(t_i^k)}\|^2] \overset{(a)}{\leq} \frac{\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \,.$$

Substituting into (20) proves the lemma. □

Proof of Theorem 3: Using the smoothness of V and update (3), we obtain:

$$V(\hat{\boldsymbol{s}}^{(k+1)}) \leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1}\langle \hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)} \,, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}\|^2 \,. \tag{21}$$

Denote $\mathsf{H}_{k+1} := \hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}$ the drift term of the fiTTEM update in (8) and $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}$. Using Lemma 7 and the additional following identity $\mathbb{E}[(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}]] = 0$, we have

$$\begin{aligned}\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \leq & -(\upsilon_{\min}\gamma_{k+1}\rho + \gamma_{k+1}\upsilon_{\max}^2)\mathbb{E}[\|\mathsf{h}_k\|^2] - \frac{\gamma_{k+1}\rho^2}{2}\xi^{(k+1)} - \frac{\gamma_{k+1}(1-\rho)^2}{2}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ & + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\|\mathsf{H}_{k+1}\|^2 \,,\end{aligned}$$

where $\xi^{(k+1)} = \mathbb{E}[\|\mathbb{E}[\eta_{i_k}^{(k+1)}|\mathcal{F}_k]\|^2]$. The remaining of the proof is similar to the one for the vrTTEM algorithm. It consists of bounding the terms (i) $\mathbb{E}[\|\mathsf{H}_{k+1}\|^2]$ and (ii) $\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]$ using respectively Lemma 8 and noting that $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k+1)}) = -\gamma_{k+1}\mathsf{H}_{k+1}$. We also recall that in expectation $\mathbb{E}[\mathsf{H}_{k+1}|\mathcal{F}_k] = \rho\mathsf{h}_k + \rho\mathbb{E}[\eta_{i_k}^{(k+1)}|\mathcal{F}_k] + (1-\rho)\mathbb{E}[S_{\text{tts}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}]$ where $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}$. As for the iSAEM method, an important step of our proof is to define the following quantity

$$\Delta^{(k)} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \,,$$

where we recall that $t_j^k = \{k' : j_{k'} = j, k' < k\}$ is the iteration index where the sample $j \in [n]$ is last drawn as j_k prior to iteration k in addition to τ_i^k which was defined w.r.t. i_k, since fiTTEM update in Line 3 requires two independently drawn indices. Then, from the bounds on (i) and (ii), we obtain

$$\begin{aligned}\Delta^{(k+1)} \leq & \left(1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2\right)\Delta^{(k)} + \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\ & + \gamma_{k+1}(1-\rho)^2\left(2\gamma_{k+1} + \frac{1}{\beta}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \gamma_{k+1}\left(2\gamma_{k+1} + \frac{\rho^2}{\beta}\right)\mathbb{E}[\left\|\eta_{i_k}^{(k+1)}\right\|^2] \,.\end{aligned}$$

Setting $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{2, 1 + 2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L_s}, \mathrm{L}_V\}$, $\gamma_{k+1} = \frac{1}{k}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \overline{L} n^{2/3}}$, then we have that $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$. Hence, we observe

$$1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 \, \mathrm{L}_{\mathbf{s}}^2 \leq 1 - \frac{1}{n} + \frac{1}{\alpha k n} + \frac{1}{\alpha^2 c_1^2 k^2 n^{\frac{4}{3}}} \leq 1 - \frac{c_1(k\alpha - 1) - 1}{k\alpha n c_1} \leq 1 - \frac{1}{k\alpha n c_1}$$

showing that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 \, \mathrm{L}_{\mathbf{s}}^2 \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} = \frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2 \rho^2 \, \mathrm{L}_{\mathbf{s}}^2$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

$$\begin{aligned}
\Delta^{(k+1)} \leq & \sum_{\ell=0}^{k} \omega_{k,\ell} \left(2\gamma_{\ell+1}^2 \rho^2 + \frac{\gamma_{\ell+1}^2 \rho^2}{\beta} \right) \mathbb{E}[\left\| \overline{\mathbf{s}}^{(\ell)} - \hat{\boldsymbol{s}}^{(\ell)} \right\|^2] \\
& + \sum_{\ell=0}^{k} \omega_{k,\ell} \gamma_{\ell+1} (1 - \rho)^2 \left(2\gamma_{\ell+1} + \frac{1}{\beta} \right) \mathbb{E}[\left\| \tilde{S}^{(\ell)} - \hat{\boldsymbol{s}}^{(\ell)} \right\|^2] + \sum_{\ell=0}^{k} \omega_{k,\ell} \gamma_{\ell+1} \tilde{\epsilon}^{(\ell+1)} \, ,
\end{aligned}$$

where $\omega_{k,\ell} = \prod_{j=\ell+1}^{k} (1 - \Lambda_{(j)})$ and $\tilde{\epsilon}^{(\ell+1)} = \left(2\gamma_{k+1} + \frac{\rho^2}{\beta} \right) \mathbb{E}[\left\| \eta_{i_k}^{(k+1)} \right\|^2]$. Summing over the total number of iterations, making assumptions on the different hyperparameters of the algorithm and injecting in the smoothness inequality (21) leads to similar final steps of the proofs as for the two other methods detailed above. For completeness, we refer readers to Appendix where our proofs are explained in greater detail.

5 Numerical Applications

This section presents several numerical applications for our proposed class of Algorithms 1. The broad range of potential applications for our scheme include Gaussian Mixture Modeling, deformable template image analysis and nonlinear mixed-effects modeling. For each example, we provide the formulation of the chosen model, detail the explicit updates for the various training methods, including the baselines, and run numerical experiments along with visual plots showing the benefits of our proposed methods.

5.1 Gaussian Mixture Models

We begin by a simple and illustrative example. The authors acknowledge that the following model can be trained using deterministic EM-type of algorithms but propose to apply stochastic methods, including theirs, in order to compare their performances. Given n observations $\{y_i\}_{i=1}^n$, the goal here is to fit a Gaussian Mixture Model (GMM) whose distribution is modeled as a mixture of M Gaussian components, each with a unit variance. Let $z_i \in [M]$ be the latent labels of each component, the complete

log-likelihood is defined as follows:

$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{m=1}^{M} \mathbb{1}_{\{m\}}(z_i)[\log(\omega_m) - \mu_m^2/2] + \sum_{m=1}^{M} \mathbb{1}_{\{m\}}(z_i)\mu_m y_i + \text{constant} \ ,$$

where $\boldsymbol{\theta} := (\boldsymbol{\omega}, \boldsymbol{\mu})$ with $\boldsymbol{\omega} = \{\omega_m\}_{m=1}^{M-1}$ are the mixing weights with the convention $\omega_M = 1 - \sum_{m=1}^{M-1} \omega_m$ and $\boldsymbol{\mu} = \{\mu_m\}_{m=1}^{M}$ are the means. We use the penalization $\mathrm{r}(\boldsymbol{\theta}) = \frac{\delta}{2} \sum_{m=1}^{M} \mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$ where $\delta > 0$ and $\mathrm{Dir}(\cdot; M, \epsilon)$ is the M dimensional symmetric Dirichlet distribution with concentration parameter $\epsilon > 0$. The constraint set is given by $\Theta = \{\omega_m,\ m = 1, ..., M-1 : \omega_m \geq 0,\ \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R},\ m = 1, ..., M\}$.

EM updates: We first recognize that the constraint set for $\boldsymbol{\theta}$ is given by $\Theta = \Delta^M \times \mathbb{R}^M$. Using the partition of the sufficient statistics as $S(y_i, z_i) = (S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, the partition $\phi(\boldsymbol{\theta}) = (\phi^{(1)}(\boldsymbol{\theta})^\top, \phi^{(2)}(\boldsymbol{\theta})^\top, \phi^{(3)}(\boldsymbol{\theta}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$ and the fact that $\mathbb{1}_{\{M\}}(z_i) = 1 - \sum_{m=1}^{M-1} \mathbb{1}_{\{m\}}(z_i)$, the complete data log-likelihood can be expressed as in (3) with

$$\begin{aligned} s_{i,m}^{(1)} &= \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{\log(\omega_m) - \frac{\mu_m^2}{2}\right\} - \left\{\log(1 - \textstyle\sum_{j=1}^{M-1}\omega_j) - \frac{\mu_M^2}{2}\right\} \ , \\ s_{i,m}^{(2)} &= \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m \ , \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M \ , \end{aligned} \tag{22}$$

and $\psi(\boldsymbol{\theta}) = -\left\{\log(1 - \sum_{m=1}^{M-1} \omega_m) - \frac{\mu_M^2}{2\sigma^2}\right\}$. We also define for each $m \in [M]$, $j \in \{1, 2, 3\}$, $s_m^{(j)} = n^{-1} \sum_{i=1}^{n} s_{i,m}^{(j)}$. Consider the following latent sample used to compute an approximation of the conditional expected value $\mathbb{E}_{\boldsymbol{\theta}}[1_{\{z_i=m\}}|y = y_i]$:

$$z_{i,m} \sim \mathbb{P}\left(z_i = m | y_i; \boldsymbol{\theta}\right) \ , \tag{23}$$

where $m \in [M]$, $i \in [n]$ and $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\mu}) \in \Theta$. In particular, given iteration $k+1$, the computation of the approximated quantity $\tilde{S}_{i_k}^{(k)}$ during the Inc-step updates, see (9), can be written as

$$\tilde{S}_{i_k}^{(k)} = (\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}) y_{i_k}, ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1}) y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\bar{\mathbf{s}}_{i_k}^{(3)}(\boldsymbol{\theta}^{(k)})})^\top. \tag{24}$$

Recall the regularizer $\frac{\delta}{2} \sum_{m=1}^{M} \mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$ we used, which also reads:

$$\mathrm{r}(\boldsymbol{\theta}) = \tfrac{\delta}{2} \textstyle\sum_{m=1}^{M} \mu_m^2 - \epsilon \sum_{m=1}^{M} \log(\omega_m) - \epsilon \log(1 - \sum_{m=1}^{M-1} \omega_m) \ . \tag{25}$$

It can be shown that the regularized M-step evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1}(s_1^{(1)} + \epsilon, \ldots, s_{M-1}^{(1)} + \epsilon)^\top \\ ((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \ldots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)})^\top \\ (1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta)^{-1}(s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} \ , \tag{26}$$

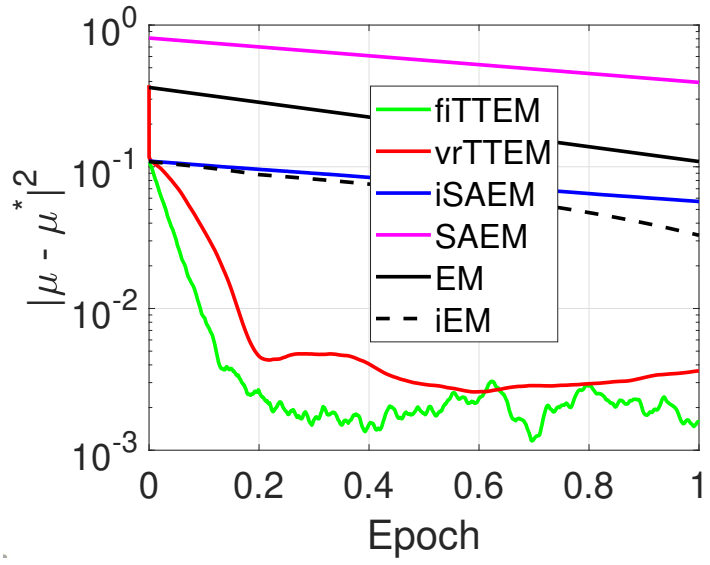

Figure 1: Precision $|\mu^{(k)} - \mu^*|^2$ for our methods (fiTTEM in green,vrTTEM in black and iSAEM in red) versus deterministic baselines (EM in dashed blue line, iEM in solid blue line) or stochastic baseline (SAEM in solid red line) against epochs elapsed. Our two variance reduced methods, *i.e.,* fiTTEM and vrTTEM are reaching the highest accuracy.

where we have defined for all $m \in [M]$ and $j \in \{1, 2, 3\}$, $s_m^{(j)} = n^{-1} \sum_{i=1}^n s_{i,m}^{(j)}$.

Synthetic data experiment: In the following experiments on synthetic data, we generate 50 synthetic datasets of size $n = 10^5$ from a GMM model with $M = 2$ components of means $\mu_1 = -\mu_2 = 0.5$.

We run the EM method until convergence (to double precision) to obtain the ML estimate $\mu^\star$ averaged on 50 datasets. We compare the EM, iEM (incremental EM), SAEM, iSAEM, vrTTEM and fiTTEM methods in terms of their precision measured by $|\mu - \mu^\star|^2$. We set the stepsize of the SA-step for all method as $\gamma_k = 1/k^\alpha$ with $\alpha = 0.5$, and the stepsize ρ_k for the vrTTEM and the fiTTEM to a constant stepsize equal to $1/n^{2/3}$. The number of MC samples is fixed to $M = 10$. Figure 1 shows the precision $|\mu - \mu^*|^2$ for the different methods through the epoch(s) (one epoch equals n iterations). The vrTTEM and fiTTEM methods outperform the other stochastic methods, supporting the benefits of our scheme.

Model Assumptions: We use the GMM example to illustrate the required assumptions. Many practical

models can satisfy the compactness of the sets as in assumption A1. For instance, the GMM example satisfies the conditions in A1 as the sufficient statistics are composed of indicator functions and observations as defined in (22). Assumptions A2 and A3 are standard for the curved exponential family models. For GMM, the following (strongly convex) regularization $\mathrm{r}(\boldsymbol{\theta})$ ensures A3:

$$\mathrm{r}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \epsilon\sum_{m=1}^{M}\log(\omega_m) - \epsilon\log(1-\sum_{m=1}^{M-1}\omega_m)\ ,$$

since it ensures $\boldsymbol{\theta}^{(k)}$ is unique and lies in $\mathrm{int}(\Delta^M)\times\mathbb{R}^M$. We remark that for A2, it is possible to define the Lipschitz constant L_p independently for each data y_i to yield a refined characterization. Again, A4 is satisfied by practical models. For GMM, it can be verified by deriving the closed form expression for $\mathrm{B}(\mathrm{s})$ and using A1. Under A1 and A3, we have $\|\hat{\boldsymbol{s}}^{(k)}\| < \infty$ since S is compact and $\hat{\boldsymbol{\theta}}^{(k)} \in \mathrm{int}(\Theta)$ for any $k \geq 0$ which thus ensure that the EM methods operate in a closed set throughout the optimization.

Algorithms updates: In the sequel, recall that, for all $i \in [n]$ and iteration k, the computed statistic $\tilde{S}_{i_k}^{(k)}$ is defined by (24). At iteration k, the several E-steps defined by (1) or (2) and (3) leads to the definition of the quantity $\hat{\mathrm{s}}^{(k+1)}$. Define the exact conditional expected value $\mathbb{E}_{\boldsymbol{\theta}}[1_{\{z_i=m\}}|y=y_i]$ as follows:

$$\widetilde{\omega}_m(y_i;\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[1_{\{z_i=m\}}|y=y_i] = \frac{\omega_m\exp(-\frac{1}{2}(y_i-\mu_i)^2)}{\sum_{j=1}^{M}\omega_j\exp(-\frac{1}{2}(y_i-\mu_j)^2)}\ .$$

Then, for the GMM example, after the initialization of the quantity $\hat{\mathrm{s}}^{(0)} = n^{-1}\sum_{i=1}^{n}\bar{\mathrm{s}}_i^{(0)}$, the E-step explicit updates are listed Table 2.

Table 2 Algorithms Updates for GMM

1:	Batch EM (EM) for all $i \in [n]$, compute $\bar{\mathrm{s}}_i^{(k)}$ and set $\hat{\mathrm{s}}^{(k+1)} = n^{-1}\sum_{i=1}^{n}\bar{\mathrm{s}}_i^{(k)}$
2:	Incremental EM (iEM) draw i_k uniformly at random on $[n]$, compute $\bar{\mathrm{s}}_{i_k}^{(k)}$ and set $\hat{\mathrm{s}}^{(k+1)} = n^{-1}\sum_{i=1}^{n}\bar{\mathrm{s}}_i^{(k)}$
3:	Batch SAEM (SAEM) for all $i \in [n]$ compute $\tilde{S}_i^{(k)}$ (24) and set $\hat{\mathrm{s}}^{(k+1)} = \hat{\mathrm{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}S_{\mathrm{tts}}^{(k)}$
4:	Variance Reduced Two-Timescale EM (vrTTEM) draw i_k uniformly at random on $[n]$, compute $\tilde{S}_{i_k}^{(k)}$ via (24) and set $\hat{\mathrm{s}}^{(k+1)} = \hat{\mathrm{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}(S_{\mathrm{tts}}^{(k)}(1-\rho) + \rho(\tilde{S}^{(\ell(k))} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))})))$
5:	Fast Incremental Two-Timescale EM (fiTTEM) draw i_k uniformly at random on $[n]$, compute $\tilde{S}_{i_k}^{(k)}$ via (24) and set $\hat{\mathrm{s}}^{(k+1)} = \hat{\mathrm{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}(S_{\mathrm{tts}}^{(k)}(1-\rho) + \rho(\overline{\boldsymbol{\mathcal{S}}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}))$.

Finally, the k-th update reads $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathrm{s}}^{(k+1)})$ where the function $\boldsymbol{s} \to \overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined by (26).

5.2 Deformable Template Model for Image Analysis

Model and EM Updates: Let $(y_i, i \in [n])$ be observed gray level images defined on a grid of pixels. Let $u \in \mathcal{U} \subset \mathbb{R}^2$ denote the pixel index on the image and $x_u \in \mathcal{D} \subset \mathbb{R}^2$ its location. The model used in this experiment suggests that each image y_i is a deformation of a template, noted $I : \mathcal{D} \to \mathbb{R}$, common to all images of the dataset:

$$y_i(u) = I\left(x_u - \Phi_i\left(x_u, z_i\right)\right) + \varepsilon_i(u) \ , \tag{27}$$

where $\Phi_i : \mathbb{R}^2 \to \mathbb{R}^2$ is a deformation function, z_i some latent variable parameterizing this deformation and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an observation error. The template model, given $\{p_k\}_{k=1}^{k_p}$ landmarks on the template, a fixed known kernel $\mathbf{K_p}$ and a vector of parameters $\beta \in \mathbb{R}^{k_p}$ is defined as follows:

$$I_\xi = \mathbf{K_p}\beta, \quad \text{where} \quad \left(\mathbf{K_p}\beta\right)(x) = \sum_{k=1}^{k_p} \mathbf{K_p}\left(x, p_k\right)\beta_k \ .$$

Given a set of landmarks $\{g_k\}_{k=1}^{k_g}$ and a fixed kernel $\mathbf{K_g}$, we parameterize the deformation Φ_i as:

$$\Phi_i = \mathbf{K_g} z_i \quad \text{where} \quad \left(\mathbf{K_g} z_i\right)(x) = \sum_{k=1}^{k_s} \mathbf{K_g}\left(x, g_k\right)\left(z_i^{(1)}(k), z_i^{(2)}(k)\right) \ ,$$

where we put a Gaussian prior on the latent variables, $z_i \sim \mathcal{N}(0, \Gamma)$ and $z_i \in \left(\mathbb{R}^{k_g}\right)^2$. Hence, the vector of parameters we want to estimate is $\boldsymbol{\theta} = (\beta, \Gamma, \sigma)$.

The complete model belongs to the curved exponential family, see Allassonnière et al. (2007), and its vector of sufficient statistics, noted $S = (S_1(z), S_2(z), S_3(z))$, reads:

$$S_1(z) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top y_i, \quad S_2(z) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top \left(\mathbf{K}_p^{z_i}\right), \quad S_3(z) = \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i \ , \tag{28}$$

where for any pixel $u \in \mathbb{R}^2$ and $j \in [k_g]$ we denote:

$$\mathbf{K}_p^{z_i}(x_u, j) = \mathbf{K}_p^{z_i}(x_u - \phi_i(x_u, z_i), p_j) \ .$$

Finally, the Two-Timescale M-step yields the following parameter updates:

$$\bar{\boldsymbol{\theta}}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n}\hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z)\beta(\hat{s}) - 2\beta(\hat{s})\hat{s}_1(z) \end{pmatrix} ,$$

where $\hat{s} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via the SA-step (8) and using the MC approximation of the sufficient statistics $(S_1(z), S_2(z), S_3(z))$ defined in (28).

Numerical Experiment on the U.S. Postal Service database: We apply model (27) and our Algorithm 1 to a collection of handwritten digits, called the US postal database (Hull, 1994), featuring $n = 1\,000$, (16×16)-pixel images for each class of digits from 0 to 9. The main challenge with this dataset stems from the geometric dispersion within each class of digit as shown Figure 2 for digit 5. Hence, we ought to use our deformable template model (27) in order to account for both sources of variability, *i.e.,* the intrinsic template of each class of digit and the small and local deformations in each observed image.

Figure 2: Training set of the USPS database (20 images for digit 5). The variability within a class in the dataset we consider for this experiment is exhibited here. We can observe that all the images present particular features both in terms of sharpness and shape.

Figure 3 shows the resulting synthetic images for digit 5 through several epochs, for the batch method, the online SAEM, the incremental SAEM and the various two-timescale methods.

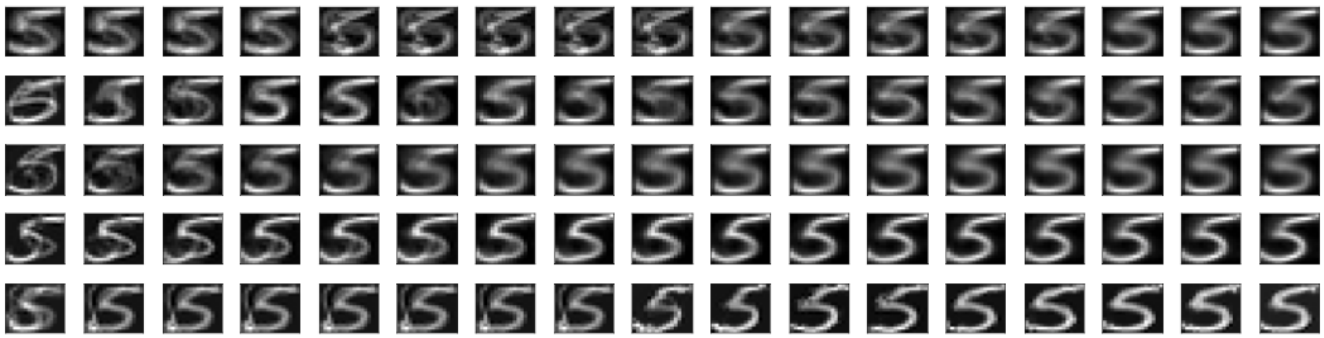

Figure 3: (USPS Digits) Estimation of the template. From top to bottom: batch, online, iSAEM, vrTTEM and fiTTEM through 7 epochs. Batch method templates are replicated in-between epochs for a fair comparison with incremental variants.

For all methods, the initialization of the template (28) is the mean of the gray level images. In our experiments, we have chosen Gaussian kernels for both, $\mathbf{K_p}$ and $\mathbf{K_g}$, defined on $\mathbb{R}^2$ and centered on the landmark points$\{p_k\}_{k=1}^{k_p}$ and $\{g_k\}_{k=1}^{k_g}$ with standard respective standard deviations of 0.12 and 0.3. We set $k_p = 15$ and $k_g = 6$ equidistributed landmarks points on the grid for the training procedure. The hyperparameters are kept the same and are set as $M = 400$, $\gamma_k = 1/k^{0.6}$ and $p = 16$. The standard deviation of the measurement errors is set to 0.1. Those hyperparameters are inspired by relevant studies (Allassonnière et al., 2010; 2013). For the sampling phase of our methods, we use the Carlin and Chib MCMC procedure, see Carlin and Chib (1995), refer to Maire et al. (2017) for more details.

In particular, the choice of the geometric covariance, indexed by g, in our study is critical since it has a direct impact on the sharpness of the templates. As for the photometric hyperparameter, indexed by p, both the template and the geometry are impacted, in the sense that with a large photometric variance, the kernel centered on one landmark spreads out to many of its neighbors.

As the iterations proceed, the templates become progressively sharper. Figure 3 displays the virtue of the vrTTEM and fiTTEM methods leading to a more contrasted and accurate template estimate. The incremental and online versions are better in the very first epochs compared to the batch method, given the high computational cost of the latter. After a few epochs, the batch SAEM estimates similar template as the incremental and online methods due to their high variance. Our variance reduced and fast incremental variants are effective in the long run and sharpen the template estimates contrasting between the background and the regions of interest in the image.

5.3 Pharmacokinetics (PK) Model with Absorption Lag Time

The following numerical example deals characterizes the pharmacokinetics (PK) of orally administered drug to simulated patients, using a population approach, *i.e.,* the training set consists of numerous drug plasmatic concentration per patient of the cohort. Specifically, $M = 50$ synthetic datasets were generated for $n = 5000$ patients with 10 observations (concentration measures) per patient. The goal is to model the evolution of the concentration of the absorbed drug using a nonlinear and latent variable model. We consider a one-compartment PK model for oral administration with an absorption lag-time (T^{lag}), assuming first-order absorption and linear elimination processes.

Model and Explicit Updates: The final model includes the following variables: ka the absorption rate constant, V the volume of distribution, k the elimination rate constant and T^{lag} the absorption lag-time. We also add several covariates to our model such as D the dose of drug administered, t the time at which measures are taken and the weight of the patient influencing the volume V. More precisely, the log-volume $\log(V)$ is a linear function of the log-weight $lw70 = \log(wt/70)$. Let $z_i = (T_i^{\text{lag}}, ka_i, V_i, k_i)$ be the vector of individual PK parameters, different for each individual i. The final model reads:

$$y_{ij} = f(t_{ij}, z_i) + \varepsilon_{ij} \quad \text{where} \quad f(t_{ij}, z_i) = \frac{D\, ka_i}{V(ka_i - k_i)}(\mathrm{e}^{-ka_i\,(t_{ij}-T_i^{\text{lag}})} - \mathrm{e}^{-k_i\,(t_{ij}-T_i^{\text{lag}})}) \;, \tag{29}$$

where y_{ij} is the j-th concentration measurement of the drug of dosage D injected at time t_{ij} for patient i. We assume in this example that the residual errors ε_{ij} are independent and normally distributed with mean 0 and variance σ^2. Lognormal distributions are used for the four PK parameters:

$$\begin{aligned}
&\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega^2_{T^{\text{lag}}}), \quad \log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega^2_{ka}) \;, \\
&\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega^2_{V}), \quad \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega^2_{k}) \;.
\end{aligned}$$

We note that the complete model $p(y, z)$ defined by the structural model in (29) belongs to the curved exponential family, which vector of sufficient statistics $S = (S_1(z), S_2(z), S_3(z))$ reads:

$$S_1(z) = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad S_2(z) = \frac{1}{n}\sum_{i=1}^{n} z_i^\top z_i, \quad S_3(z) = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(t_i, z_i))^2 \;, \tag{30}$$

where we have noted y_i and t_i the vector of observations and time for each patient $i \in [n]$. At iteration k, and setting the number of MC samples to 1 for the sake of clarity, the MC sampling $z_i^{(k)} \sim p(z_i|y_i; \boldsymbol{\theta}^{(k)})$ is performed using a Metropolis-Hastings procedure detailed in Algorithm 2. The quantities $S_{\text{tts}}^{(k+1)}$ and

$\hat{\mathrm{s}}^{(k+1)}$ are then updated according to the different methods introduced in our paper, see Table 1. Finally the maximization step yields:

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} \hat{\mathbf{s}}_1^{(k+1)} \\ \hat{\mathbf{s}}_2^{(k+1)} - \hat{\mathbf{s}}_1^{(k+1)} \left(\hat{\mathbf{s}}_1^{(k+1)}\right)^\top \\ \hat{\mathbf{s}}_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{z}_{\mathbf{pop}}}(\hat{\mathbf{s}}^{(k+1)}) \\ \overline{\boldsymbol{\omega}_{\boldsymbol{z}}}(\hat{\mathbf{s}}^{(k+1)}) \\ \overline{\boldsymbol{\sigma}}(\hat{\mathbf{s}}^{(k+1)}) \end{pmatrix} , \tag{31}$$

where z_{pop} denotes the vector of fixed effects $(T_{\text{pop}}^{\text{lag}}, ka_{\text{pop}}, V_{\text{pop}}, k_{\text{pop}})$.

Monte Carlo study: We conduct a Monte Carlo study to showcase the benefits of our scheme. $M = 50$ datasets have been simulated using the following PK parameters values: $T_{\text{pop}}^{\text{lag}} = 1$, $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$, $\omega_{T^{\text{lag}}} = 0.4$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. We define the mean square distance over the M replicates as $E_k(\ell) = \frac{1}{M}\sum_{m=1}^{M}\left(\boldsymbol{\theta}_k^{(m)}(\ell) - \boldsymbol{\theta}^*\right)^2$, and plot it against the epochs (passes over the data) in Figure 4. Note that the MC-step (6) is performed using a Metropolis-Hastings procedure since the posterior distribution under the model $\boldsymbol{\theta}$ noted $p(z_i|y_i;\boldsymbol{\theta})$ is intractable, mainly due to the nonlinearity of the model (29). The Metropolis-Hastings (MH) algorithm (Meyn and Tweedie, 2012) leverages a proposal distribution $q(z_i, \delta)$ where $\boldsymbol{\theta} = (z_{\text{pop}}, \omega_z)$ and δ is the vector of parameters of the proposal distribution. Generally, and for simplicity, a Gaussian proposal is used. The MH algorithm employed to sample from each individual posterior distribution $(p(z_i|y_i;\boldsymbol{\theta}), i \in [n])$ is summarized in Algorithm 2.

Algorithm 2 Metropolis-Hastings algorithm

1: **Input:** initialization $z_{i,0} \sim q(z_i;\boldsymbol{\delta})$
2: **for** $m = 1, \cdots, M$ **do**
3: Sample $z_{i,m} \sim q(z_i;\boldsymbol{\delta})$.
4: Sample $u \sim \mathcal{U}([0,1])$.
5: Calculate the ratio $r = \frac{\pi(z_{i,m};\boldsymbol{\theta})/q(z_{i,m};\boldsymbol{\delta})}{\pi(z_{i,m-1};\boldsymbol{\theta})/q(z_{i,m-1});\boldsymbol{\delta})}$.
6: **if** $u < r$ **then** accept $z_{i,m}$ **else** $z_{i,m} \leftarrow z_{i,m-1}$
7: **end for**
8: **Output:** $z_{i,M}$

Figure 4 shows clear advantage of variance reduced methods (vrTTEM and fiTTEM) avoiding the twists and turns displayed by the incremental and the batch methods (iSAEM and SAEM).

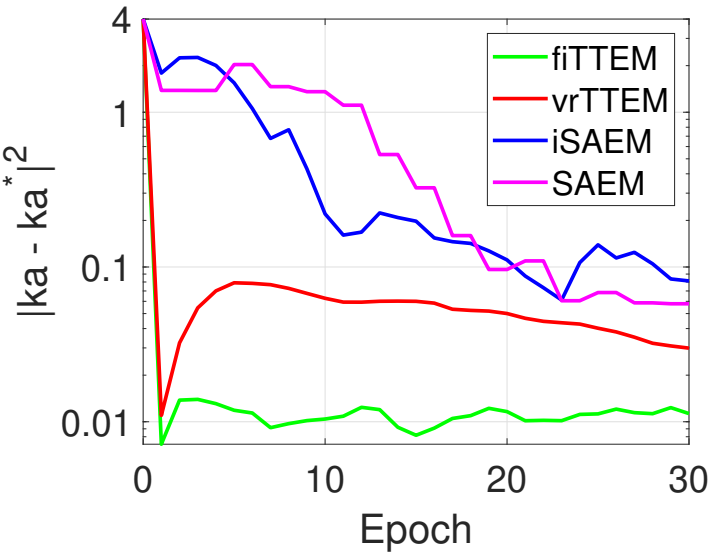

Figure 4: Mean square errors $|ka^{(k)} - ka^*|^2$ for our methods (fiTTEM in green,vrTTEM in black and iSAEM in blue) versus the SAEM baseline in red, against epochs elapsed. The errors have been averaged over $M = 50$ synthetic datasets for robustness. The fiTTEM appears to be the best method among the four. The other variance reduced TTSEM method, namely vrTTEM is quickly reaching a similar accuracy but exhibits overfitting rather quickly. The two other plain SAEM methods are the slowest.

Both our newly proposed EM methods quickly reaches a neighborhood of the solution while baselines slowly converge to it empirically stressing on the benefits of our two-timescale methods that not only temper the noise of the incremental update but also reduce the MC noise stemming from a required approximation of the expectations.

6 Conclusion

This paper introduces a new class of two-timescale EM methods for learning latent variable models. In particular, the models dealt with in this paper belong to the curved exponential family and are possibly nonconvex. The nonconvexity of the problem is tackled using a Robbins-Monro type of update, which represents the first level of our class of methods. The scalability with the number of samples is performed through a variance reduced and incremental update, the second and last level of the scheme we introduce in this paper. The various algorithms are interpreted as scaled gradient methods, in the space of the sufficient statistics, and our convergence results are global, in the sense of independence of the

initial values, and non-asymptotic, *i.e.,* true for any termination iteration index. A panoply of numerical examples illustrate the benefits of our scheme on synthetic and real datasets.

7 Supplemental Material

The supplementary material of this paper can be consulted in a separate file and contains the proofs for our theoretical results. Along the proofs of our main Theorems detailed Section 3, we also provide the statements and proofs of important intermediary Lemmas.

References

Allassonnière, S., Amit, Y., and Trouvé, A. (2007). Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29.

Allassonnière, S., Bigot, J., Glaunès, J. A., Maire, F., and Richard, F. J. (2013). Statistical models for deformable templates in image and shape analysis. *Annales mathématiques Blaise Pascal*, 20(1):1–35.

Allassonnière, S., Kuhn, E., and Trouvé, A. (2010). Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678.

Baey, C., Trevezas, S., and Cournède, P.-H. (2016). A non linear mixed effects model of plant growth and estimation via stochastic variants of the EM algorithm. *Communications in Statistics-Theory and Methods*, 45(6):1643–1669.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Cappé, O. (2011). Online EM algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749.

Cappé, O. and Moulines, E. (2009). On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484.

Chakraborty, A. and Das, K. (2010). Inferences for joint modelling of repeated ordinal scores and time to event data. *Computational and mathematical methods in medicine*, 11(3):281–295.

Chen, J., Zhu, J., Teh, Y. W., and Zhang, T. (2018). Stochastic expectation maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988.

Delyon, B., Lavielle, M., and Moulines, É. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242.

Fort, G., Moulines, É., and Wai, H.-T. (2020). A stochastic path integral differential estimator expectation maximization algorithm. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16972–16982. Curran Associates, Inc.

Fort, G., Moulines, É., and Wai, H.-T. (2021). Geom-spider-em: Faster variance reduced stochastic expectation maximization for nonconvex finite-sum optimization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3135–3139. IEEE.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Hughes, J. P. (1999). Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55(2):625–629.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554.

Jain, P. and Kar, P. (2017). Non-convex optimization for machine learning. *Found. Trends Mach. Learn.*, 10(3-4):142–336.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Karimi, B., Wai, H.-T., Moulines, É., and Lavielle, M. (2019). On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems*, pages 2833–2843.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131.

Kuhn, E., Matias, C., and Rebafka, T. (2020). Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Stat. Comput.*, 30(6):1725–1739.

Liang, P. and Klein, D. (2009). Online EM for unsupervised models. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 611–619, Boulder, CO.

Maire, F., Moulines, É., and Lefebvre, S. (2017). Online EM for functional data. *Comput. Stat. Data Anal.*, 111:27–47.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Ng, S. and McLachlan, G. J. (2003). On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Stat. Comput.*, 13(1):45–55.

Nguyen, H. D., Forbes, F., and McLachlan, G. J. (2020). Mini-batch learning of exponential family finite mixture models. *Stat. Comput.*, 30(4):731–748.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. (2016). Fast incremental method for smooth nonconvex optimization. In *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, pages 1971–1977, Las Vegas, NV.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Wei, G. C. and Tanner, M. A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103.

Zhu, R., Wang, L., Zhai, C., and Gu, Q. (2017). High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 4180–4188, Sydney, Australia.