Fast Two-Timescale Stochastic EM Algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Expectation-Maximization (EM) algorithm is a popular choice for learning latent variable models. Variants of the EM have been initially introduced by [28], using incremental updates to scale to large datasets, and by [33, 12], 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 are also presented to illustrate our findings.

1 Introduction

Learning latent variable models is critical for modern machine learning problems, see (e.g.,) [26] for references. We formulate the training of such model as an empirical risk minimization problem:

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

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 function $\overline{\mathsf{L}}(\boldsymbol{\theta})$ is possibly *nonconvex* and is assumed to be lower bounded. In the latent variable model, the likelihood $g(y_i;\boldsymbol{\theta})$, is the marginal of the complete data likelihood defined as $f(z_i, y_i;\boldsymbol{\theta})$, *i.e.,* $g(y_i;\boldsymbol{\theta}) = \int_{\mathsf{Z}} f(z_i, y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i)$, where $\{z_i\}_{i=1}^n$ are the latent variables. In this paper, we assume that the complete model belongs to the curved exponential family [14]:

$$f(z_i, y_i;\boldsymbol{\theta}) = h(z_i, y_i)\exp\big(\big\langle S(z_i, y_i) \,|\, \phi(\boldsymbol{\theta})\big\rangle - \psi(\boldsymbol{\theta})\big) \;, \tag{2}$$

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 [13, 34], the method of reference for (1), is comprised of two steps. The E-step computes the conditional expectation of the sufficient statistics of (2), noted $\overline{\mathbf{s}}(\boldsymbol{\theta})$:

$$\text{E-step: } \overline{\mathbf{s}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \overline{\mathbf{s}}_i(\boldsymbol{\theta}) \quad \text{where} \quad \overline{\mathbf{s}}_i(\boldsymbol{\theta}) = \int_{\mathsf{Z}} S(z_i, y_i) p(z_i|y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i) \;, \tag{3}$$

and the M-step is given by

$$\text{M-step: } \hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\overline{\mathbf{s}}(\boldsymbol{\theta})) := \operatorname*{arg\,min}_{\vartheta\in\Theta} \big\{ \mathrm{r}(\vartheta) + \psi(\vartheta) - \big\langle \overline{\mathbf{s}}(\boldsymbol{\theta}) \,|\, \phi(\vartheta)\big\rangle \big\} \;. \tag{4}$$

Two caveats of this method are the following: (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 (3) intractable. So far, and to the best of our knowledge, both challenges have been addressed separately, as detailed in the sequel.

Prior Work: Inspired by stochastic optimization procedures, [28] and [8] develop 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 [30, 23, 7]. Some improvements of those methods have been provided and analyzed, globally and in finite-time, in [20] where variance reduction techniques taken from the optimization literature have been efficiently applied to scale the EM algorithm to large datasets. Regarding the computation of the expectation under the posterior distribution, the Monte Carlo EM (MCEM) has been introduced in the seminal paper [33] where an MC approximation for this expectation is computed. A variant of that algorithm is the Stochastic Approximation of the EM (SAEM) in [12] leveraging the power of Robbins-Monro update [32] 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 [25, 16, 4] or to do inference for joint modeling of time to event data coming from clinical trials in [10], unsupervised clustering in [29], variational inference of graphical models in [5] among other applications. Recently, an incremental variant of the SAEM was proposed in [22] showing positive empirical results but its analysis is limited to asymptotic consideration.

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 modern examples and settings using the M-step of the EM algorithm. The main contributions of the paper are:

- We propose a two-timescale method based on (i) Stochastic Approximation (SA), to alleviate the problem of computing MC approximations, and on (ii) Incremental updates, to scale to large datasets. We describe in details the edges of each level of our method based on variance reduction arguments. 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 nonlinear learning tasks. Then, it gives a simple formulation as a *scaled-gradient method* which makes the global analysis and the 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 in the nonconvex literature [18, 15], *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 stress the advantages of introducing variance reduction in a *Stochastic Approximation* [32] scheme.

In Section 2 we formalize both incremental and Monte Carlo variants of the EM. Then, we introduce our two-timescale class of EM algorithms for which we derive several global statistical guarantees in Section 3 for possibly *nonconvex* functions. Section 4 is devoted to numerical illustrations. The supplementary material of this paper includes proofs of our theoretical results.

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 and the large-scale problem. We then provide the general framework of our method that efficiently tackles the optimization problem (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 (3) is not tractable. In that case, the first solution involves computing a Monte Carlo integration of that latter. For all $i \in [n]$, where $[n] := \{1, \cdots, n\}$, draw $\{z_{i,m} \sim p(z_i|y_i;\theta)\}_{m=1}^M$ samples and compute the MC integration of S_{MC} of $\overline{\mathbf{s}}(\boldsymbol{\theta})$ defined by (3):

$$\textsf{MC-step}: \; S_{\text{MC}} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}, y_i) \; . \tag{5}$$

Then update the parameter via the maximization function $\overline{\boldsymbol{\theta}}(S_{\text{MC}})$. This algorithm bypasses the intractable expectation issue but is rather computationally expensive in order to reach point wise convergence (M needs to be large). An alternative to that stochastic algorithm is to use a Robbins-Monro (RM) type of update. We denote, at iteration k, the number of samples M_k and the following MC approximation $\tilde{S}^{(k+1)}$:

$$S_{\text{MC}}^{(k+1)} := \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{M_k}\sum_{m=1}^{M_k} S(z_{i,m}^{(k)}, y_i) \quad \text{where} \quad z_{i,m}^{(k)} \sim p(z_i|y_i;\theta^{(k)}) \; . \tag{6}$$

Then, the RM update of the sufficient statistics $\hat{\mathbf{s}}^{(k+1)}$ reads:

$$\textsf{SA-step}: \; \hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(S_{\text{MC}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \; , \tag{7}$$

where $\{\gamma_k\}_{k>1} \in (0,1)$ is a sequence of decreasing stepsizes to ensure asymptotic convergence. This 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 [12]. In simple scenarios, the samples $\{z_{i,m}\}_{m=1}^{M}$ are conditionally independent and identically distributed with distribution $p(z_i, \theta)$. 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 [27, 6]. It has been proved in [21] that (7) converges almost surely when coupled with an MCMC sampling 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 M_k bypassing the computationally involved sampling step in (5). 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 today's applications and formalize the *second level* of our class of two-timescale EM methods.

2.2 Incremental and Two-Stage Stochastic EM Methods

Efficient strategies to scale to large datasets include incremental [28] and variance reduced [11, 19] methods. We will 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 $S_{\text{MC}}^{(k+1)}$ in (6). Instead of computing its full batch, the MC approximation is evaluated as:

$$\textsf{Incremental-step}: \; \tilde{S}^{(k+1)} = \tilde{S}^{(k)} + \rho_{k+1}\big(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k)}\big) \; . \tag{8}$$

Note that $\{\rho_k\}_{k>1} \in (0,1)$ is a sequence of stepsizes, $\boldsymbol{\mathcal{S}}^{(k)}$ is a proxy for $S_{\text{MC}}^{(k)}$. If the stepsize is equal to 1 and the proxy $\boldsymbol{\mathcal{S}}^{(k)} = S_{\text{MC}}^{(k)}$, i.e., computed in a full batch manner as in (6), then we recover the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\boldsymbol{\mathcal{S}}^{(k)} = S_{\text{MC}}^{(k)}$, then we recover the MCEM [33].

Remarks on Table 1: For all methods, we define a random index drawn at iteration k, noted $i_k \in [n]$, 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. The proposed fiTTEM method draws *two* indices *independently* and uniformly as $i_k, j_k \in [n]$. Thus, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample $j \in [n]$ is last drawn as j_k prior to iteration k in addition to τ_i^k which was defined *w.r.t.* i_k. Recall $\tilde{S}_{i_k}^{(k)} = \frac{1}{M_k}\sum_{m=1}^{M_k} S(z_{i_k,m}^{(k)}, y_{i_k})$ where $z_{i_k,m}^{(k)}$ are samples drawn from $p(z_{i_k}|y_{i_k};\theta^{(k)})$. The stepsize in (8) is set to $\rho_{k+1} = 1$ for

Table 1 Proxies for the Incremental-step (8)

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

the iSAEM method and we initialize with $\boldsymbol{\mathcal{S}}^{(0)} = \tilde{S}^{(0)}$; $\rho_{k+1} = \rho$ is constant for the vrTTEM and fiTTEM methods. Note that we initialize as follows $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \tilde{S}^{(0)}$ for the fiTTEM which can be seen as a slightly modified version of SAGA inspired by [31]. 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.

Two-Timescale Stochastic EM methods: We now introduce the general method derived using the two variance reduction techniques described above. Algorithm 1 leverages both levels (7) and (8) in order to output a vector of fitted parameters $\hat{\boldsymbol{\theta}}^{(\mathsf{K_m})}$ where $\mathsf{K_m}$ is the total number of iterations.

Algorithm 1 Two-Timescale Stochastic EM methods.

Input: $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathrm{s}}^{(0)} \leftarrow \tilde{S}^{(0)}$, $\{\gamma_k\}_{k>0}$, $\{\rho_k\}_{k>0}$ and $\mathsf{K_m} \in \mathbb{N}^*$.
for $k = 0, 1, 2, \ldots, \mathsf{K_m} - 1$ **do**
    Draw index $i_k \in [n]$ uniformly (and $j_k \in [n]$ for fiTTEM).
    Compute $\tilde{S}_{i_k}^{(k)}$ using the MC-step (5), for the drawn indices.
    Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using Lines 1, 2 or 3 in Table 1.
    Compute $\tilde{S}^{(k+1)}$ and $\hat{\mathrm{s}}^{(k+1)}$ using respectively (8) and (7):

$$
\begin{aligned}
\tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho_{k+1}\big(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k)}\big) \\
\hat{\mathrm{s}}^{(k+1)} &= \hat{\mathrm{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathrm{s}}^{(k)})
\end{aligned}
\tag{9}
$$

    Compute $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathrm{s}}^{(k+1)})$ via the M-step.
end for

The update in (9) 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 $\hat{\mathrm{s}}^{(k+1)}$, 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 the Two-Timescale Scheme

Following [8], it can be shown that stationary points of the objective function (1) corresponds to the stationary points of the following *nonconvex* Lyapunov function:

$$
\min_{\mathrm{s} \in \mathsf{S}} V(\mathrm{s}) := \overline{\mathrm{L}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathrm{s})) + \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathrm{s})) \,, \tag{10}
$$

that we propose to study in this article.

3.1 Assumptions and Intermediate Lemmas

Several important assumptions required to derive convergence guarantees read as follows:

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

$$
C_{\mathsf{S}} := \max_{\mathrm{s},\mathrm{s}' \in \mathsf{S}} \|\mathrm{s} - \mathrm{s}'\| < \infty, \quad C_{\mathsf{Z}} := \max_{i\in[n]} \int_{\mathsf{Z}} |S(z, y_i)| \mu(\mathrm{d}z) < \infty. \tag{11}
$$

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

We also recall that we consider curved exponential family models assuming the following:

A3. *For any $s \in \mathsf{S}$, the function $\boldsymbol{\theta} \mapsto L(s, \boldsymbol{\theta}) := \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \langle \mathrm{s} \,|\, \phi(\boldsymbol{\theta}) \rangle$ admits a unique global minimum $\overline{\boldsymbol{\theta}}(\mathrm{s}) \in \mathrm{int}(\Theta)$.*

In addition, $\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))$, the Jacobian of the function ϕ at $\boldsymbol{\theta}$, is full rank, L_p-Lipschitz and $\overline{\boldsymbol{\theta}}(\mathrm{s})$ is L_t-Lipschitz.

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

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 algorithms we develop in this paper is composed of two levels where the second stage corresponds to the variance reduction trick used in [20] in order to accelerate incremental methods and reduce the variance introduced by the index sampling. The first stage is the Robbins-Monro update that aims at reducing the Monte Carlo noise of $\tilde{S}^{(k+1)}$ at iteration k denoted as follows:

$$\eta_i^{(k)} := \tilde{S}_i^{(k)} - \overline{\mathbf{s}}_i(\vartheta^{(k)}) \quad \text{for all} \quad i \in [n] \quad \text{and} \quad k > 0 \,. \tag{12}$$

For instance, we consider that the MC approximation is unbiased if for all $i \in [n]$ and $m \in [M]$, the samples $z_{i,m} \sim p(z_i|y_i;\theta)$ 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:* $\mathbb{E}[\|\eta_i^{(k)}\|^2] < \infty$ *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. *[20] Assume A1-A4. For all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}$ *and* $i \in [n]$*, we have*

$$\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \, \|\mathbf{s} - \mathbf{s}'\|, \;\; \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \mathrm{L}_V \, \|\mathbf{s} - \mathbf{s}'\| \,, \tag{13}$$

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

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* $\mathbf{s} \in \mathsf{S}$*,*

$$\upsilon_{\min}^{-1} \big\langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \big\rangle \geq \|\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2 \,. \tag{14}$$

We present in the following sections a finite-time and global (independent of the initialization) analysis of both the incremental and two-timescale variants our method.

3.2 Global Convergence of Incremental Stochastic EM Algorithms

The following result for the iSAEM algorithm is derived under the control of the Monte Carlo fluctuations as described by Assumption A5 and is built upon an intermediary Lemma, characterizing the quantity of interest $(\hat{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)})$:

Lemma 3. *Assume A1. The iSAEM update (1) is equivalent to the following update on the statistics* $\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}\big(\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \hat{\mathbf{s}}^{(k)}\big)$*. Also:*

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

where $\overline{\mathbf{s}}^{(k)}$ *is defined by (3) and* $\tau_i^k = \max\{k' : i_{k'} = i, \; k' < k\}$.

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

Theorem 1. *Assume A1-A5. Consider the iSAEM sequence* $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ *obtained with* $\rho_{k+1} = 1$ *for any* $k \leq \mathsf{K}_{\mathsf{m}}$ *where* K_{m} *is a positive integer. 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:*

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

Note that, in Theorem 1, the convergence bound is composed of an initialization term $V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(\mathsf{K}_{\mathsf{m}})})$ 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 will be included in our bounds.

3.3 Global Convergence of Two-Timescale Stochastic EM Algorithms

We now deal with the analysis of Algorithm 1 when variance reduction is applied *i.e.,* $\rho < 1$. Two important intermediate Lemmas are needed in order to establish finite-time bounds for the vrTTEM and the fiTTEM methods. We first derive an identity for the drift term of the vrTTEM :

Lemma 4. *Consider the vrTTEM update (2) with $\rho_k = \rho$, it holds for all $k > 0$*

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

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

The second one derives an identity for the quantity $\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2]$ using the fiTTEM update:

Lemma 5. *Consider the fiTTEM update (3) with $\rho_k = \rho$. It holds for all $k > 0$ that*

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

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

Let K be an independent discrete r.v. drawn from $\{1, \dots, \mathsf{K}_\mathsf{m}\}$ with distribution $\{\gamma_{k+1}/\mathsf{P}_\mathsf{m}\}_{k=0}^{\mathsf{K}_\mathsf{m}-1}$, then, for any $\mathsf{K}_\mathsf{m} > 0$, the convergence criterion used in our study reads

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

where $\mathsf{P}_\mathsf{m} = \sum_{\ell=0}^{\mathsf{K}_\mathsf{m}-1} \gamma_\ell$ and the expectation is over the stochasticity of the algorithm. Denote $\Delta V = V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(\mathsf{K}_\mathsf{m})})$. We now state the main result regarding the vrTTEM method:

Theorem 2. *Assume A1-A5. Consider the vrTTEM sequence $\{\hat{\mathrm{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K}_\mathsf{m}$ where K_m 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:*

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

Furthermore, the fiTTEM method has the following convergence rate:

Theorem 3. *Assume A1-A5. Consider the fiTTEM sequence $\{\hat{\mathrm{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K}_\mathsf{m}$ where K_m 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 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$. Then:*

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2] \leq \frac{4\alpha\overline{L}n^{2/3}}{\mathsf{P}_\mathsf{m} \upsilon_{\min}^2 \upsilon_{\max}^2} \left(\mathbb{E}\big[\Delta V\big] + \sum_{k=0}^{\mathsf{K}_\mathsf{m}-1} \Xi^{(k+1)} + \Gamma^{(k+1)}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \right) \, .$$

Note that in those two bounds, the quantities $\tilde{\eta}^{(k+1)}$ 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 $\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]$. Indeed, the following remarks are worth doing on this quantity: (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 $k > 0$

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] = \mathbb{E}[\|\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k+1)}\|^2] = 0 \quad \text{with} \quad \hat{\boldsymbol{s}}^{(0)} = \tilde{S}^{(0)} = 0 \, ,$$

which strengthen the fact that this quantity characterizes the impact of the variance reduction technique introduced in our class of methods. The following Lemma characterizes this gap:

Lemma 6. *Considering a decreasing stepsize $\gamma_k \in (0,1)$ and a constant $\rho \in (0,1)$, we have*

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \leq \frac{\rho}{1-\rho} \sum_{\ell=0}^{k} (1-\gamma_\ell)^2 (\boldsymbol{\mathcal{S}}^{(\ell)} - \tilde{S}^{(\ell)}) ,$$

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

4 Numerical Examples

This section presents several numerical applications for our proposed class of Algorithms 1.

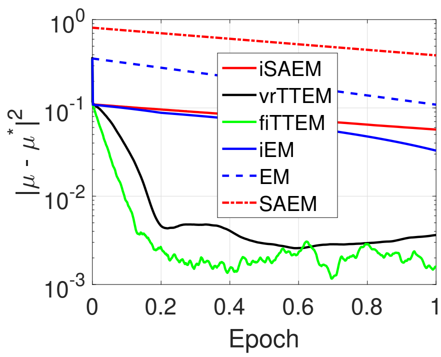

Figure 1: Precision $|\mu^{(k)} - \mu^*|^2$ per epoch

4.1 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 $\{y_i\}_{i=1}^n$, we want 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:

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

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 $\Theta = \{\omega_m,\ m = 1, ..., M-1 : \omega_m \geq 0,\ \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R},\ m = 1, ..., M\}$. 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.

4.2 Deformable Template Model for Image Analysis

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:

$$y_i(u) = I\left(x_u - \Phi_i\left(x_u, z_i\right)\right) + \varepsilon_i(u) \tag{15}$$

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:

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

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:

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

where we put a Gaussian prior on the latent variables, $z_i \sim \mathcal{N}(0,\Gamma)$ and $z_i \in \left(\mathbb{R}^{k_g}\right)^2$. The vector of parameters we estimate is thus $\boldsymbol{\theta} = \left(\beta, \Gamma, \sigma\right)$. The complete model (15) belongs to the curved exponential family, see [2], which vector of sufficient statistics for all $i \in [n]$ is defined by $S(y_i, z_i) = (\mathbf{K}_{p,z_i}^\top y_i, \mathbf{K}_{p,z_i}^\top \mathbf{K}_{p,z_i}, z_i^t z_i)$ where we denote $\mathbf{K}_{p,z_i} = \mathbf{K}_{p,z_i}(x_u - \phi_i(x_u, z_i), p_j)$. Then, the two-timescale M-step (4) yields the following parameter updates $\bar{\boldsymbol{\theta}}(\hat{s}) = \left(\boldsymbol{\beta}(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z), \boldsymbol{\Gamma}(\hat{s}) = \hat{s}_3(z)/n, \boldsymbol{\sigma}(\hat{s}) = \boldsymbol{\beta}(\hat{s})^\top \hat{s}_2(z)\boldsymbol{\beta}(\hat{s}) - 2\boldsymbol{\beta}(\hat{s})\hat{s}_1(z)\right)$ where $\hat{s} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via update (9) in Algorithm 1.

Numerical Experiment: We apply model (15) and our Algorithm 1 to a collection of handwritten digits, called the US postal database [17], 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. We thus ought to use our deformable template model (15) in order to account for both sources of variability: the intrinsic template to 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)

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. For all methods, the initialization of the template (16) 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. Those hyperparameters are inspired by relevant studies [1, 3]. In particular, the choice of the geometric covariance, indexed by g, in such 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.

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

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

5 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 our newly introduced scheme. 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 random termination number. Numerical examples illustrate the benefits of our scheme on synthetic and real tasks.

6 Broader Impact

Our work aims at improving training procedures for latent data models. Latent data models are particularly interesting in several impactful domains such as sociology, economy or pharmacology. Indeed, in those latter domains, a special instance of latent data models, namely mixed-effect models, can be employed. It considers a latent structure in order to take into account the variability between subjects, which can be individuals, companies or patients. In that case, our class of algorithms becomes useful as demonstrated in the additional experiment in the supplementary material. It is worth noting that other types of latent variable models could be used for impactful research, such as the missing data framework when the considered observations are sensible, yet missing.

References

[1] Stéphanie Allassonnière and Estelle Kuhn. Stochastic algorithm for parameter estimation for dense deformable template mixture model. *arXiv preprint arXiv:0802.1521*, 2008.

[2] Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

[3] Stéphanie Allassonnière, Estelle Kuhn, Alain Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16 (3):641–678, 2010.

[4] Charlotte Baey, Samis Trevezas, and Paul-Henry Cournède. 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, 2016.

[5] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112(518):859–877, JUN 2017. ISSN 0162-1459. doi: {10.1080/01621459.2017.1285773}.

[6] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

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

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

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

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

[11] Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. Stochastic expectation maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988, 2018.

[12] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL https://doi.org/10.1214/aos/1018031103.

[13] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

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

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

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

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

[18] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *arXiv preprint arXiv:1712.07897*, 2017.

[19] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[20] Belhal Karimi, Hoi-To Wai, Éric Moulines, and Marc Lavielle. On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems*, pages 2833–2843, 2019.

[21] Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[22] Estelle Kuhn, Catherine Matias, and Tabea Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.

[23] Percy Liang and Dan Klein. Online em for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619, 2009.

[24] Florian Maire, Eric Moulines, and Sidonie Lefebvre. Online em for functional data, 2016. URL http://arxiv.org/abs/1604.00570. cite arxiv:1604.00570v1.pdf.

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

[26] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[27] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[28] Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[29] SK Ng and GJ McLachlan. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Statistics and Computing*, 13(1):45–55, FEB 2003. ISSN 0960-3174. doi: {10.1023/A:1021987710829}.

[30] Hien D Nguyen, Florence Forbes, and Geoffrey J McLachlan. Mini-batch learning of exponential family finite mixture models. *Statistics and Computing*, pages 1–18, 2020.

[31] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*, 2016.

[32] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[33] Greg CG Wei and Martin A Tanner. 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, 1990.

[34] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

A Proofs for the iSAEM Algorithm

A.1 Proof of Lemma 2

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

$$\upsilon_{\min}^{-1}\langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\rangle \geq \|\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq \upsilon_{\max}^{-2}\|\nabla V(\mathbf{s})\|^2, \tag{16}$$

Proof Using A3 and the fact that we can exchange integration with differentiation and the Fisher's identity, we obtain

$$\begin{aligned}
\nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \Big(\nabla_{\boldsymbol{\theta}}\, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathsf{L}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big) \\
&= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \Big(\nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}}\, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big) \\
&= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top\, \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top (\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})))\,,
\end{aligned} \tag{17}$$

Consider the following vector map:

$$\mathbf{s} \to \nabla_{\boldsymbol{\theta}} L(\mathbf{s}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\mathbf{s})} = \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}}\, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{s}\,.$$

Taking the gradient of the above map *w.r.t.* $\mathbf{s}$ and using assumption A3, we show that:

$$\mathbf{0} = -\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \Big(\underbrace{\nabla_{\boldsymbol{\theta}}^2 \big(\psi(\boldsymbol{\theta}) + \mathrm{r}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) \,|\, \mathbf{s}\rangle\big)}_{=\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s};\boldsymbol{\theta})}|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\mathbf{s})}\Big)\, \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})\,.$$

The above yields

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{B}(\mathbf{s})(\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})))\,,$$

where we recall $\mathrm{B}(\mathbf{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}; \overline{\boldsymbol{\theta}}(\mathbf{s}))\Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top$. The proof of (16) follows directly from the assumption A4. □

A.2 Proof of Theorem 1

Beforehand, We present two intermediary Lemmas important for the analysis of the incremental update of the iSAEM algorithm. The first one gives a characterization of the quantity $\mathbb{E}[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}]$:

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

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}\big(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\big)\,.$$

Also:

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

where $\overline{\mathbf{s}}^{(k)}$ *is defined by* (3) *and* $\tau_i^k = \max\{k' : i_{k'} = i,\ k' < k\}$.

Proof From update (1), we have:

$$\begin{aligned}
\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= \tilde{S}^{(k)} - \hat{\mathbf{s}}^{(k)} + \frac{1}{n}\left(\tilde{S}_{i_k}^{(k+1)} - \tilde{S}_{i_k}^{(\tau_i^k)}\right) \\
&= \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{S}^{(k)} - \overline{\mathbf{s}}^{(k)} - \frac{1}{n}\left(\tilde{S}_{i_k}^{(\tau_i^k)} - \tilde{S}_{i_k}^{(k+1)}\right)\,.
\end{aligned}$$

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

$$\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{S}^{(k)} - \overline{\mathbf{s}}^{(k)} - \frac{1}{n}\left(\tilde{S}_{i_k}^{(\tau_i^k)} - \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})\right) + \frac{1}{n}\eta_{i_k}^{(k+1)}\,.$$

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

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

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 $\mathbb{E}\left[\overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})|\mathcal{F}_k\right] = \overline{\mathbf{s}}^{(k)}$, we conclude the proof of the Lemma. □

We also derived the following auxiliary Lemma which sets an upper bound for the quantity $\mathbb{E}[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2]$:

Lemma 7. *For any $k \geq 0$ and consider the iSAEM update in (1), it holds that*

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

Proof Applying the iSAEM update yields:

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

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.

□

Theorem. *Assume A1-A5. Consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ obtained with $\rho_{k+1} = 1$ for any $k \leq \mathsf{K}_\mathsf{m}$ where K_m is a positive integer. 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:*

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

Proof Under the smoothness of the Lyapunov function V (cf. Lemma 1), we can write:

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

Taking the expectation on both sides yields:

$$\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k+1)})\right] \leq \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)})\right] + \gamma_{k+1}\mathbb{E}\left[\langle\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right] \,.$$

Using Lemma 3, we obtain:

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

where (a) is due to the growth condition (2) and (b) is due to Young's inequality (with $\beta \to 1$). Note $a_k = \gamma_{k+1}\left(\upsilon_{\min} - \upsilon_{\max}^2\frac{\beta(n-1)+1}{2n}\right)$ and

$$
\begin{aligned}
a_k\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \leq& \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right] \\
&+ \frac{\gamma_{k+1}(1-\frac{1}{n})}{2\beta}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{\gamma_{k+1}}{2n}\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \,.
\end{aligned}
\tag{18}
$$

We now give an upper bound of $\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right]$ using Lemma 7 and plug it into (18):

$$
\begin{aligned}
&(a_k - 2\gamma_{k+1}^2 \mathrm{L}_V)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\
\leq& \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] \\
&+ \gamma_{k+1}\left(\frac{1}{2\beta}(1-\frac{1}{n}) + 2\gamma_{k+1}\mathrm{L}_V\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+ \gamma_{k+1}\left(\gamma_{k+1}\mathrm{L}_V + \frac{1}{2n}\right)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\
&+ \frac{\gamma_{k+1}^2 \mathrm{L}_V \mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2] \,.
\end{aligned}
\tag{19}
$$

Next, we observe that

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

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

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\big\rangle\Big] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 - 2\gamma_{k+1}\big\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\big\rangle\Big] \\
\leq&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 + \gamma_{k+1}\beta\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2\Big] \ ,
\end{aligned}
$$

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

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

Observe that $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)})$. Applying Lemma 7 yields

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

Let us define

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

From the above, we get

$$
\begin{aligned}
\Delta^{(k+1)} \leq& \big(1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})\big)\Delta^{(k)} + 4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\Big[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2\Big] \\
&+ 2\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] + 4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \ .
\end{aligned}
$$

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

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

Note $\omega_{k,\ell} = \prod_{j=\ell+1}^{k} \left(1 - \Lambda_{(j)}\right)$ Summing on both sides over $k = 0$ to $k = \mathsf{K}_{\mathsf{m}} - 1$ yields:

$$
\begin{aligned}
& \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \Delta^{(k+1)} \\
= & 4\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} (\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})\omega_{k,1}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + 2\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} (\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})\omega_{k,1}\mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\
& + \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} 4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})\omega_{k,1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
\leq & \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \frac{4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})}{\Lambda_{(k+1)}}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \frac{2(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})}{\Lambda_{(k+1)}}\mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\
& + \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \frac{4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})}{\Lambda_{(k+1)}}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] .
\end{aligned}
\tag{20}
$$

We recall (19) where we have summed on both sides from $k = 0$ to $k = \mathsf{K}_{\mathsf{m}} - 1$:

$$
\begin{aligned}
& \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \left(a_k - 2\gamma_{k+1}^2\,\mathrm{L}_V\right)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\
\leq & \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(K)})\right] \\
& + \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \gamma_{k+1}\left(\frac{1}{2\beta}(1 - \frac{1}{n}) + 2\gamma_{k+1}\,\mathrm{L}_V\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
& + \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \gamma_{k+1}\left(\gamma_{k+1}\,\mathrm{L}_V + \frac{1}{2n}\right)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\
& + \sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1} \frac{\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2}{n^2}\Delta^{(k)} .
\end{aligned}
\tag{21}
$$

Plugging (20) into (21) results in:

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

where

$$
\begin{aligned}
\tilde{\alpha}_k &= a_k - 2\gamma_{k+1}^2 \, \mathrm{L}_V - \frac{\gamma_{k+1}^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2}{n^2} \frac{4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)}{\Lambda_{(k+1)}} \ , \\
\tilde{\beta}_k &= \gamma_{k+1}\left(\frac{1}{2\beta}(1 - \frac{1}{n}) + 2\gamma_{k+1} \, \mathrm{L}_V\right) - \frac{\gamma_{k+1}^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2}{n^2} \frac{4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)}{\Lambda_{(k+1)}} \ , \\
\tilde{\Gamma}_k &= \gamma_{k+1}\left(\gamma_{k+1} \, \mathrm{L}_V + \frac{1}{2n}\right) + \frac{\gamma_{k+1}^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2}{n^2} \frac{2\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)}{\Lambda_{(k+1)}} \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
&a_k = \gamma_{k+1}\left(\upsilon_{\min} - \upsilon_{\max}^2 \frac{\beta(n-1)+1}{2n}\right) \ , \\
&\Lambda_{(k+1)} = \frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1} \, \mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \ , \\
&c_1 = \upsilon_{\min}^{-1}, \alpha = \max\{8, 1 + 6\upsilon_{\min}\}, \overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}, \gamma_{k+1} = \frac{1}{k\alpha c_1 \overline{L}}, \beta = \frac{c_1 \overline{L}}{n} \ .
\end{aligned}
$$

When, for any $k > 0$, $\tilde{\alpha}_k \geq 0$, we have by Lemma 2 that:

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

which yields an upper bound of the gradient of the Lyapunov function V along the path of the iSAEM update and concludes the proof of the Theorem. □

B Proofs for the vrTTEM and the fiTTEM Algorithms

B.1 Proofs of Auxiliary Lemmas (Lemma 4, Lemma 5 and Lemma 6)

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

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

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:

$$
\begin{aligned}
\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) \\
&= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho \boldsymbol{\mathcal{S}}^{(k+1)}) \\
&= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\right]\right) \ .
\end{aligned}
\tag{22}
$$

We observe, using the identity (22), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]+2\rho^2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\boldsymbol{\mathcal{S}}^{(k+1)}\|^2]+2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))}-\tilde{S}^{(k)}\|^2]. \tag{23}$$

For the latter term, we obtain its upper bound as

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

where (a) uses the variance inequality and (b) uses Lemma 1. Substituting into (23) proves the lemma. □

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

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

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

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

$$\begin{aligned}
\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) \\
&= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho\boldsymbol{\mathcal{S}}^{(k+1)}) \\
&= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\right]\right) \\
&= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)} - \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)\right]\right) \,.
\end{aligned} \tag{24}$$

We observe, using the identity (24), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]+2\rho^2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\boldsymbol{\mathcal{S}}^{(k+1)}\|^2]+2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))}-\tilde{S}^{(k)}\|^2] \,. \tag{25}$$

For the latter term, we obtain its upper bound as

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

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 (25) proves the lemma. □

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

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

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

Proof We begin by writing the two-timescale update:

$$\begin{aligned}\tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho\big(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k)}\big) \;, \\ \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \;,\end{aligned} \tag{26}$$

where $\boldsymbol{\mathcal{S}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(t_i^k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)$ according to (3). Denote $\delta^{(k+1)} = \hat{\boldsymbol{s}}^{(k+1)} - \tilde{S}^{(k+1)}$. Then from (26), doing the subtraction of both equations yields:

$$\delta^{(k+1)} = (1-\gamma_{k+1})\delta^{(k)} + \frac{\rho}{1-\rho}(1-\gamma_{k+1})(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k+1)}) \;.$$

Using the telescoping sum and noting that $\delta^{(0)} = 0$, we have

$$\delta^{(k+1)} \leq \frac{\rho}{1-\rho}\sum_{\ell=0}^{k}(1-\gamma_{\ell+1})^2(\boldsymbol{\mathcal{S}}^{(\ell+1)} - \tilde{S}^{(\ell+1)}) \;.$$

□

B.2 Additional Intermediary Result

Lemma 8. *At iteration $k+1$,the drift term of update (3), with $\rho_{k+1} = \rho$, is equivalent to the following :*

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

where we recall that $\eta_{i_k}^{(k+1)}$, defined in (12), which is the gap between the MC approximation and the expected statistics.

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

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

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 (12), 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 $\overline{\mathbf{s}}_{i_k}^{(k)}$. □

B.3 Proof of Theorem 2

Theorem. *Assume A1-A5. Consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K}_\mathsf{m}$ where K_m 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:*

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

Proof Using the smoothness of V and update (2), we obtain:

$$
\begin{aligned}
V(\hat{\boldsymbol{s}}^{(k+1)}) &\leq V(\hat{\boldsymbol{s}}^{(k)}) + \langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle + \frac{\mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \\
&\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \, .
\end{aligned}
\tag{27}
$$

Denote $\mathsf{H}_{k+1} := \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}$ the drift term of the fiTTEM update in (7) and $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathrm{s}}^{(k)}$. Taking expectations on both sides show that

$$
\begin{aligned}
&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \\
\overset{(a)}{\leq} &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}(1-\rho)\mathbb{E}\Big[\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle\Big] \\
&- \gamma_{k+1}\rho\mathbb{E}\Big[\langle \hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle\Big] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\
\overset{(b)}{\leq} &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}\rho\mathbb{E}\Big[\langle \mathsf{h}_k \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle\Big] - \gamma_{k+1}(1-\rho)\mathbb{E}\Big[\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle\Big] \\
&-\gamma_{k+1}\rho\mathbb{E}\Big[\langle \eta_{i_k}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle\Big] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\
\overset{(c)}{\leq} &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \left(\gamma_{k+1}\rho\upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2\right)\mathbb{E}\Big[\|\mathsf{h}_k\|^2\Big] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\
&-\gamma_{k+1}\rho\mathbb{E}\left[\left\|\eta_{i_k}^{(k+1)}\right\|^2\right] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \, ,
\end{aligned}
\tag{28}
$$

where we have used (22) in (a) and $\mathbb{E}\left[\boldsymbol{\mathcal{S}}^{(k+1)}\right] = \overline{\mathrm{s}}^{(k)} + \mathbb{E}[\eta_{i_k}^{(k+1)}]$ in (b), the growth condition in Lemma 2 and Young's inequality with the constant equal to 1 in (c).

Furthermore, for $k+1 \leq \ell(k) + m$ (*i.e.*, $k+1$ is in the same epoch as k), we have

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] = \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} + \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + 2\langle \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))} \,|\, \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \rangle\Big] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2\|\mathsf{H}_{k+1}\|^2 \\
&-2\gamma_{k+1}\langle \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))} \,|\, \rho(\mathsf{h}_k - \eta_{i_k}^{(k+1)}) + (1-\rho)(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}) \rangle\Big] \\
\leq&\mathbb{E}\Big[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2\|\mathsf{H}_{k+1}\|^2 + \frac{\gamma_{k+1}\rho}{\beta}\|\mathsf{h}_k\|^2 \\
&+\frac{\gamma_{k+1}\rho}{\beta}\|\eta_{i_k}^{(k+1)}\|^2 + \frac{\gamma_{k+1}(1-\rho)}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\Big] \, ,
\end{aligned}
$$

where we first used (22) and the last inequality is due to Young's inequality.

Consider the following sequence

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

where $b_k := \overline{b}_{k \bmod m}$ 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 $\overline{b}_i$ is decreasing with i and this implies

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

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

$$
\begin{aligned}
R_{k+1} \leq & \mathbb{E}\Big[V(\hat{\boldsymbol{s}}^{(k)}) - \big(\gamma_{k+1}\rho\upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2\big)\|\mathsf{h}_k\|^2 + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\|\mathsf{H}_{k+1}\|^2\Big] \\
& + \gamma_{k+1}\mathbb{E}\left[\rho\left\|\eta_{i_k}^{(k+1)}\right\|^2 - (1-\rho)\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \\
& + b_{k+1}\mathbb{E}\left[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2\|\mathsf{H}_{k+1}\|^2 + \frac{\gamma_{k+1}\rho}{\beta}\|\mathsf{h}_k\|^2\right] \\
& + b_{k+1}\mathbb{E}\left[\frac{\gamma_{k+1}\rho}{\beta}\|\eta_{i_k}^{(k+1)}\|^2 + \frac{\gamma_{k+1}(1-\rho)}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] .
\end{aligned}
$$

And using Lemma 4 we obtain:

$$
\begin{aligned}
& R_{k+1} \\
\leq & \mathbb{E}\Big[V(\hat{\boldsymbol{s}}^{(k)}) - \big(\gamma_{k+1}\rho\upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2 - \gamma_{k+1}^2\rho^2\mathrm{L}_V\big)\|\mathsf{h}_k\|^2 + \gamma_{k+1}^2\rho^2\mathrm{L}_V\mathrm{L}_{\mathbf{s}}^2\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2\Big] \\
& + b_{k+1}\mathbb{E}\left[(1+\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2\mathrm{L}_{\mathbf{s}}^2)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + (\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\|\mathsf{h}_k\|^2\right] \\
& + \gamma_{k+1}\mathbb{E}\left[(\rho + \rho^2\gamma_{k+1}\mathrm{L}_V)\left\|\eta_{i_k}^{(k+1)}\right\|^2 - (1-\rho-(1-\rho)^2\gamma_{k+1}\mathrm{L}_V)\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \\
& + b_{k+1}\mathbb{E}\left[(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\|\eta_{i_k}^{(k+1)}\|^2 + (\frac{\gamma_{k+1}(1-\rho)}{\beta} + 2\gamma_{k+1}^2(1-\rho)^2)\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] .
\end{aligned}
$$

Rearranging the terms yields:

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

where

$$
\begin{aligned}
\tilde{\eta}^{(k+1)} &= \left(\gamma_{k+1}(\rho + \rho^2\gamma_{k+1}\mathrm{L}_V) + b_{k+1}(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\right)\mathbb{E}\left[\left\|\eta_{i_k}^{(k+1)}\right\|^2\right] \\
\chi^{(k+1)} &= \left(b_{k+1}(\frac{\gamma_{k+1}(1-\rho)}{\beta} + 2\gamma_{k+1}^2(1-\rho)^2) - \gamma_{k+1}(1-\rho-(1-\rho)^2\gamma_{k+1}\mathrm{L}_V)\right) \\
\tilde{\chi}^{(k+1)} &= \chi^{(k+1)}\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] .
\end{aligned}
$$

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

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

We first remark that

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

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

$$
\begin{aligned}
&1-\gamma_{k+1}c_1\rho\,\mathrm{L}_V-b_{k+1}\big(\frac{c_1}{\beta}+2\gamma_{k+1}\rho c_1\big)\\
\geq& 1-\frac{\mu}{n^{\frac{2}{3}}}-\bar{b}_0\big(\frac{n^{\frac{1}{3}}}{\overline{L}}+\frac{2\mu}{\overline{L}n^{\frac{2}{3}}}\big)\\
\geq& 1-\frac{\mu}{n^{\frac{2}{3}}}-\frac{\mathrm{L}_V\,\mu^2}{c_1^2 n^{\frac{4}{3}}}\frac{(1+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2)^m-1}{\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}\big(\frac{n^{\frac{1}{3}}}{\overline{L}}+\frac{2\mu}{\overline{L}n^{\frac{2}{3}}}\big)\\
\overset{(a)}{\geq}& 1-\frac{\mu}{n^{\frac{2}{3}}}-\frac{\mu}{c_1^2}(\mathrm{e}-1)\big(1+\frac{2\mu}{n}\big)\geq 1-\mu-\mu(1+2\mu)\frac{\mathrm{e}-1}{c_1^2}\overset{(b)}{\geq}\frac{1}{2}\;,
\end{aligned}
$$

where the simplification in (a) is due to

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

and the required μ in (b) can be found by solving the quadratic equation.

Finally, these results yield:

$$
\upsilon_{\max}^2\sum_{k=0}^{\mathsf{K_m}-1}\gamma_{k+1}\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2]\leq\frac{2(R_0-R_{\mathsf{K_m}})}{\upsilon_{\min}\rho}+2\sum_{k=0}^{\mathsf{K_m}-1}\frac{\tilde{\eta}^{(k+1)}+\tilde{\chi}^{(k+1)}}{\upsilon_{\min}\rho}\;.
$$

Note that $R_0=\mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)})]$ and if $\mathsf{K_m}$ is a multiple of m, then $R_{\mathsf{max}}=\mathbb{E}[V(\hat{\boldsymbol{s}}^{(\mathsf{K_m})})]$. Under the latter condition, we have

$$
\begin{aligned}
\sum_{k=0}^{\mathsf{K_m}-1}\gamma_{k+1}\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2]\leq&\frac{2n^{2/3}\overline{L}}{\mu\upsilon_{\min}^2\upsilon_{\max}^2}\mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)})-V(\hat{\boldsymbol{s}}^{(\mathsf{K_m})})]\\
&+\frac{2n^{2/3}\overline{L}}{\mu\upsilon_{\min}^2\upsilon_{\max}^2}\sum_{k=0}^{\mathsf{K_m}-1}\Big[\tilde{\eta}^{(k+1)}+\tilde{\chi}^{(k+1)}\Big]\;.
\end{aligned}
$$

This concludes our proof.

□

B.4 Proof of Theorem 3

Theorem. *Assume A1-A5. Consider the fiTTEM sequence $\{\hat{\mathrm{s}}^{(k)}\}_{k>0}\in\mathcal{S}$ for any $k\leq\mathsf{K_m}$ where $\mathsf{K_m}$ 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 $c_1(k\alpha-1)\geq c_1(\alpha-1)\geq 2$, $\alpha\geq 2$. Then:*

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

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

$$
\begin{aligned}
V(\hat{\boldsymbol{s}}^{(k+1)})\leq& V(\hat{\boldsymbol{s}}^{(k)})+\big\langle\hat{\boldsymbol{s}}^{(k+1)}-\hat{\boldsymbol{s}}^{(k)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle+\frac{\mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k+1)}-\hat{\boldsymbol{s}}^{(k)}\|^2\\
\leq& V(\hat{\boldsymbol{s}}^{(k)})-\gamma_{k+1}\big\langle\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle+\frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}\|^2\;.
\end{aligned}
\tag{29}
$$

Denote $\mathsf{H}_{k+1}:=\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}$ the drift term of the fiTTEM update in (7) and $\mathsf{h}_k=\hat{\boldsymbol{s}}^{(k)}-\overline{\mathbf{s}}^{(k)}$. Using Lemma 8 and the additional following identity:

$$
\mathbb{E}\left[\big(\overline{\mathbf{s}}_{i_k}^{(k)}-\tilde{S}_{i_k}^{(t_{i_k}^k)}\big)-\mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)}-\tilde{S}_{i_k}^{(t_{i_k}^k)}]\right]=0\;,
\tag{30}
$$

we have:

$$
\begin{aligned}
&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})]\\
\leq&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}\rho\mathbb{E}[\langle \mathsf{h}_k \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle]\\
&- \gamma_{k+1}\mathbb{E}\left[\left\langle \rho\mathbb{E}[\eta_{i_k}^{(k+1)}|\mathcal{F}_k] + (1-\rho)\mathbb{E}[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}] \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle\right] + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\mathsf{H}_{k+1}\|^2\\
\overset{(a)}{\leq}& - \upsilon_{\min}\gamma_{k+1}\rho\mathbb{E}[\|\mathsf{h}_k\|^2] - \gamma_{k+1}\mathbb{E}\left[\left\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\right\|^2\right]\\
&- \frac{\gamma_{k+1}\rho^2}{2}\xi^{(k+1)} - \frac{\gamma_{k+1}(1-\rho)^2}{2}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\mathsf{H}_{k+1}\|^2\\
\overset{(b)}{\leq}& - (\upsilon_{\min}\gamma_{k+1}\rho + \gamma_{k+1}\upsilon_{\max}^2)\mathbb{E}[\|\mathsf{h}_k\|^2] - \frac{\gamma_{k+1}\rho^2}{2}\xi^{(k+1)} - \frac{\gamma_{k+1}(1-\rho)^2}{2}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\\
&+ \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\mathsf{H}_{k+1}\|^2 ,
\end{aligned}
$$

where $\xi^{(k+1)} = \mathbb{E}[\|\mathbb{E}[\eta_{i_k}^{(k+1)}|\mathcal{F}_k]\|^2]$.

Bounding $\mathbb{E}\left[\|\mathsf{H}_{k+1}\|^2\right]$ Using Lemma 5, we obtain:

$$
\begin{aligned}
&\gamma_{k+1}(\upsilon_{\min}\rho + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2\,\mathrm{L}_V)\mathbb{E}[\|\mathsf{h}_k\|^2]\\
\leq&\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] + \tilde{\xi}^{(k+1)} + \left((1-\rho)^2\gamma_{k+1}^2\,\mathrm{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\\
&+ \frac{\gamma_{k+1}^2\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] ,
\end{aligned}
\tag{31}
$$

where $\tilde{\xi}^{(k+1)} = \gamma_{k+1}^2\rho^2\,\mathrm{L}_V\,\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] - \frac{\gamma_{k+1}\rho^2}{2}\xi^{(k+1)}$. Next, we observe that

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]\right) ,
\tag{32}
$$

where the equality holds as i_k and j_k are drawn independently. Then,

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]\\
&= \mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\big\rangle\Big] .
\end{aligned}
$$

Note that $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) = -\gamma_{k+1}\mathsf{H}_{k+1}$ and that in expectation we recall that $\mathbb{E}[\mathsf{H}_{k+1}|\mathcal{F}_k] = \rho\mathsf{h}_k + \rho\mathbb{E}[\eta_{i_k}^{(k+1)}|\mathcal{F}_k] + (1-\rho)\mathbb{E}[\tilde{S}^{(k)} - \hat{\boldsymbol{s}}^{(k)}]$ where $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}$. Thus, for any $\beta > 0$, it holds

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]\\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\big\rangle\Big]\\
\leq&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + (1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathsf{h}_k\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\left\|\eta_{i_k}^{(k+1)}\right\|^2]\\
&+ \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\Big] ,
\end{aligned}
$$

where the last inequality is due to Young's inequality. Plugging this into (32) yields:

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)} \big\rangle\Big] \\
\leq&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + (1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathsf{h}_k\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] \\
&+ \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}\left[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2\right]\Big] .
\end{aligned}
$$

Subsequently, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] \\
\leq&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n^2}\sum_{i=1}^{n}\mathbb{E}\Big[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathsf{h}_k\|^2 \\
&+ \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] + \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}\left[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2\right]\Big] .
\end{aligned}
$$

We now use Lemma 5 on $\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 = \gamma_{k+1}^2\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2$ and obtain:

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

Let us define

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

From the above, we get

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

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

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

which shows 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:

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

where $\omega_{k,\ell}=\prod_{j=\ell+1}^{k}\left(1-\Lambda_{(j)}\right)$ 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 on both sides over $k=0$ to $k=\mathsf{K}_\mathsf{m}-1$ yields:

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

We recall (31) where we have summed on both sides from $k=0$ to $k=\mathsf{K}_\mathsf{m}-1$:

$$
\begin{aligned}
& \mathbb{E}\big[V(\hat{\mathbf{s}}^{(\mathsf{K}_\mathsf{m})})-V(\hat{\mathbf{s}}^{(0)})\big] \\
\leq & \sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\left\{\gamma_{k+1}(-(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}\rho^2\,\mathrm{L}_V)\mathbb{E}[\|\mathsf{h}_k\|^2]+\gamma^2\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2\,\Delta^{(k)}\right\} \\
& + \sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\left\{\tilde{\xi}^{(k+1)}+\left((1-\rho)^2\gamma_{k+1}^2\,\mathrm{L}_V-\frac{\gamma_{k+1}(1-\rho)^2}{2}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k)}\|^2]\right\} \\
\leq & \sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\left\{\left[-\gamma_{k+1}(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}^2\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right]\mathbb{E}[\|\mathsf{h}_k\|^2]\right\} \\
& + \sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\Xi^{(k+1)}+\sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\Gamma^{(k+1)}\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k)}\|^2\right]\,,
\end{aligned}
\tag{33}
$$

where

$$
\Xi^{(k+1)}=\tilde{\xi}^{(k+1)}+\frac{\gamma_{k+1}^3\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2}{\Lambda_{(k+1)}}\tilde{\epsilon}^{(k+1)}
$$

and

$$
\Gamma^{(k+1)}=\left((1-\rho)^2\gamma_{k+1}^2\,\mathrm{L}_V-\frac{\gamma_{k+1}(1-\rho)^2}{2}\right)+\frac{\gamma_{k+1}^3\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2(1-\rho)^2\left(2\gamma_{k+1}+\frac{1}{\beta}\right)}{\Lambda_{(k+1)}}\,.
$$

We now analyse the following quantity

$$
\begin{aligned}
& -\gamma_{k+1}(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}^2\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \\
& =\gamma_{k+1}\left[-(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right]\,.
\end{aligned}
\tag{34}
$$

Furthermore, we recall that $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{2, 1 + 2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\gamma_{k+1} = \frac{1}{k}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \overline{L} n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$.Then,

$$
\begin{aligned}
&\gamma_{k+1}\rho^2 \mathrm{L}_V + \frac{\rho^2 \gamma_{k+1} \mathrm{L}_V \mathrm{L}_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2 \rho^2 \mathrm{L}_{\mathbf{s}}^2} \\
\leq & \frac{1}{k\alpha^2 c_1^2 \overline{L} n^{4/3}} + \frac{\overline{L}(k\alpha^2 c_1^2 n^{4/3})^{-1}\left(\frac{2}{k^2\alpha^2 c_1^2 \overline{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \overline{L}^2 n^{1/3}}\right)}{\frac{1}{n} - \frac{1}{k\alpha n} - \frac{1}{k^2\alpha^2 c_1^2 n^{4/3}}} \\
= & \frac{1}{k\alpha^2 c_1^2 \overline{L} n^{4/3}} + \frac{\overline{L}\left(\frac{2}{k^2\alpha^2 c_1^2 \overline{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \overline{L}^2 n^{1/3}}\right)}{(k\alpha c_1 n^{1/3})(k\alpha - 1)c_1 - 1} \\
\overset{(a)}{\leq} & \frac{1}{k\alpha^2 c_1^2 \overline{L} n^{4/3}} + \frac{\frac{1}{k\alpha c_1^2 \overline{L} n^{1/3}}\left(\frac{2}{k\alpha n} + 1\right)}{2(\alpha c_1 n^{1/3}) - 1} \\
\leq & \frac{1}{k^2 \alpha c_1^2 \overline{L} n^{4/3}} + \frac{1}{4k\alpha^2 c_1^3 \overline{L} n^{2/3}} \\
\leq & \frac{3/4}{\alpha c_1^2 \overline{L} n^{2/3}} ,
\end{aligned}
\tag{35}
$$

where (a) is due to $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$ and $k\alpha c_1 n^{1/3} \geq 1$. Note also that

$$
-(\upsilon_{\min}\rho + \upsilon_{\max}^2) \leq -\rho\upsilon_{\min} = -\frac{1}{\alpha c_1^2 \overline{L} n^{2/3}} ,
$$

which yields that

$$
\left[-(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}\rho^2 \mathrm{L}_V + \frac{\rho^2 \gamma_{k+1} \mathrm{L}_V \mathrm{L}_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \right] \leq -\frac{1/4}{\alpha c_1^2 \overline{L} n^{2/3}} .
$$

Using the Lemma 2, we know that $\upsilon_{\max}^2 \|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2 \leq \|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2$ and using (35) on (33) yields:

$$
\begin{aligned}
&\upsilon_{\max}^2 \sum_{k=0}^{\mathsf{K_m}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \\
\leq & \frac{4\alpha \overline{L} n^{2/3}}{\upsilon_{\min}^2} \left[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(\mathsf{K_m})})\right] \\
& + \frac{4\alpha \overline{L} n^{2/3}}{\upsilon_{\min}^2} \sum_{k=0}^{\mathsf{K_m}-1} \Xi^{(k+1)} + \sum_{k=0}^{\mathsf{K_m}-1} \Gamma^{(k+1)} \mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] ,
\end{aligned}
$$

proving the bound on the second order moment of the gradient of the Lyapunov function:

$$
\begin{aligned}
\sum_{k=0}^{\mathsf{K_m}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq & \frac{4\alpha \overline{L} n^{2/3}}{\upsilon_{\min}^2 \upsilon_{\max}^2} \left[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(\mathsf{K_m})})\right] \\
& + \frac{4\alpha \overline{L} n^{2/3}}{\upsilon_{\min}^2 \upsilon_{\max}^2} \sum_{k=0}^{\mathsf{K_m}-1} \Xi^{(k+1)} + \sum_{k=0}^{\mathsf{K_m}-1} \Gamma^{(k+1)} \mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] .
\end{aligned}
$$

$\square$

C Practical Implementations of Two-Timescale EM Methods

C.1 Application on GMM

C.1.1 Explicit 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 (2) with

$$\begin{aligned} &s_{i,m}^{(1)} = \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{\log(\omega_m) - \frac{\mu_m^2}{2}\right\} - \left\{\log(1 - \textstyle\sum_{j=1}^{M-1}\omega_j) - \frac{\mu_M^2}{2}\right\} , \\ &s_{i,m}^{(2)} = \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m , \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M , \end{aligned} \tag{36}$$

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 $m \in [\![1, M]\!]$, $j \in [\![1, 3]\!]$, $s_m^{(j)} = n^{-1}\sum_{i=1}^n s_{i,m}^{(j)}$. Consider the following latent sample used to compute an approximation of the conditional expected value $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y = y_i]$:

$$z_{i,m} \sim \mathbb{P}\left(z_i = m | y_i; \boldsymbol{\theta}\right) \tag{37}$$

where $m \in [\![1, M]\!]$, $i \in [n]$ 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 Incremental-step updates, see (8) can be written as

$$\tilde{S}_{i_k}^{(k)} = (\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1})y_{i_k}, ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\bar{\mathbf{s}}_{i_k}^{(3)}(\boldsymbol{\theta}^{(k)})})^\top . \tag{38}$$

Recall that we have used the following regularizer:

$$\mathrm{r}(\boldsymbol{\theta}) = \tfrac{\delta}{2}\textstyle\sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log\left(1 - \sum_{m=1}^{M-1}\omega_m\right) , \tag{39}$$

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

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1+\epsilon M)^{-1}\big(s_1^{(1)} + \epsilon, \ldots, s_{M-1}^{(1)} + \epsilon\big)^\top \\ \big((s_1^{(1)} + \delta)^{-1}s_1^{(2)}, \ldots, (s_{M-1}^{(1)} + \delta)^{-1}s_{M-1}^{(2)}\big)^\top \\ \big(1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta\big)^{-1}\big(s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}\big) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{40}$$

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

C.1.2 Model Assumptions (GMM example)

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 (11) as the sufficient statistics are composed of indicator functions and observations as defined Section C.1 Equation (36).

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:

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

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 $\mathrm{B}(\mathbf{s})$ 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 process.

C.1.3 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 (38). At iteration k, the several E-steps defined by (1) or (2) and (3) leads to the definition of the quantity $\hat{\mathbf{s}}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{\mathbf{s}}^{(0)} = n^{-1} \sum_{i=1}^n \overline{\mathbf{s}}_i^{(0)}$, those E-steps break down as follows:

Batch EM (EM): for all $i \in [n]$, compute $\overline{\mathbf{s}}_i^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = n^{-1} \sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(k)} \ .$$

where $\overline{\mathbf{s}}_i^{(k)}$ are computed using the exact conditional expected balue $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y=y_i]$:

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

Incremental EM (iEM): draw an index i_k uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \frac{1}{n}\big(\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\tau_i^k)}\big) = n^{-1} \sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(\tau_i^k)} \ .$$

batch SAEM (SAEM): draw an index i_k uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\tilde{S}^{(k)} \ .$$

where $= \frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(k)}$ with $\tilde{S}_i^{(k)}$ defined in (38).

Incremental SAEM (iSAEM): draw an index i_k uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)} + \frac{1}{n}(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)})\big) \ .$$

Variance Reduced Two-Timescale EM (vrTTEM): draw an index i_k uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)}(1-\rho) + \rho(\tilde{S}^{(\ell(k))} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}\big))\big) \ .$$

Fast Incremental Two-Timescale EM (fiTTEM): draw an index i_k uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)}(1-\rho) + \rho(\overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)\big) \ .$$

Finally, the k-th update reads $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$ where the function $\boldsymbol{s} \to \overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined by (40).

C.2 Deformable Template Model for Image Analysis

C.2.1 Model and Updates

The complete model belongs to the curved exponential family, see [2], which vector of sufficient statistics $S = \big(S_1(z), S_2(z), S_3(z)\big)$ read:

$$\begin{aligned}
S_1(z) &= \frac{1}{n}\sum_{i=1}^{n} S_1(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top y_i \; , \\
S_2(z) &= \frac{1}{n}\sum_{i=1}^{n} S_2(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top \left(\mathbf{K}_p^{z_i}\right) \; , \\
S_3(z) &= \frac{1}{n}\sum_{i=1}^{n} S_3(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i \; ,
\end{aligned} \tag{41}$$

where for any pixel $u \in \mathbb{R}^2$ and $j \in [\![1, k_g]\!]$ 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:

$$\bar{\boldsymbol{\theta}}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n}\hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z)\beta(\hat{s}) - 2\beta(\hat{s})\hat{s}_1(z) \end{pmatrix} \; , \tag{42}$$

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 (7) and using the MC approximation of the sufficient statistics $\big(S_1(z), S_2(z), S_3(z)\big)$ defined in (41).

C.2.2 Numerical Applications

For the inference of the template, we use the Matlab code (online SAEM) used in [24] and implement our own batch, incremental, Variance reduced and Fast Incremental variants. The hyperparameters are kept the same and reads as follows $M = 400$, $\gamma_k = 1/k^{0.6}$ and $p = 16$. The number of landmarks for the template is $k_p = 15$ points and for the deformation $k_g = 6$ points. Both have Gaussian kernels with respectively standard deviation of 0.12 and 0.3. The standard deviation of the measurement errors is set to 0.1.

For the simulation part, we use the Carlin and Chib MCMC procedure, see [9]. Refer to [24] for more details.

D Additional Experiment: Pharmacokinetics (PK) Model with Absorption Lag Time

This numerical example was conducted in order to characterize the pharmacokinetics (PK) of orally administered drug to simulated patients, using a population pharmacokinetics approach. $M = 50$ synthetic datasets were generated for $n = 5000$ patients with 10 observations (concentration measures) per patient. The goal tis to model the evolution of the concentration of the absorbed drug using a nonlinear and latent variable model.

Model and Explicit Updates: 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. 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:

$$y_{ij} = f(t_{ij}, z_i) + \varepsilon_{ij} \quad \text{where} \quad f(t_{ij}, z_i) = \frac{D\, ka_i}{V(ka_i - k_i)}(\mathrm{e}^{-ka_i\,(t_{ij} - T_i^{\text{lag}})} - \mathrm{e}^{-k_i\,(t_{ij} - T_i^{\text{lag}})}) \ , \tag{43}$$

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.

Lognormal distributions are used for the four PK parameters:

$$\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega_{T^{\text{lag}}}^2) \ , \log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2) \ ,$$
$$\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2) \ , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2) \ .$$

We recall that the complete model (y, z) defined by (43) belongs to the curved exponential family, which vector of sufficient statistics $S = \big(S_1(z), S_2(z), S_3(z)\big)$ read:

$$S_1(z) = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad S_2(z) = \frac{1}{n}\sum_{i=1}^{n} z_i^\top z_i, \quad S_3(z) = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(t_i, z_i))^2 \tag{44}$$

where we have noted y_i and t_i the vector of observations and time for each patient i. At iteration k, and setting the number of MC samples to 1 for the sake of clarity, the MC sampling $z_i^{(k)} \sim p(z_i | y_i, \theta^{(k)})$ is performed using a Metropolis-Hastings procedure detailed in Algorithm 2. The quantities $\tilde{S}^{(k+1}$ and $\hat{\mathbf{s}}^{(k+1)}$ are then updated according to the different methods. Finally the maximization step yields:

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} \hat{\mathbf{s}}_1^{(k+1)} \\ \hat{\mathbf{s}}_2^{(k+1)} - \hat{\mathbf{s}}_1^{(k+1)} \left(\hat{\mathbf{s}}_1^{(k+1)}\right)^\top \\ \hat{\mathbf{s}}_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{z_{\text{pop}}}}(\hat{\mathbf{s}}^{(k+1)}) \\ \overline{\boldsymbol{\omega_z}}(\hat{\mathbf{s}}^{(k+1)}) \\ \overline{\boldsymbol{\sigma}}(\hat{\mathbf{s}}^{(k+1)}) \end{pmatrix} \ . \tag{45}$$

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

Metropolis Hastings algorithm. During the simulation step of the MISSO method, the sampling from the target distribution $\pi(z_i, \boldsymbol{\theta}) := p(z_i | y_i, \boldsymbol{\theta})$ is performed using a Metropolis Hastings (MH) algorithm [27] with 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. Commonly they parameterize a Gaussian proposal. The MH algorithm is summarized in 2.

Algorithm 2 MH aglorithm

Input: initialization $z_{i,0} \sim q(z_i; \boldsymbol{\delta})$
for $m = 1, \cdots, M$ **do**
  Sample $z_{i,m} \sim q(z_i; \boldsymbol{\delta})$
  Sample $u \sim \mathcal{U}(\llbracket 0, 1 \rrbracket)$
  Calculate the ratio $r = \frac{\pi(z_{i,m}; \boldsymbol{\theta})/q(z_{i,m}); \boldsymbol{\delta})}{\pi(z_{i,m-1}; \boldsymbol{\theta})/q(z_{i,m-1}); \boldsymbol{\delta})}$
  if $u < r$ **then**
    Accept $z_{i,m}$
  else
    $z_{i,m} \leftarrow z_{i,m-1}$
  end if
end for
Output: $z_{i,M}$

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 $E_k(\ell) = \frac{1}{M}\sum_{m=1}^{M}\left(\theta_k^{(m)}(\ell) - \theta^*\right)^2$ and plot it against the epochs (passes over the data) Figure 4. Note that the MC-step (5) is performed using a Metropolis Hastings procedure since the posterior distribution under the model θ noted $p(z_i|y_i,\theta)$ is intractable due to the nonlinearity of the model (43). 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.

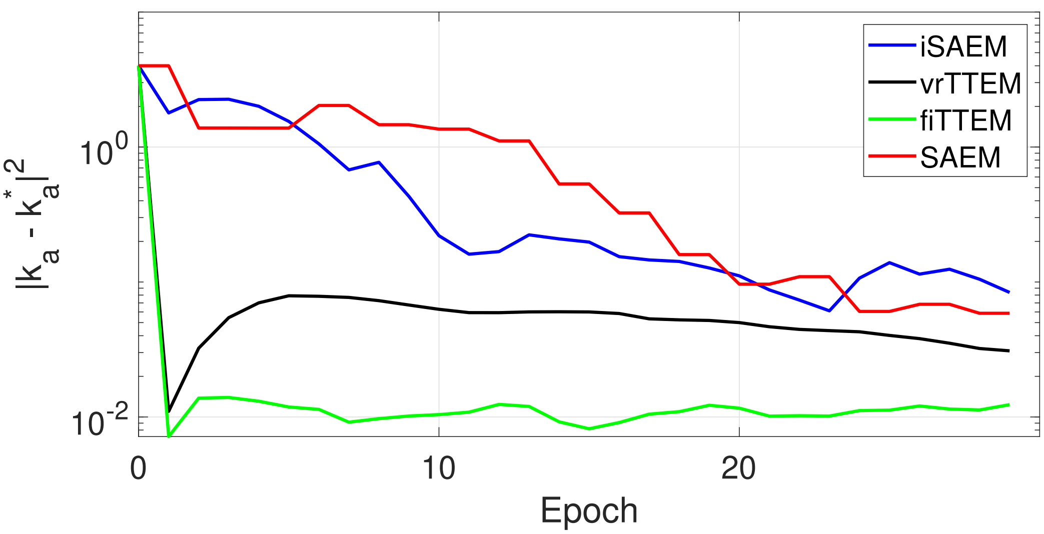

Figure 4: Precision $|ka^{(k)} - ka^*|^2$ per epoch